FROM LOCAL SCORES TO POST HOC CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Data selection often scores examples using local derivatives at a reference model, although subset quality is set by retraining. We ask when these scores certify that no alternative in a declared exchange neighbourhood improves a returned subset beyond a tolerance. Existing uniform approximation bounds support such guarantees, but symmetric corrections also account for overestimated gains. Since overprediction preserves an upper bound, we focus on correcting underestimated gains. To bound these gains, we represent an exchange as continuous weight transfer, with its local prediction giving the starting slope of the retrained-utility curve. Under verifiable smoothness and curvature conditions, we derive a one-sided correction that retains target curvature's suppression while controlling training-curvature variation and curvature differences between exchanged examples. Combined with scores recomputed at the returned subset's trained model, this correction covers the neighbourhood in one computation, without retraining or enumerating alternatives. Exhaustive CIFAR-10 one-swap auditing demonstrates certification despite prediction error, with 14.3% local overprediction of the maximum retrained gain and an analytic correction adding 0.011% to that prediction. On a fixed empirical detection target, our verified certificate covers four exchanges at tolerance 0.02, versus two for symmetric and one for convex tangent bounds.

1 INTRODUCTION

Training on a carefully chosen subset of the available data can come close to training on all of it in many large-scale settings. Common selection methods compute local scores at a reference model and keep the highest ranked examples (Koh & Liang, 2017; Xia et al., 2024). After a subset has been returned, we ask whether an allowed exchange can improve the target beyond a given tolerance after retraining. This post hoc question compares subsets within a declared exchange neighbourhood under the same training rule and data budget.

A local score describes the response to an infinitesimal change in training weight. As weight moves from the returned subset toward an alternative, the retrained target utility traces a curve, and the score gives only its slope at the start. Certification requires an upper bound on the gain at the endpoint. The number of alternative subsets makes retraining and comparing them individually impractical. We therefore seek conditions under which local scores at the returned subset's trained model can bound these gains over the whole neighbourhood.

GLISTER and SELCON give guarantees tied to their selection objectives and searches (Killamsetty et al., 2021; Sivasubramanian et al., 2021). These guarantees do not directly provide post hoc certificates for arbitrary returned subsets. Influence functions and the infinitesimal jackknife provide error bounds for approximating retraining under finite data changes (Giordano et al., 2019b; Broderick et al., 2020; Koh et al., 2019). Uniform error bounds already support decisions, and specialized regression audits provide certificates without enumeration (Freund & Hopkins, 2023; Rubinstein & Hopkins, 2025). The direction of the error matters for how informative such a certificate can be. Overpredicting an alternative's gain preserves an upper bound's validity, while a symmetric correction can charge for this error. Only gains above the local prediction need correction. Finite exchanges change both the model and training curvature, so even a convex target loss cannot rule out such gains.

Our key insight is that target curvature can suppress gains above the local prediction. For concave retrained utility, the starting tangent already bounds the curve and no correction is needed. Beyond concavity, we decompose curvature into the target's curvature, the variation of training curvature as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the model moves, and curvature differences between exchanged examples. When target curvature outweighs training curvature variation, the remaining quadratic suppression absorbs the effect of example differences, which grows at most linearly with the model response. A certified response cap provides control in the remaining cases. This yields a one-sided correction without requiring accurate endpoint predictions or concave retrained utility.

To make this control uniform, we verify a region containing every relevant retrained solution and derive shared curvature and response bounds within it. Combining these bounds with local scores recomputed at the returned subset's trained model bounds the maximum gain over the declared neighbourhood. The computation covers all allowed exchanges without retraining or enumerating alternatives, regardless of which selection procedure returned the subset. The procedure issues a certificate when the required conditions are verified and the gain upper bound does not exceed the tolerance. Otherwise, it withholds certification.

1. **Certification with nonzero prediction error.** We derive a one-sided curvature correction that separates target curvature, training curvature variation and curvature differences between exchanged examples. This correction can vanish on instances where every valid uniform two-sided error bound for the same local predictor is positive.
2. **Computable guarantees over an exchange neighbourhood.** We give regional verification conditions that turn pathwise control into one shared computation over the declared neighbourhood. The resulting upper bound applies to any returned subset for which these conditions hold.
3. **Decision guarantees under estimation and training errors.** We combine the geometric bound with controls for target sampling, numerical computation and deviations from the reference training rule. These controls specify when to certify a subset, confirm an improving alternative, or abstain.

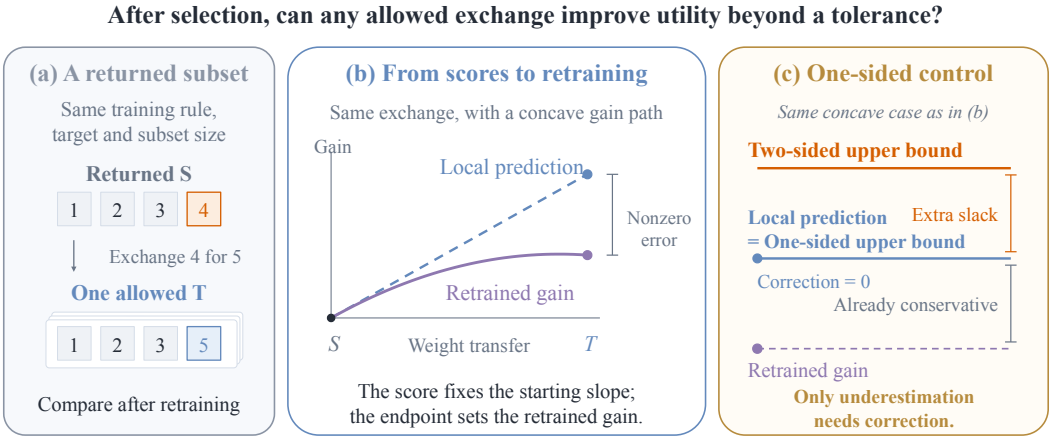

Figure 1: Post hoc certification and the one-sided viewpoint. **(a)** An allowed exchange preserves the subset size, training rule, and target. **(b)** Scores recomputed at the returned subset's model give the starting slope along the weight-transfer path; the endpoint gives the retrained gain. **(c)** In this concave example, the prediction is a valid one-sided upper bound with zero correction. A symmetric absolute-error correction instead raises the upper bound, charging for a conservative overestimate. Figure 2 shows how verification covers the whole declared family. All curves are schematic.

2 Background and Motivation

2.1 Background

Scoring signals and reference models. Influence functions describe how a fitted model and target respond to infinitesimal example reweighting (Koh & Liang, 2017); LESS uses optimizer-aware gradient similarity to match instruction data to target examples (Xia et al., 2024). Target-

Aligned Paths scores normalized endpoint-loss drops along target-induced paths; BLADE updates the influence objective's reference model during selection (Yang et al., 2026; Wang et al., 2026). CHIPS adds curvature-aware Newton-style alignment to CLIP selection scores and analyses its proxy approximation (Zhuang et al., 2026). These methods vary in scoring signal and reference point; comparing subsets also requires retrained outcomes under joint data changes.

Finite data changes and retraining approximations. Group influence and deletion studies examine interactions among jointly changed examples (Koh et al., 2019; Hu et al., 2024; Huang et al., 2025); higher-order and interaction methods approximate those effects (Basu et al., 2020; Heo et al., 2026). Infinitesimal jackknife analyses give finite-sample error bounds for reweighted fits, including expansions of arbitrary order under regularity conditions (Giordano et al., 2019b;a). AMIP searches influential deletion sets and bounds approximation error uniformly over the declared changes (Broderick et al., 2020). Such uniform bounds already support decisions through an approximation. Self-concordant influence analysis supplies statistical and computational controls using local Hessian geometry (Fisher et al., 2023).

Selection guarantees and post hoc verification. GLISTER and SELCON derive optimization guarantees from their objectives' structure (Killamsetty et al., 2021; Sivasubramanian et al., 2021); Hanneke et al. (2025) study fixed learning rules under a data budget. Linear regression audits study worst-case deletion effects and include non-enumerative certificates (Moitra & Rohatgi, 2023; Freund & Hopkins, 2023; Rubinstein & Hopkins, 2025); Karchmer et al. (2025) interactively verify predictions of attribution models. Abstract Gradient Training bounds parameter sets reachable under data perturbations to certify first-order training (Sosnin et al., 2026). For concave retrained utility, the Frank-Wolfe gap supplies the relevant regret bound (Jaggi, 2013). Our analysis extends the curvature control to heterogeneous training Hessians by retaining target curvature's suppression of missed gains and verifying shared constants over the exchange neighbourhood.

2.2 Motivation

Given a returned subset and its trained model, we seek an upper bound on retrained gains over a declared exchange neighbourhood. The correction combines target curvature, variation of training curvature, and differences between example Hessians. Its one-sided form retains favorable curvature while controlling gains above the local prediction. Regional verification makes this control simultaneous over all exchange paths, allowing a shared computation without alternative retraining. Table 9 compares the guarantee objects and required inputs of the related tools.

3 Selection Decisions after Retraining

We certify that no allowed replacement improves a returned subset's target utility beyond tolerance under the same training rule.

3.1 Decision target and exchange neighbourhood

Let $[N] = \{1, \dots, N\}$ index candidate examples and $1 \le k \le N$ be the selection budget. The capped simplex

$$\mathcal{W}_k = \left\{ w \in \mathbb{R}_+^N : \mathbf{1}^\top w = 1, \quad w_i \le \frac{1}{k} \right\} \tag{1}$$

is the convex hull of uniform size-k subset weights $w_S = k^{-1}\mathbf{1}_S$. Define the weighted training problem and utility by

$$L(\theta, w) = \sum_{i=1}^{N} w_i \ell_i(\theta) + r_0(\theta), \qquad \theta(w) = \arg\min_{\theta} L(\theta, w), \qquad F(w) = -R(\theta(w)), \tag{2}$$

Here R is the target loss and $\theta(w)$ a single-valued exact minimizer. Path derivatives require Assumption 4.1, including $H(w) \succ 0$ along the chords; Section 5.3 covers finite updates. For size-k subsets S, T, the exchange radius $q = |T \setminus S|$ gives

$$\|w_T - w_S\|_1 = \frac{2q}{k}, \qquad q \le q_{\max} = \min(k, N-k). \tag{3}$$

For a returned subset S, define its size-k selection regret by

$$\mathrm{Reg}_k(S) = \max_{T:\,|T|=k} \{F(w_T) - F(w_S)\}. \tag{4}$$

Restricting to $|T \setminus S| \le Q$ defines $\mathrm{Reg}_{k,\le Q}(S)$ for $1 \le Q \le q_{\max}$; $Q = q_{\max}$ recovers $\mathrm{Reg}_k(S)$. A declared family $\mathcal{C} \ni S$ similarly restricts regret. An upper bound at most ϵ certifies the declared regret; a positive gain lower bound establishes improvement.

3.2 Local scores and the endpoint identity

Regardless of the selector, recompute scores at $\theta_S = \theta(w_S)$, with $H_S = \nabla^2_\theta L(\theta_S, w_S)$, as

$$h_i(S) = \nabla \ell_i(\theta_S)^\top H_S^{-1} \nabla R(\theta_S), \qquad \Gamma_q(S) = \max_{|T|=k,\,|T\setminus S|=q} h(S)^\top (w_T - w_S). \tag{5}$$

Sorting the q largest outside and q smallest inside scores gives $\Gamma_q(S)$. The weight path $f_{S,T}(t) = F(w_S + t(w_T - w_S))$ follows retrained solutions and differs from an optimizer trajectory. Its initial slope is $h(S)^\top (w_T - w_S)$. Assumption 4.1 gives

$$\underbrace{f_{S,T}(1) - f_{S,T}(0)}_{\text{retrained gain}} = \underbrace{f'_{S,T}(0)}_{\text{local prediction}} + \underbrace{\int_0^1 (1-t) f''_{S,T}(t)\, dt}_{\text{accumulated curvature}}. \tag{6}$$

Section 4 bounds the curvature integral from above uniformly over allowed paths to certify the neighbourhood.

4 From Local Scores to a Shared Certificate

4.1 Zero correction under concavity

Under concavity, the curvature integral in Equation (6) is nonpositive, so $\max\{0, \max_q \Gamma_q(S)\}$ bounds regret with zero correction. This is the Frank-Wolfe gap of $-F$ over $\mathcal{W}_k$ (Jaggi, 2013), with Γ_q its radius-q component. Quadratic training with a shared Hessian and convex R gives affine $\theta(w)$ and concave F (Corollary A.2).

For gain Δ_T and local prediction ℓ_T, two-sided error control bounds $|\Delta_T - \ell_T|$ (Broderick et al., 2020), whereas certification needs only $\rho_T \ge \Delta_T - \ell_T$ (Proposition B.5; Figure 1(b,c)). Remark B.6 gives an exact instance where zero correction certifies zero regret while the smallest symmetric error radius yields a positive bound.

4.2 Curvature correction and the main certificate

To bound the curvature integral beyond concavity, write, for $\mathbf{1}^\top d = 0$ at $\theta = \theta(w)$,

$$H = \nabla^2_\theta L(\theta, w), \qquad q_d = \sum_i d_i \nabla \ell_i(\theta), \qquad u_d = H^{-1} q_d,$$
$$H_d = \sum_i d_i \nabla^2 \ell_i(\theta), \qquad \mathcal{T} = \nabla^3_\theta L(\theta, w), \qquad H_R = \nabla^2 R(\theta), \tag{7}$$

where q_d is the exchange gradient change and u_d its first-order parameter response. Differentiating stationarity twice (Theorem A.1) gives, with $r = \nabla R(\theta(w))$ and $v = H^{-1} r$,

$$D^2 F(w)[d,d] = -u_d^\top H_R u_d + \langle v, \mathcal{T}[u_d, u_d] - 2 H_d u_d \rangle. \tag{8}$$

For convex R, target curvature suppresses gains quadratically, training variation has either sign, and exchange Hessian differences contribute at most linearly (Figure 2a). Define response energy E_H, target size a_H, net curvature margin μ_H and exchange strength β_H by

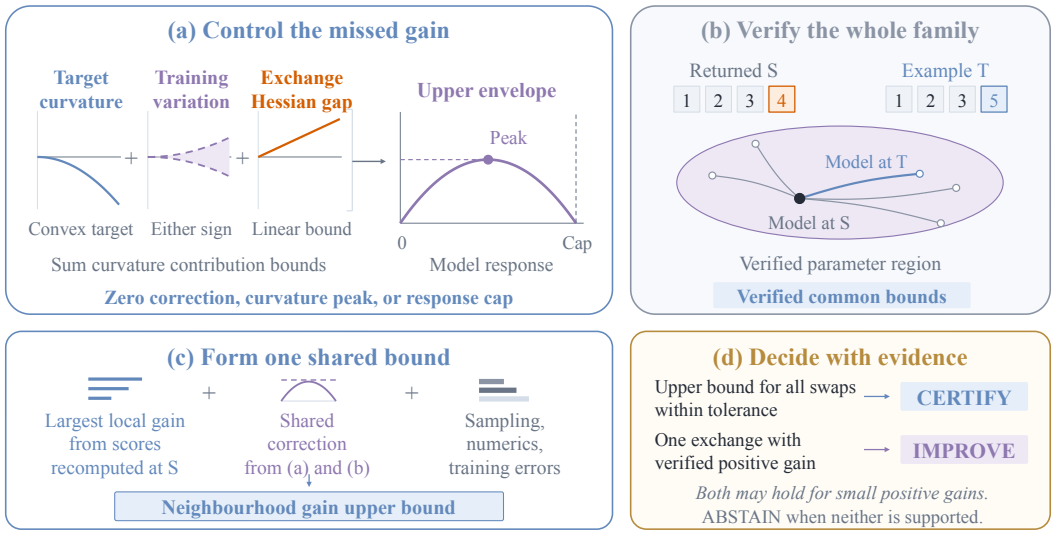

Figure 2: From pathwise control to a shared certificate. **(a)** Bounds on the three signed curvature contributions give a scalar upper envelope; concavity, an interior peak, or a verified response cap controls its maximum. **(b)** A verified parameter region contains all allowed solution paths and supplies common bounds. **(c)** The largest local gain, shared correction, and jointly valid error allowances form the neighbourhood upper bound. **(d)** An upper bound within tolerance supports CERTIFY. A separately verified positive lower bound on one replacement supports IMPROVE. Both checks may hold for small positive gains; abstain when neither is supported. All curves are schematic.

$$E_H = u_d^\top H u_d = q_d^\top H^{-1} q_d, \qquad a_H^2 = r^\top H^{-1} r,$$
$$\mu_H = \frac{u_d^\top H_R u_d - \langle v, \mathcal{T}[u_d, u_d]\rangle}{E_H}, \qquad \beta_H = \frac{2|\langle v, H_d u_d\rangle|}{\|d\|_1 \sqrt{E_H}}. \tag{9}$$

Ratios require $E_H > 0$; all quantities are invariant under invertible affine reparameterization (Proposition B.1). With normalized response $z = \sqrt{E_H}/\|d\|_1$,

$$D^2 F(w)[d,d] \le \|d\|_1^2 \left(-\mu_H z^2 + \beta_H z\right). \tag{10}$$

A shared certificate requires bounds on these varying quantities valid along every allowed path at each radius.

Assumption 4.1 (Certified path geometry)**.** On a relative-open neighborhood of the declared comparator chords, L is C^3, R is C^2, the solution map $\theta(w)$ is single-valued and $H(w) \succ 0$. For every feasible q-swap chord and every $t \in [0,1]$ with $E_H > 0$, the declared constants satisfy $\mu_H \ge m_q$, $\beta_H \le \beta_q < \infty$ and $\sqrt{E_H} \le \bar{z}_q \|d\|_1$; no sign is required of m_q. A radius whose chords all have $E_H = 0$ carries no penalty.

Theorem 4.2 (One-sided selection certificate)**.** *Under Assumption 4.1 at every radius $1 \le q \le q_{\max}$, the size-k selection regret of S obeys*

$$\boxed{\mathrm{Reg}_k(S) \le \max\left\{0, \max_q \left[\Gamma_q(S) + 2P_q\left(\frac{q}{k}\right)^2\right]\right\},} \qquad P_q = \max_{0 \le z \le \bar{z}_q} \{-m_q z^2 + \beta_q z\}. \tag{11}$$

The bound is invariant under invertible affine reparameterization, requires no concavity of F, and retrains no alternative (Theorem B.4). For $\mathrm{Reg}_{k,\le Q}(S)$, the same bound takes the maximum over $1 \le q \le Q$ and requires the path conditions only for those radii.

Integration gives one correction per radius; sorting maximizes local predictions to obtain $\Gamma_q(S)$, and maximizing over radii bounds neighbourhood regret.

Evaluating the envelope. For $m_q > 0$ and $\beta_q \le 2m_q\bar{z}_q$, the peak at $\beta_q/(2m_q)$ gives $P_q = \beta_q^2/(4m_q)$, independent of response magnitude. Otherwise $P_q = \beta_q\bar{z}_q - m_q\bar{z}_q^2$. The gradient diameter gives $\sqrt{E_H} \le \frac{1}{2}\|d\|_1 \max_{i,j}\|\nabla\ell_i - \nabla\ell_j\|_{H^{-1}}$ (Lemmas B.3 and C.17).

4.3 Computing the certificate from a verified region

A region containing all allowed solution paths supplies the shared bounds in Assumption 4.1.

Region verification. Choose a region around θ_S and verify the boundary condition of Theorem C.10 using anchor gradients, Hessian diameters and relative third-order bounds. Under convexity and coercivity, this places all minimizers, including intermediate path solutions, inside the region. Hessian comparison certifies positive definiteness throughout.

Shared constants. Regional derivative bounds give

$$\beta_H \le a_H\zeta_H, \qquad \mu_H \ge \lambda_{\min}(H_R, H) - Ma_H,$$
$$\zeta_H = \max_{i,j}\left\|H^{-1/2}\big(\nabla^2\ell_i(\theta) - \nabla^2\ell_j(\theta)\big)H^{-1/2}\right\|_{\mathrm{op}}. \tag{12}$$

Here $\lambda_{\min}(H_R, H) = \inf_{u\neq 0}(u^\top H_R u)/(u^\top H u)$ for $H \succ 0$, and M bounds relative third-order variation (Corollary B.10). Theorem C.12 converts uniform regional bounds into m_q, β_q and $\bar{z}_q = \bar{\gamma}_q/2$, where $\bar{\gamma}_q$ bounds the gradient diameter along all paths.

Shared path certificate for a returned subset S. (1) Fit θ_S and factor H_S for inverse actions. **(2)** Compute scores, a_H, $\lambda_{\min}(H_R, H_S)$ and the diameters in Equation (158) with $O(N)$ certified norm calls. **(3)** Verify containment at each radius. **(4)** Bound m_q, β_q, $\bar{z}_q$ and P_q on the verified tube. **(5)** Sort scores for $\Gamma_q(S)$ and return Equation (11).

The $O(N)$ count covers norm calls; exact diameters use $O(N^2)$ pairs. Fitting, factorization and enclosures depend on dimension. Failed containment sets $U_q = +\infty$ (Section 5.1); nonpositive margins invoke the response cap. Theorem C.5 gives hardness at specified additive precisions.

Corollary C.20 gives sufficient conditions for nonvacuous ridge-softmax certificates; experiments verify their own regions.

Endpoint refinement. In fixed coordinates, endpoint stationarity gives another gain upper bound using a chosen response p, retaining matrix curvature and charging its residual (Proposition I.5). Sorted support, regional bounds and numerical checks must be verified (Appendix I) before use in Section 5.1. Section 6.2 compares this extension with the fixed response.

5 Certification with Estimation and Training Errors

5.1 The decision rule

Let $\widehat{\Gamma}_q$ be the largest estimated local gain, ε_q a joint sampling and numerical score error radius, $D_q^+ = 2P_q(q/k)^2$ the one-sided correction, $D_q^\pm$ a two-sided remainder, and Δ the training discrepancy (Section 5.2). Define

$$U_q = \widehat{\Gamma}_q + \varepsilon_q + D_q^+ + 2\Delta, \qquad L_q = \widehat{\Gamma}_q - \varepsilon_q - D_q^\pm - 2\Delta. \tag{13}$$

Under jointly valid inputs, Theorem B.18 bounds all radius-q gains by U_q; L_q lower-bounds the gain of the replacement maximizing the estimated local prediction. Verified witnesses may supply separate lower bounds. Report Certify if $\max\{0, \max_q U_q\} \le \epsilon$ and Improve if a replacement has a positive lower bound. Both may hold for positive gains within tolerance; abstain if neither holds. Missing controls set $U_q = +\infty$ or $L_q = -\infty$ (Appendix B.5).

Jointly valid upper bounds for the same training rule, target and family may be minimized at each q, including Proposition I.5, charging each required allowance once. Endpoint experiments use fixed empirical targets; population claims require joint control of all estimates used to choose and verify the response.

5.2 Estimation and implementation errors

Target samples. For n target rows independent of selection and training, Theorem A.3 controls target-gradient error projected onto feasible responses under centered sub-Gaussian errors, up to a universal constant (Appendix B.7). Curvature and geometry need separate envelopes; target reuse requires uniform control over induced selections (Appendix B.6).

Numerical computation. Verified residual bounds for the score linear solves contribute to ε_q (Lemma B.12); regional bounds require valid numerical enclosures.

Deployed training. For a fixed family $\mathcal{C}$, compare smooth surrogate utility $\widetilde{F}$ with expected utility F_{real} under deployed, possibly randomized training through

$$\Delta_{\mathrm{cpl}}(\mathcal{C}) = \sup_{T \in \mathcal{C}} |F_{\mathrm{real}}(T) - \widetilde{F}(w_T)|. \tag{14}$$

Each endpoint contributes an error, so $\Delta \geq \Delta_{\mathrm{cpl}}(\mathcal{C})$ gives the 2Δ allowance (Theorem B.14). For a fixed finite family, Corollary B.16 bounds this discrepancy using independent paired repetitions for every member and a valid sub-Gaussian scale.

5.3 Finite training and comparator-wise bounds

Finite updates change the training map; comparator-wise bounds change family coverage. Both supply inputs to Section 5.1.

Finite update maps. Let z_t contain parameters and optimizer state, with $z_{t+1} = G_t(z_t, w)$ and $F_K(w) = -R(z_K(w))$ at fixed horizon K. Differentiating smooth updates propagates the weight response without requiring an exact stationary endpoint or positive definite training Hessian. Uniform derivative bounds on updates and target over a region containing all declared trajectories bound $|D^2 F_K(w)[d, d]|$ (Theorem B.17). Integration supplies a two-sided remainder for the local derivative of F_K. The bound depends on the norm and horizon; nonsmooth components require Equation (14).

Comparator-wise signed bounds. For a small declared family, Proposition G.1 encloses each curvature integral in Equation (6) as $[J - r, J + r]$ using midpoint derivatives and a cellwise fourth-derivative bound. The gain upper bound $f'(0) + J + r$ improves on the symmetric $f'(0) + |J| + r$ when $J < 0$. Each comparator needs its own path computation. Add the required error allowances, then maximize these bounds and zero for a family upper bound. All estimation, solver and path controls must hold jointly (Appendix G).

6 Experimental Evaluation

We test validity, coverage, endpoint refinement and cost, followed by decision error budgets and prediction on modern models. Tables 7 and 12 and Appendix K.1 specify each training rule, target and comparison family.

6.1 Validity against exhaustive retraining

Table 1: Bounds against exhaustive retraining over all $1{,}835{,}008$ one-swap alternatives of one CIFAR-10 anchor.

Quantity	Best gain or bound	Ratio to truth
best gain after retraining every alternative	7.6465×10^{-5}	1
largest local prediction	8.7399×10^{-5}	1.143
shared analytic certificate	8.7409×10^{-5}	1.143
outward-rounded verification of the certificate	8.7948×10^{-5}	1.150
strict second-order certificate	7.8527×10^{-5}	1.027

Coverage and tightness. For one size-512 CIFAR-10 anchor (Krizhevsky, 2009), all 1,835,008 one-swap alternatives are retrained with pinned ResNet-18 features and a ridge-softmax head. The local maximum exceeds the best gain by 14.3%; the analytic correction adds 0.011% and outward rounding adds 0.63% (Table 1). All 337 prospective endpoint audits satisfy their committed bounds (Appendix D.5).

Cost and comparison scope. Analytic construction takes 1.7 seconds, exhaustive retraining 283 seconds on two GPUs, and outward verification 67.21 seconds on one CPU thread. Matched CPU runs give strict second order 1.20 times the analytic cost and tighter bounds on all ten seeds. HOIJ2's mean prediction error is 2.95×10^{-8}; its certified scope is 160 fixed witnesses, while shared geometry covers the neighborhood (Appendices H.4 and J.3).

6.2 Returned subsets, radius and tolerance

Certification across returned subsets. E4 audits ten size-512 subsets for a 17-parameter detection head on Qwen features under one fixed empirical target. Exhaustive retraining of each subset's 262,144 one-swap alternatives gives float64 gains with Arb upper/gain ratios of 1.13 to 1.35. At $Q = 4$ and illustrative tolerance 0.005, Arb bounds certify all five iterative subsets; symmetric and convex tangent baselines certify none (Table 20, Figure 7).

How far the certificate extends. The retrospective CIFAR envelope certifies all 160 cells (ten subsets, two targets, $Q = 1, \ldots, 8$) at 0.02 under conditional float64 evaluation. For the regularized target, the correction share grows from 0.02% to 65%; the cap raises two-swap coverage from $7/10$ to $10/10$ (Figure 3(a)). Unregularized cross-entropy has negative curvature margins. At $Q = 8$, all stability bounds exceed 0.02; the largest envelope is 7.63×10^{-3}.

Tightness and confirmation on a fresh target. At $Q = 8$, feasible CIFAR CE witnesses reach 95.57% to 96.05% of the conditional directional endpoint bounds across all ten anchors. A separate detection confirmation freezes an anchor and uses 145 previously unused target answers. At tolerance 0.02, its fixed response endpoint bound certifies four exchanges, versus two for the symmetric and one for the convex target bound (Appendix J.2).

Endpoint refinement over complete four-exchange neighborhoods. We compare Proposition I.5 with a fixed response matrix baseline on twelve detection anchors from four seeds, including two new seeds confirming the frozen method on the same 512 empirical target answers. All 512 selected and 512 outside candidates participate. At $Q = 4$, verified upper bounds divided by strict witness lower bounds give ratios of 1.421 to 1.564, tighter at every anchor. These ratios bound the ratios to true regret from above. Three runs per anchor give median construction costs of 114.90 seconds for the baseline and 144.35 seconds including the extension, with median paired increase 25.3%. These exclude anchor fitting and witness search (Appendix K).

6.3 Error budgets and abstention

Target evidence. For fixed HelpSteer3 families, diagnostics abstain through 448 target rows; at 896, all ten families certify tolerance 0.02 and none confirms improvement. The support range term supplies a median 80% of the sampling radius. The separate 4096-row confirmation supports both decisions on all ten families; its median sampling and quadrature allowances for selected over base are 0.011329 and 0.000455 (Appendices F.2 and H.3).

Training error. Worst-case HelpSteer3 coupling yields ten abstentions through 304 steps, ten certificates without improvement at 336, and both decisions at 384 (Figure 3(c)). For nonconvex heads, direct enclosures of four GD steps (56 parameters) and eight momentum steps (596 parameters) certify all ten declared families per study at 0.02, testing the finite update route of Section 5.3. Their descriptive certification thresholds are 0.00242 and 0.00851 (Figure 5). Sampling dominates the GD path widths; momentum has visible path and sampling contributions.

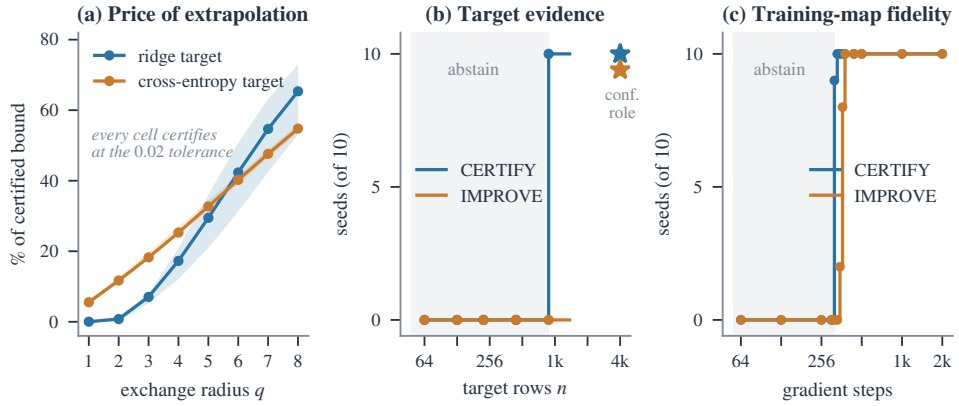

Figure 3: Contributions to the certification budget. **(a)** Retrospective correction share across ten frozen CIFAR subsets; bands span the ten. **(b)** HelpSteer3 target sampling diagnostics; the 4096-row confirmation uses a separate role. **(c)** Decisions after adding worst-case coupling to finite training.

6.4 Retraining audits and prediction on modern models

Observed gains after retraining. We audit Qwen3 (8B) and OLMo3 (7B) with LESS-style and random selectors, each returning 256 of 1,024 examples for 64 AdamW updates. Retraining 122 declared exchanges per selector and model finds a positive audit NLL gain in each of the four arms (Table 2). The comparisons retain the original initialization and training rule (Appendix J.1).

Prediction with midpoint feedback. On twelve preselected exchanges through the same 64 updates, midpoint feedback reduces MAE relative to affine feedback by 95.6% on Qwen3 and 91.2% on OLMo3. Both meet the predeclared 75% requirement; all twelve gain signs are correct (Appendix J.1).

Table 2: Modern retraining audits and empirical prediction accuracy. Gains and MAEs are in units of 10^{-3} audit NLL. Each gain is the best of 122 retrainings per arm; each MAE averages six fixed exchanges per model. Reduction is relative to affine MAE.

	Best observed gain		Prediction on fixed cases		
Model	LESS	Random	Affine MAE	Midpoint MAE	Reduction
Qwen3	4.491	8.586	0.5373	**0.02354**	95.6%
OLMo3	1.224	3.174	0.2039	**0.01797**	91.2%

7 Conclusion

We asked when scores computed at one trained model certify that no alternative improves on the returned subset beyond a stated tolerance once the alternative is retrained. Certification needs no accurate retraining prediction, but it must control gains above the local prediction while retaining the suppression from favorable curvature. When target curvature outweighs training-curvature variation, the remaining quadratic suppression absorbs the at most linear effect of example differences; otherwise a certified response cap provides control. Combining local scores at the trained model with a verified region, the certificate covers the declared exchange neighbourhood in one shared computation, whichever selector produced the subset, without retraining alternatives and with abstention when conditions cannot be verified. On an image instance the resulting certificates hold over neighbourhoods far too large to retrain, adding little to the local prediction for small exchanges and more as they widen, and on language instances the error budget still adds up under limited data and finite-step training.

AI USE STATEMENT

We used generative AI tools to assist with mathematical derivations, proof drafting and checking, literature search, experimental planning, method implementation, and result interpretation. We also used these tools for translation, manuscript drafting and editing, and figure preparation. The authors checked the statements, assumptions, derivations, and proof obligations and take responsibility for the final text, claims, results, and artifacts.

ETHICS STATEMENT

The study uses declared target losses and disjoint data roles. Its decision ledger makes target reuse and every empirical claim boundary explicit.

REPRODUCIBILITY STATEMENT

Assumptions and proofs are given in the main text and appendices. Appendices D, E, F, and J document data roles, representations, training rules, comparison families, and evaluation protocols. Appendix H records decision thresholds and error budgets. Appendix K gives the seed configurations, numerical verification procedures, and repeated timing protocol for the complete exchange experiments.

REFERENCES

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

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 715–724. PMLR, 2020. URL https://proceedings.mlr.press/v119/basu20b.html.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: When can dropping a little data make a big difference?, 2020. URL https://arxiv.org/abs/2011.14999.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29, 2015. doi: 10.1214/EJP.v20-3760. URL https://projecteuclid.org/journals/electronic-journal-of-probability/volume-20/issue-none/Tail-bounds-via-generic-chaining/10.1214/EJP.v20-3760.full.

Jillian Fisher, Lang Liu, Krishna Pillutla, Yejin Choi, and Zaid Harchaoui. Influence diagnostics under self-concordance. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10028–10076. PMLR, 2023. URL https://proceedings.mlr.press/v206/fisher23a.html.

Daniel Freund and Samuel B. Hopkins. Towards practical robustness auditing for linear regression, 2023. URL https://arxiv.org/abs/2307.16315.

Ryan Giordano, Michael I. Jordan, and Tamara Broderick. A higher-order Swiss army infinitesimal jackknife, 2019a. URL https://arxiv.org/abs/1907.12116.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A Swiss army infinitesimal jackknife. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1139–1147. PMLR, 2019b. URL https://proceedings.mlr.press/v89/giordano19a.html.

Steve Hanneke, Shay Moran, Alexander Shlimovich, and Amir Yehudayoff. Data selection for ERMs. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2634–2665. PMLR, 2025. URL https://proceedings.mlr.press/v291/hanneke25a.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.

Jaeseung Heo, Kyeongheung Yun, Youngbin Choi, Sehyun Hwang, Jungseul Ok, and Dongwoo Kim. Interaction-aware influence functions for group attribution, 2026. URL https://arxiv.org/abs/2605.15675.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuzheng Hu, Pingbang Hu, Han Zhao, and Jiaqi W. Ma. Most influential subset selection: Challenges, promises, and beyond. In *Advances in Neural Information Processing Systems 37*, 2024. URL https://openreview.net/forum?id=qWi33pPecC.

Jenny Y. Huang, David R. Burt, Yunyi Shen, Tin D. Nguyen, and Tamara Broderick. Approximations to worst-case data dropping: Unmasking failure modes. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=m6EQ6YdPXV.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435. PMLR, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Ari Karchmer, Seth Neel, and Martin Pawelczyk. Efficiently verifiable proofs of data attribution. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ffbfecfb9230f70a2938b1ec5502155c-Abstract-Conference.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, 2021. doi: 10.1609/aaai.v35i9.16988. URL https://ojs.aaai.org/index.php/AAAI/article/view/16988.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a78482ce76496fcf49085f2190e675b4-Abstract.html.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Ankur Moitra and Dhruv Rohatgi. Provably auditing ordinary least squares in low dimensions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DlpCotqdTy.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Ittai Rubinstein and Samuel B. Hopkins. Robustness auditing for linear regression: To singularity and beyond. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07916.

Durga Sivasubramanian, Rishabh Iyer, Ganesh Ramakrishnan, and Abir De. Training data subset selection for regression with controlled generalization error. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9202–9212. PMLR, 2021. URL https://proceedings.mlr.press/v139/s21a.html.

Philip Sosnin, Matthew Wicker, Josh Collyer, and Calvin Tsay. Abstract gradient training: A unified certification framework for data poisoning, unlearning, and differential privacy. *Journal of Machine Learning Research*, 27(128):1–58, 2026. URL https://www.jmlr.org/papers/v27/25-2206.html.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

Jiaxing Wang, Deping Xiang, Jin Xu, Zirui Liu, Zicheng Zhang, Guoqiang Gong, Jun Fang, Chao Liu, Pengzhang Liu, Tongxuan Liu, Ke Zhang, and Qixia Jiang. BLADE: Scalable bi-level adaptive data selection for LLM training, 2026. URL https://arxiv.org/abs/2606.18650.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/3e0271cf7df2cdb3b91565ad1f525f3a-Abstract-Datasets_and_Benchmarks_Track.html.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Huitao Yang, Hengzhi He, and Guang Cheng. Let the target select for itself: Data selection via target-aligned paths, 2026. URL https://arxiv.org/abs/2605.09404.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Xinlin Zhuang, Yichen Li, Xiwei Liu, Haolin Yang, Yifan Lu, Ziyun Zou, Yulong Li, Huifa Li, Dongliang Chen, Qinglei Wang, Weiyang Liu, Ying Qian, Jiangming Shi, and Imran Razzak. CHIPS: Efficient CLIP adaptation via curvature-aware hybrid influence-based data selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29483–29493, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhuang_CHIPS_Efficient_CLIP_Adaptation_via_Curvature-aware_Hybrid_Influence-based_Data_Selection_CVPR_2026_paper.html.

A Core setup, exact calculus, and fixed-index target evidence

Fix $N \geq 2$ and $1 \leq k \leq N$. For a size-k subset $S \subseteq [N]$, write $w_S = k^{-1}\mathbf{1}_S$, set $q_{\max} = \min\{k, N-k\}$, and define

$$\mathcal{W}_k = \{w \in \mathbb{R}^N_+ : \mathbf{1}^\top w = 1,\ w_i \leq 1/k\}, \qquad L(\theta, w) = r_0(\theta) + \sum_{i=1}^{N} w_i \ell_i(\theta). \tag{15}$$

Let $\mathcal{D}$ be a relatively open subset of the affine hyperplane $\{w : \mathbf{1}^\top w = 1\}$ containing every chord under study. Assume $\ell_i, r_0 \in C^3$, $R \in C^2$, and that a fixed C^2 branch satisfies

$$\nabla_\theta L(\theta(w), w) = 0, \qquad H(w) = \nabla^2_\theta L(\theta(w), w) \succ 0 \tag{16}$$

throughout $\mathcal{D}$. Set $F(w) = -R(\theta(w))$. If $\mathcal{D}$ contains every size-k vertex, define

$$\mathrm{Reg}_k(S) = \max_{T:\,|T|=k} \{F(w_T) - F(w_S)\}. \tag{17}$$

All weight derivatives below are relative to the simplex affine hull. A boundary vertex therefore requires the declared branch to extend to a relatively open neighborhood of each certified chord.

For tangent directions d, e, define

$$q_d = \sum_i d_i \nabla \ell_i(\theta), \quad H_d = \sum_i d_i \nabla^2 \ell_i(\theta), \quad u_d = H^{-1} q_d, \tag{18}$$

and analogously for e. Let $\mathcal{T} = \nabla^3_\theta L(\theta(w), w)$ and $H_R = \nabla^2 R(\theta(w))$.

Theorem A.1 (Exact data-weight derivatives)**.** *For tangent d, e,*

$$D\theta(w)[d] = -u_d, \tag{19}$$

$$D^2\theta(w)[d,e] = H^{-1}(H_d u_e + H_e u_d - \mathcal{T}[u_d, u_e])\,, \tag{20}$$

$$DF(w)[d] = \nabla R(\theta)^\top u_d, \tag{21}$$

$$D^2F(w)[d,e] = -u_d^\top H_R u_e + \left\langle H^{-1}\nabla R, \mathcal{T}[u_d, u_e] - H_d u_e - H_e u_d \right\rangle. \tag{22}$$

In particular,

$$D^2F(w)[d,d] = -u_d^\top H_R u_d + \left\langle H^{-1}\nabla R, \mathcal{T}[u_d, u_d] - 2H_d u_d \right\rangle. \tag{23}$$

Proof. Set $G(\theta, w) = \nabla_\theta L(\theta, w)$. Differentiating $G(\theta(w), w) = 0$ in direction d gives $HD\theta[d] + q_d = 0$, proving (19). Differentiating that equation in direction e gives

$$HD^2\theta[d,e] + \mathcal{T}[u_d, u_e] - H_e u_d - H_d u_e = 0, \tag{24}$$

which proves (20). Finally, $F = -R \circ \theta$ gives

$$DF[d] = -\nabla R^\top D\theta[d], \tag{25}$$

$$D^2F[d,e] = -D\theta[d]^\top H_R D\theta[e] - \nabla R^\top D^2\theta[d,e]. \tag{26}$$

Substitution and symmetry of H prove the remaining identities. □

Corollary A.2 (Common-quadratic boundary case)**.** *Suppose on $\mathcal{W}_k$ that*

$$L(\theta, w) = \tfrac{1}{2}\theta^\top H\theta - b(w)^\top\theta + c(w), \qquad H \succ 0, \tag{27}$$

where b is affine, and suppose R is convex. With $h = \nabla F(w_S)$,

$$\mathrm{Reg}_k(S) \le \max_{v \in \mathcal{W}_k} h^\top(v - w_S) = \frac{1}{k}\sum_{i \in \mathrm{Top}_k(h)} h_i - \frac{1}{k}\sum_{i \in S} h_i. \tag{28}$$

Thus a zero gap certifies global size-k optimality.

Proof. Here $\theta(w) = H^{-1}b(w)$ is affine. Since R is convex, $F = -R \circ \theta$ is concave, so $F(v) - F(w_S) \le h^\top(v - w_S)$. Linear optimization over $\mathcal{W}_k$ places mass $1/k$ on the k largest coordinates of h. □

A.1 An honest fixed-subset target process

Let $\mathscr{A}$ contain every object observed before certification, including the selected subset S, training pool, fitted parameters, Hessian, gradients, and training randomness. Write $\theta_S = \theta(w_S)$ and $H_S = H(w_S)$. Let $Z_1, \ldots, Z_n$ be conditionally independent target rows given $\mathscr{A}$, and suppose

$$\widehat{r}_S = r_S + \bar{\xi}, \qquad r_S = \nabla R(\theta_S), \qquad \bar{\xi} = \frac{1}{n}\sum_{j=1}^n \xi_j, \tag{29}$$

where the ξ_j are conditionally centered and, for a deterministic or $\mathscr{A}$-measurable positive semidefinite Σ,

$$\log \mathbb{E}[\exp(\lambda z^\top \xi_j) \mid \mathscr{A}] \le \frac{\lambda^2}{2} z^\top \Sigma z \tag{30}$$

for every z and λ. For a q-swap comparator T, define

$$\begin{aligned} x_T &= H_S^{-1} \sum_i (w_T - w_S)_i \nabla \ell_i(\theta_S), \\ \mathcal{X}_q(S) &= \{x_T : |T| = k,\ |T \setminus S| = q\}, \\ \mathcal{X}_q^{\pm}(S) &= \{0\} \cup \mathcal{X}_q(S) \cup [-\mathcal{X}_q(S)], \end{aligned} \tag{31}$$

and

$$d_\Sigma(x, y) = n^{-1/2} \left\| \Sigma^{1/2}(x - y) \right\|_2, \qquad \Delta_q = \sup_{x \in \mathcal{X}_q} d_\Sigma(x, 0). \tag{32}$$

For a pseudometric space (T, d), let

$$\gamma_2(T, d) = \inf_{(T_s)} \sup_{t \in T} \sum_{s \ge 0} 2^{s/2} d(t, T_s), \tag{33}$$

where $|T_0| = 1$ and $|T_s| \le 2^{2^s}$.

Theorem A.3 (Conditional fixed-index exchange deviation)**.** *Under* (29)–(30)*, there is a universal C such that, conditionally on $\mathscr{A}$, with probability at least $1 - \delta$, simultaneously for $0 \le q \le q_{\max}$,*

$$\sup_{x \in \mathcal{X}_q(S)} \left| (\widehat{r}_S - r_S)^\top x \right| \le C \left[\gamma_2(\mathcal{X}_q^{\pm}, d_\Sigma) + \Delta_q \sqrt{\log \frac{e(q_{\max} + 1)}{\delta}} \right]. \tag{34}$$

The event controls only the terminal target-gradient contrasts. Any curvature, covariance, or geometry quantity estimated from these rows needs its own simultaneous bound.

Proof. Condition on $\mathscr{A}$ and write $Z_x = \bar{\xi}^\top x$. Independence and (30) give

$$\log \mathbb{E}[e^{\lambda(Z_x - Z_y)} \mid \mathscr{A}] \le \frac{\lambda^2}{2} d_\Sigma(x, y)^2. \tag{35}$$

Apply the generic-chaining tail inequality (Dirksen, 2015) to the fixed, zero-anchored symmetric set $\mathcal{X}_q^{\pm}$. Its supremum is the absolute supremum over $\mathcal{X}_q$, and its diameter is at most $2\Delta_q$. Choosing $u = \sqrt{\log(e(q_{\max} + 1)/\delta)}$ and union-bounding the feasible radii proves (34), after absorbing universal constants. □

B Affine-invariant geometry and decision certificates

This section uses the weighted-ERM setup and exact identity from Appendix A. Write

$$v = H^{-1} \nabla R(\theta(w)), \qquad E_H(w, d) = u_d^\top H u_d = q_d^\top H^{-1} q_d. \tag{36}$$

Define the weight-point quantity

$$a_H(w)^2 = \nabla R(\theta(w))^\top H^{-1} \nabla R(\theta(w)). \tag{37}$$

For $E_H(w, d) > 0$, define

$$\mu_H(w, d) = \frac{u_d^\top H_R u_d - \langle v, \mathcal{T}[u_d, u_d] \rangle}{E_H(w, d)}, \tag{38}$$

$$\beta_H(w, d) = \frac{2\,|\langle v, H_d u_d \rangle|}{\|d\|_1 \sqrt{E_H(w, d)}}. \tag{39}$$

The ratio μ_H combines target curvature and the third-order training response in the form needed by the one-sided bound. The ratios are not assigned a continuous extension at $E_H = 0$. In that case $u_d = 0$ and the exact identity directly gives $D^2 F(w)[d, d] = 0$. For bookkeeping only, set $\beta_H(w, d) = 0$ on such a degenerate direction.

For a symmetric matrix B, define its relative norm

$$\|B\|_{H,\mathrm{op}} = \sup_{z\neq 0} \frac{|z^\top B z|}{z^\top H z} = \left\| H^{-1/2} B H^{-1/2} \right\|_{\mathrm{op}}, \qquad \zeta_H(w) = \max_{i,j} \|A_i - A_j\|_{H,\mathrm{op}}, \tag{40}$$

where $A_i = \nabla^2 \ell_i(\theta(w))$. The numerator is an absolute quadratic form; equivalently, the definition controls both generalized spectral signs when B is indefinite.

Proposition B.1 (Affine reparameterization invariance)**.** *Let P be invertible and b be fixed. Under the pullback*

$$\widetilde{L}(z,w) = L(Pz+b,w), \qquad \widetilde{R}(z) = R(Pz+b), \tag{41}$$

the hyperobjective $F(w)$ and the quantities $E_H, a_H, \mu_H, \beta_H, \zeta_H$ are unchanged. This is invariance under invertible affine parameter changes, not under arbitrary nonlinear reparameterizations.

Proof. Quantities in the z coordinates satisfy

$$\begin{aligned} &\widetilde{H} = P^\top H P, \qquad \widetilde{r} = P^\top r, \qquad \widetilde{q}_d = P^\top q_d, \qquad \widetilde{u}_d = P^{-1} u_d, \\ &\widetilde{H}_d = P^\top H_d P, \qquad \widetilde{H}_R = P^\top H_R P, \qquad \widetilde{\mathcal{T}}[x,y] = P^\top \mathcal{T}[Px, Py]. \end{aligned} \tag{42}$$

Substitution proves invariance of E_H, a_H, μ_H, and β_H. For the matrix diameter, the generalized Rayleigh quotient gives

$$\sup_{z\neq 0} \frac{|z^\top P^\top (A_i - A_j) P z|}{z^\top P^\top H P z} = \sup_{x\neq 0} \frac{|x^\top (A_i - A_j) x|}{x^\top H x}, \tag{43}$$

where $x = Pz$. Maximizing over i, j completes the proof. □

Theorem B.2 (Intrinsic localized subset certificate)**.** *Fix a size-k subset S. For $1 \le q \le q_{\max}$, let*

$$\mathfrak{P}_q(S) = \left\{ (w_S + t d_{S,T}, d_{S,T}) : \begin{array}{l} |T| = k,\ |T \setminus S| = q,\ t \in [0,1], \\ d_{S,T} = w_T - w_S,\ E_H(w_S + t d_{S,T}, d_{S,T}) > 0 \end{array} \right\}. \tag{44}$$

Assume every comparator chord has a C^2 solution-map extension to a relatively open neighborhood. Define

$$m_q = \inf_{(w,d)\in\mathfrak{P}_q(S)} \mu_H(w,d), \qquad \beta_q = \sup_{(w,d)\in\mathfrak{P}_q(S)} \beta_H(w,d), \tag{45}$$

with $m_q = +\infty, \beta_q = 0$ if the path family is empty; in that case the radius-q curvature penalty in (46) *is defined to be zero. If $m_q > 0$ and $\beta_q < \infty$ for every nonempty radius, then*

$$\mathrm{Reg}_k(S) \le \max\left\{ 0, \max_{1\le q\le q_{\max}} \left[\Gamma_q(S) + \frac{\beta_q^2}{2m_q} \left(\frac{q}{k}\right)^2 \right] \right\}, \tag{46}$$

where the inner maximum is omitted if $q_{\max} = 0$ and, for $q \ge 1$,

$$\Gamma_q(S) = \max_{T:\,|T|=k,\ |T\setminus S|=q} DF(w_S)[w_T - w_S]. \tag{47}$$

Writing $h_i = \nabla \ell_i(\theta_S)^\top H_S^{-1} \nabla R(\theta_S)$, Γ_q is computed by taking the q largest h_i outside S and removing the q smallest inside S.

Proof. At a nondegenerate (w,d), Theorem A.1 and (38)–(39) give

$$D^2F(w)[d,d] \le -m_q E_H(w,d) + \beta_q \|d\|_1 \sqrt{E_H(w,d)}. \tag{48}$$

For $x = \sqrt{E_H(w,d)}$, completing the square yields

$$D^2F(w)[d,d] \le \frac{\beta_q^2}{4m_q} \|d\|_1^2. \tag{49}$$

The same inequality is automatic when $E_H = 0$. Taylor integration along the entire chord therefore gives

$$F(w_T) - F(w_S) \le DF(w_S)[d_{S,T}] + \frac{\beta_q^2}{8m_q} \|d_{S,T}\|_1^2. \tag{50}$$

Since $\|d_{S,T}\|_1 = 2q/k$, maximizing first within q and then over feasible radii proves (46). The sorting formula follows from linear optimization with exactly q exchanges. □

Lemma B.3 (Response energy is bounded by the gradient diameter)**.** *Let $\mathbf{1}^\top d = 0$ and write $\gamma(\theta, H) = \max_{i,j} \|\nabla\ell_i(\theta) - \nabla\ell_j(\theta)\|_{H^{-1}}$. Then*

$$\sqrt{E_H} = \|q_d\|_{H^{-1}} \le \tfrac{1}{2} \|d\|_1 \, \gamma(\theta, H). \tag{51}$$

Proof. Write $d = d^+ - d^-$ with $d_i^\pm = \max(\pm d_i, 0)$. Since $\mathbf{1}^\top d = 0$, $\sum_i d_i^+ = \sum_i d_i^- =: \sigma$ and $\|d\|_1 = 2\sigma$; if $\sigma = 0$ then $d = 0$ and the claim is trivial. Put $\alpha = d^+/\sigma$ and $\beta = d^-/\sigma$, both in the probability simplex. Using $\sum_j \beta_j = 1$ in the first sum and $\sum_i \alpha_i = 1$ in the second,

$$q_d = \sigma\Big(\sum_i \alpha_i \nabla\ell_i(\theta) - \sum_j \beta_j \nabla\ell_j(\theta)\Big) = \sigma \sum_{i,j} \alpha_i\beta_j\big(\nabla\ell_i(\theta) - \nabla\ell_j(\theta)\big), \tag{52}$$

a convex combination of pairwise gradient differences scaled by σ. The triangle inequality in $\|\cdot\|_{H^{-1}}$ gives $\|q_d\|_{H^{-1}} \le \sigma\gamma(\theta, H)$. □

Theorem B.4 (Truncated absorbed envelope)**.** *Adopt the hypotheses of Theorem B.2 except that m_q is not required to be positive, and suppose $\bar z_q$ satisfies $\sqrt{E_H} \le \bar z_q \|d\|_1$ at every point of $\mathfrak{P}_q(S)$. Set*

$$P_q = \max_{0\le z\le \bar z_q} \{-m_q z^2 + \beta_q z\} = \begin{cases} \dfrac{\beta_q^2}{4m_q}, & m_q > 0 \text{ and } \beta_q \le 2m_q\bar z_q, \\ \beta_q\bar z_q - m_q\bar z_q^2, & \text{otherwise.} \end{cases} \tag{53}$$

Then

$$\mathrm{Reg}_k(S) \le \max\Big\{0, \max_{1\le q\le q_{\max}} \big[\Gamma_q(S) + 2P_q(q/k)^2\big]\Big\}. \tag{54}$$

The radius-q penalty never exceeds the one in Equation (46)*, and equals it whenever $m_q > 0$ and $\beta_q \le 2m_q\bar z_q$.*

Proof. Fix a chord of radius q and a point of $\mathfrak{P}_q(S)$, and write $s = \|d\|_1$ and $z = \sqrt{E_H}/s$, so $z \in [0, \bar z_q]$ by hypothesis. The pointwise curvature bound in the proof of Theorem B.2 reads $D^2F(w)[d,d] \le -m_qE_H + \beta_q s\sqrt{E_H} = s^2(-m_qz^2 + \beta_q z) \le s^2P_q$, where the last step maximizes over the admissible interval rather than over $z \ge 0$; no sign condition on m_q is used. Taylor integration against $(1-t)$ contributes a factor $\frac{1}{2}$, and $s = \|w_T - w_S\|_1 = 2q/k$ gives the radius-q penalty $2P_q(q/k)^2$. Taking the maximum over comparators of each radius, then over radii and with 0, gives (54). For the closed form, the map $z \mapsto -m_qz^2 + \beta_q z$ has stationary point $z^\star = \beta_q/(2m_q)$ when $m_q > 0$; it is nondecreasing on $[0,\infty)$ when $m_q \le 0$, and on $[0, \bar z_q]$ the maximum is attained at $z^\star$ if $z^\star \le \bar z_q$ and at $\bar z_q$ otherwise. Since $[0, \bar z_q] \subseteq [0,\infty)$, $P_q \le \beta_q^2/(4m_q)$ whenever the latter is defined, which gives the comparison. □

Proposition B.5 (Absolute versus one-sided endpoint control)**.** *Let $\mathcal{C}$ be a finite comparator family containing S, and define*

$$\Delta_T = F(w_T) - F(w_S), \qquad \ell_T = DF(w_S)[w_T - w_S], \tag{55}$$

so that $\Delta_S = \ell_S = 0$.

1. *If nonnegative ρ_T satisfy $\Delta_T - \ell_T \le \rho_T$ for every $T \in \mathcal{C}$, then*

$$\mathrm{Reg}_{\mathcal{C}}(S) := \max_{T\in\mathcal{C}} \Delta_T \le \max\left\{0, \max_{T\in\mathcal{C}\setminus\{S\}} (\ell_T + \rho_T)\right\}. \tag{56}$$

2. *If instead $|\Delta_T - \ell_T| \le \varepsilon$ uniformly over $\mathcal{C}$, then the corresponding fixed-anchor bound is*

$$\mathrm{Reg}_{\mathcal{C}}(S) \le \max\left\{0, \max_{T\ne S}(\ell_T + \varepsilon)\right\}. \tag{57}$$

If $\widehat T \in \arg\max_{T\in\mathcal{C}} \ell_T$, its endpoint decision regret obeys

$$\max_{T\in\mathcal{C}}\{F(w_T) - F(w_{\widehat T})\} \le 2\varepsilon. \tag{58}$$

3. *Consider the common-Hessian quadratic inner problem*

$$L(\theta, w) = \tfrac{1}{2}\theta^\top A\theta - \left(\sum_i w_i b_i\right)^\top \theta + c(w), \qquad A \succ 0, \tag{59}$$

and the strictly convex quadratic target

$$R(\theta) = \tfrac{1}{2}(\theta - \theta_\star)^\top Q(\theta - \theta_\star), \qquad Q \succ 0. \tag{60}$$

Writing $\delta_T = \theta(w_T) - \theta(w_S)$, one has exactly

$$\Delta_T = \ell_T - \tfrac{1}{2}\delta_T^\top Q \delta_T. \tag{61}$$

Consequently, $\rho_T = 0$ is valid for every comparator, whereas the smallest uniform absolute remainder is

$$\varepsilon^\star_{\mathrm{abs}} = \tfrac{1}{2}\max_{T\in\mathcal{C}} \delta_T^\top Q \delta_T, \tag{62}$$

which is strictly positive whenever some comparator has $\theta(w_T) \neq \theta(w_S)$.

Proof. The first claim follows by maximizing $\Delta_T \le \ell_T + \rho_T$ and retaining the exact anchor $\Delta_S = 0$. The fixed-anchor absolute bound is the special case $\rho_T = \varepsilon$. If $T^\star$ maximizes the exact endpoint, then

$$\begin{aligned} F(w_{T^\star}) - F(w_{\widehat{T}}) &= \Delta_{T^\star} - \Delta_{\widehat{T}} \\ &\le \ell_{T^\star} + \varepsilon - (\ell_{\widehat{T}} - \varepsilon) \le 2\varepsilon, \end{aligned} \tag{63}$$

where the last inequality uses maximality of $\widehat{T}$ for the linear surrogate. For the quadratic model, $\theta(w) = A^{-1}\sum_i w_i b_i$ is affine. Expanding the target at $\theta(w_S)$ gives

$$R(\theta(w_T)) - R(\theta(w_S)) = \nabla R(\theta(w_S))^\top \delta_T + \tfrac{1}{2}\delta_T^\top Q \delta_T. \tag{64}$$

Because $F = -R$ and $\ell_T = -\nabla R(\theta(w_S))^\top \delta_T$, this proves (61); positive definiteness of Q proves the last claim. □

Remark B.6 (A strict two-point witness). Take $N = 2$, $k = 1$, scalar losses $\ell_1(\theta) = \frac{1}{2}(\theta - 1)^2$ and $\ell_2(\theta) = \frac{1}{2}(\theta - 4)^2$, target $R(\theta) = \frac{1}{2}\theta^2$, and $S = \{1\}$. For $T = \{2\}$, $\ell_T = -3$, $\Delta_T = -15/2$, and $\Delta_T - \ell_T = -9/2$. Thus S is uniquely self-top-1 and the one-sided certificate returns the exact value $\mathrm{Reg}_1(S) = 0$, while the optimal uniform absolute radius is $9/2$ and the corresponding sign-blind anchor bound is $3/2 > 0$. This separates certification sharpness by the sign of the remainder; it is not a claim that first-order control dominates higher-order IJ.

Proposition B.7 (Target curvature is not optional)**.** *Fix the training data $\{\ell_i\}, r_0$, a returned size-k subset S, and a vector g, and let*

$$\mathcal{R}(g) = \{R \in C^2 : \nabla R(\theta_S) = g\}. \tag{65}$$

Every input the certificate reads other than the target's second derivative is constant on $\mathcal{R}(g)$: the losses and their derivatives, the solution map $\theta(\cdot)$ and hence every endpoint θ_T, the base objects H_S and $\mathcal{T}$, the scores $h_i(S) = \nabla\ell_i(\theta_S)^\top H_S^{-1} g$, and every $\Gamma_q(S)$. Suppose some feasible T has $\delta_T = \theta_T - \theta_S \neq 0$. Then

$$\sup_{R\in\mathcal{R}(g)} \mathrm{Reg}_k(S) = +\infty, \qquad \inf_{R\in\mathcal{R}(g)} \mathrm{Reg}_k(S) = 0, \tag{66}$$

and on the convex subclass the supremum equals $\Gamma^{\mathrm{lin}} := \max_T\{-g^\top\delta_T\}$ while the infimum is still 0, both attained. Hence no function of the listed quantities is a finite valid upper bound on $\mathrm{Reg}_k(S)$ over $\mathcal{R}(g)$, and if $\Gamma^{\mathrm{lin}} > 0$ none is valid with a bounded ratio to $\mathrm{Reg}_k(S)$ over the convex subclass.

Proof. The training problem does not involve R, so $\theta(\cdot)$ and every derivative of L are independent of the target, and $h_i(S)$ and $\Gamma_q(S)$ depend on R only through $\nabla R(\theta_S) = g$. For $c \in \mathbb{R}$ set $R_c(\theta) = g^\top(\theta - \theta_S) + \frac{c}{2}\|\theta - \theta_S\|_2^2$, so that $R_c \in \mathcal{R}(g)$, and R_c is convex exactly when $c \ge 0$. Because $F(w) = -R(\theta(w))$,

$$F(w_T) - F(w_S) = R_c(\theta_S) - R_c(\theta_T) = -g^\top\delta_T - \tfrac{c}{2}\|\delta_T\|_2^2. \tag{67}$$

With $\delta_T \neq 0$, letting $c \to -\infty$ sends the right-hand side to $+\infty$, and letting $c \to +\infty$ makes every feasible gain negative, so $\mathrm{Reg}_k(S) = 0$; this proves (66). For $c \geq 0$ each gain decreases in c, so the supremum over the convex subclass is attained at the affine target $c = 0$, where $\mathrm{Reg}_k(S) = \max\{0, \Gamma^{\mathrm{lin}}\}$, and every gain is nonpositive once $c \geq \max_T\{-2g^\top \delta_T / \|\delta_T\|_2^2\}$, where $\mathrm{Reg}_k(S) = 0$. A bound that is constant on $\mathcal{R}(g)$ must dominate the supremum to be valid, which gives both consequences. □

Remark B.8 (A one-dimensional witness). Take $N = 2$, $k = 1$, $\ell_1(\theta) = \frac{1}{2}\theta^2$ and $\ell_2(\theta) = \frac{a}{2}(\theta - 1)^2$ with $a > 0$, so $\theta_S = 0$, $\theta_T = 1$, $H_S = 1$, $H_d = a - 1$, and $H(t) = 1 - t + ta$ along the chord. With $g = -1$ the scores are $h_1 = 0$ and $h_2 = a$, so $\Gamma_1(S) = a$ for every target in $\mathcal{R}(g)$, while $\mathrm{Reg}_1(S) = \max\{0, 1 - c/2\}$ under R_c. Choosing $a = 1/2$ makes the tangent term alone invalid: $\mathrm{Reg}_1(S) \to 1$ as $c \downarrow 0$, against $\Gamma_1(S) = 1/2$. Here $\mathcal{T} = 0$ gives $\mu_H(t) = c/H(t)$ and $\beta_H(t) = |r(t)|\,|a - 1|/H(t)^{3/2}$. The pre-truncation penalty $\beta_q^2/(2m_q)$ would grow like $1/c$ as the target flattens; under Theorem B.4 the admissible energy is capped by $\bar{z}_q = \bar{\gamma}_q/2 = 1/\sqrt{2}$. Since $H(t) = 1 - t/2 \in [1/2, 1]$, valid envelopes are $m = c$ and $\beta_q = \sqrt{2}$ for $0 \leq c \leq 1$, and $m = 2c$ and $\beta_q = \sqrt{2}(1 - c)$ for $c < 0$; both are attained at $t = 1$. The maximizer sits at the cap in either case, giving the finite values $5/2 - c$ and $5/2 - 4c$. The certificate still moves with the one quantity that distinguishes these instances, and diverges only as $c \to -\infty$, where the regret itself is unbounded.

Lemma B.9 (Factorized intrinsic coupling)**.** *For every tangent d,*

$$\|H_d\|_{H,\mathrm{op}} \leq \frac{\zeta_H(w)}{2}\,\|d\|_1\,. \tag{68}$$

For $E_H(w, d) > 0$ this implies

$$\beta_H(w, d) \leq a_H(w)\zeta_H(w). \tag{69}$$

Proof. Write $d = d^+ - d^-$ and $s = \sum_i d_i^+ = \sum_i d_i^- = \|d\|_1/2$. If $s > 0$, the matrices $P_+ = s^{-1}\sum_i d_i^+ A_i$ and $P_- = s^{-1}\sum_i d_i^- A_i$ are convex combinations and

$$H_d = s(P_+ - P_-) = s\sum_{i,j} \frac{d_i^+ d_j^-}{s^2}(A_i - A_j). \tag{70}$$

Convexity of the relative norm proves (68). Moreover,

$$\begin{aligned} 2\left|\left\langle H^{-1}r, H_d u_d\right\rangle\right| &\leq 2\left\|H^{-1/2}r\right\|_2 \left\|H^{-1/2}H_d H^{-1/2}\right\|_{\mathrm{op}} \left\|H^{1/2}u_d\right\|_2 \\ &\leq a_H\zeta_H\,\|d\|_1\,\sqrt{E_H}, \end{aligned} \tag{71}$$

which proves (69). The case $s = 0$ is trivial. □

Corollary B.10 (Relative third-order envelope)**.** *Suppose, pointwise with the same local $H(w)$-norm, that the symmetric third derivative satisfies*

$$\left|D_\theta^3 L(\theta(w), w)[x, x, y]\right| \leq M\,\|x\|_H^2\,\|y\|_H \tag{72}$$

for all x, y. Then, for $E_H > 0$,

$$\frac{\left\langle H^{-1}r, \mathcal{T}[u_d, u_d]\right\rangle}{E_H} \leq M a_H(w). \tag{73}$$

Thus a pathwise generalized-eigenvalue lower bound

$$\lambda_{\min}(H_R, H) - M a_H \geq \underline{m} > 0 \tag{74}$$

together with $a_H \leq \bar{a}$ and $\zeta_H \leq \bar{\zeta}$ yields the valid factorized remainder $\bar{a}^2\bar{\zeta}^2(2\underline{m})^{-1}(q/k)^2$.

Proof. Use symmetry to write the numerator in (73) as $D_\theta^3 L[u_d, u_d, H^{-1}r]$, then apply (72) and $\left\|H^{-1}r\right\|_H = a_H$. Lemma B.9 and Theorem B.2 finish the argument. □

B.1 Replicated full-swap sharpness for every budget

The following construction tests the scale of the heterogeneity correction. Replication makes the geometry of these two losses available at every budget k. The asymptotic comparison varies the Hessian perturbation δ toward zero, and establishes the necessity of its quadratic contribution on this family. The finite-δ comparison also uses envelopes valid along the whole path.

Theorem B.11 (Replicated quadratic sharpness)**.** *Fix any $k \geq 1$, let $N = 2k$, and take k identical copies of each of the two losses*

$$\begin{aligned} \ell_1(\theta) &= \tfrac{1}{2}\theta^\top A_1\theta - b_1^\top\theta, & A_1 &= \mu_0 I, & b_1 &= \mu_0 e_1, \\ \ell_2(\theta) &= \tfrac{1}{2}\theta^\top A_2\theta - b_2^\top\theta, & A_2 &= \mu_0(I + \delta J), & b_2 &= \mu_0(e_1 + 2\delta e_2), \end{aligned} \tag{75}$$

where $\mu_0, \alpha > 0$, $0 < \delta \leq 1/4$, and $J = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Let $R(\theta) = \alpha \|\theta\|_2^2/2$, and let S contain all type-1 copies. Then every local hypergradient score at S is zero, yet the unique best mixture count is the subset T containing all type-2 copies, and

$$\mathrm{Reg}_k(S) = F(w_T) - F(w_S) = \frac{\alpha\delta^2(1-3\delta^2)}{2(1-\delta^2)^2} > 0. \tag{76}$$

At S,

$$a_H^2 = \frac{\alpha^2}{\mu_0}, \qquad \zeta_H = \delta, \qquad m_{\mathrm{base}} := \inf_{d:\, E_H(S,d)>0} \mu_H(S,d) = \frac{\alpha}{\mu_0}, \qquad \frac{a_H^2\zeta_H^2}{2m_{\mathrm{base}}} = \frac{\alpha\delta^2}{2}. \tag{77}$$

Thus the factorized intrinsic remainder is matched to first order as $\delta \downarrow 0$ for every budget k.

Uniformly over the full-swap path, the valid envelopes

$$\bar{a}_{\mathrm{env}}^2 = \frac{\alpha^2}{\mu_0(1-\delta)}, \quad \bar{\zeta}_{\mathrm{env}} = \frac{\delta}{1-\delta}, \quad \underline{m}_{\mathrm{env}} = \frac{\alpha}{\mu_0(1+\delta)} \tag{78}$$

satisfy the finite-δ comparison

$$\mathrm{Reg}_k(S) \geq \frac{39}{125}\frac{\bar{a}_{\mathrm{env}}^2\bar{\zeta}_{\mathrm{env}}^2}{2\underline{m}_{\mathrm{env}}}. \tag{79}$$

This is an explicit replicated-family sharpness result, not a nontrivial k-dependent combinatorial minimax theorem.

Proof. Any size-k subset containing q type-2 copies has $t = q/k$ and the same objective as the two-type mixture. With $x = t\delta$, direct inversion gives

$$\theta_t = \frac{1}{1-x^2}\begin{pmatrix} 1-2x^2 \\ x \end{pmatrix}. \tag{80}$$

At $t = 0$, the type-1 gradient is zero, the type-2 gradient is $-\mu_0\delta e_2$, and $H_0^{-1}\nabla R(\theta_0) = (\alpha/\mu_0)e_1$. Hence all scores are zero.

The utility gain at mixture level t is

$$F(t) - F(0) = \frac{\alpha}{2}\frac{x^2(1-3x^2)}{(1-x^2)^2}. \tag{81}$$

Its derivative with respect to x has the sign of $2x(1-5x^2)/(1-x^2)^3$, which is positive for $0 < x \leq 1/4$. Therefore $t = 1$ is globally best among all feasible mixture counts, and (76) follows. The base-point quantities in (77) follow by substitution into (37) and (40); the inner objective is quadratic, so its third-order term is zero.

Along the path, $H_t \succeq \mu_0(1-\delta)I$. Equation (80) gives $\|\theta_t\|_2 \leq 1$ because $x^2 \leq 1/16$. This proves the first envelope. Simultaneous diagonalization of H_t and J gives $\zeta_H(t) = \delta/(1-t\delta)$ and the generalized target-curvature lower bound $\alpha/[\mu_0(1+t\delta)]$, proving (78). Finally,

$$\frac{\mathrm{Reg}_k(S)}{\bar{a}_{\mathrm{env}}^2\bar{\zeta}_{\mathrm{env}}^2/(2\underline{m}_{\mathrm{env}})} = \frac{(1-3\delta^2)(1-\delta)}{(1+\delta)^3}. \tag{82}$$

Its derivative is $4(3\delta^2 - \delta - 1)/(1+\delta)^4 < 0$ on $(0, 1/4]$, and its value at $\delta = 1/4$ is $39/125$. □

B.2 NUMERICAL AND TWO-SIDED MODULES

Define the gradient diameter in the dual Hessian norm by

$$\gamma_H(w) = \max_{i,j} \sqrt{(g_i - g_j)^\top H^{-1}(g_i - g_j)}. \tag{83}$$

Lemma B.12 (Certified solve residual)**.** *Let $v = H^{-1}r$, let $\widetilde{v}$ be an approximation, and set $s = H\widetilde{v} - r$. If a valid one-sided bound $\|s\|_{H^{-1}} \le \eta_{\mathrm{solve}}$ is available, then*

$$\left|q_d^\top(\widetilde{v} - v)\right| \le \eta_{\mathrm{solve}}\sqrt{E_H(w,d)} \le \frac{\eta_{\mathrm{solve}}\gamma_H(w)}{2}\|d\|_1. \tag{84}$$

For a q-swap this is at most $\eta_{\mathrm{solve}}\gamma_H(w)q/k$. In particular, a certified $H \succeq \underline{\lambda} I$ implies $\|s\|_{H^{-1}} \le \|s\|_2/\sqrt{\underline{\lambda}}$.

Proof. Since $\widetilde{v} - v = H^{-1}s$, Cauchy–Schwarz in the dual Hessian norm proves the first inequality. The second follows from the same positive/negative decomposition as Lemma B.9, now applied to the convex hull of the example gradients. □

For $E_H > 0$ define

$$\Lambda_H(w,d) = \frac{\left|-u_d^\top H_R u_d + \left\langle H^{-1}r, \mathcal{T}[u_d,u_d]\right\rangle\right|}{E_H(w,d)}. \tag{85}$$

The numerator is an absolute value, so Λ_H is a two-sided curvature magnitude, not a one-sided ratio.

Lemma B.13 (Two-sided intrinsic remainder)**.** *Suppose, at every chord point in a q-swap family with $E_H > 0$, $\Lambda_H \le \bar{\Lambda}_q$ and $\beta_H \le \bar{\beta}_q$, and suppose on the whole chord that $\gamma_H \le \bar{\gamma}_q$. Then*

$$\left|D^2F(w)[d,d]\right| \le L_q^{\pm}\|d\|_1^2, \qquad L_q^{\pm} = \frac{\bar{\Lambda}_q\bar{\gamma}_q^2}{4} + \frac{\bar{\beta}_q\bar{\gamma}_q}{2}, \tag{86}$$

and every q-swap satisfies

$$|F(w_T) - F(w_S) - DF(w_S)[w_T - w_S]| \le 2L_q^{\pm}(q/k)^2. \tag{87}$$

Proof. At a point with $E_H = 0$, positive definiteness of H gives $u_d = 0$, so the exact identity yields $D^2F(w)[d,d] = 0$ and the bound is automatic. At a point with $E_H > 0$, the following ratio bounds apply. The first two terms of the exact identity are bounded by $\bar{\Lambda}_q E_H$. The heterogeneity term is bounded by $\bar{\beta}_q\|d\|_1\sqrt{E_H}$. The gradient-diameter argument gives $\sqrt{E_H} \le \bar{\gamma}_q\|d\|_1/2$. Substitute and integrate along the whole chord. □

B.3 TRANSFER TO A REAL TRAINING RECIPE

Let $\mathcal{C}$ be a fixed family of size-k subsets containing S. Let $F_{\mathrm{real}}(T)$ be expected utility under the deployed randomized recipe and let $\widetilde{F}(w)$ be a smooth surrogate utility. Define restricted regret and endpoint coupling below. In particular, Δ_{cpl} is the supremum of the *absolute* endpoint discrepancy.

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{real},\mathcal{C}}(S) &= \max_{T\in\mathcal{C}}\{F_{\mathrm{real}}(T) - F_{\mathrm{real}}(S)\},\\
\mathrm{Reg}_{\widetilde{F},\mathcal{C}}(S) &= \max_{T\in\mathcal{C}}\{\widetilde{F}(w_T) - \widetilde{F}(w_S)\},\\
\Delta_{\mathrm{cpl}}(\mathcal{C}) &= \sup_{T\in\mathcal{C}}\left|F_{\mathrm{real}}(T) - \widetilde{F}(w_T)\right|.
\end{aligned} \tag{88}$$

Theorem B.14 (Coupling transfer)**.** *For every $T \in \mathcal{C}$,*

$$\mathrm{Reg}_{\mathrm{real},\mathcal{C}}(S) \le \mathrm{Reg}_{\widetilde{F},\mathcal{C}}(S) + 2\Delta_{\mathrm{cpl}}(\mathcal{C}), \tag{89}$$

$$F_{\mathrm{real}}(T) - F_{\mathrm{real}}(S) \ge \widetilde{F}(w_T) - \widetilde{F}(w_S) - 2\Delta_{\mathrm{cpl}}(\mathcal{C}). \tag{90}$$

The factor 2 is sharp for endpoint sup errors. The family $\mathcal{C}$ gives $\mathcal{C}$-restricted expected-utility regret, and the full size-k family gives global expected-utility regret. The fixed-seed version is recorded below.

Remark B.15 (Fixed-seed diagnostic). The proof is an algebraic endpoint comparison and hence can be instantiated for any fixed deterministic map $F_{\mathrm{real}}^{(s)}$. If all endpoints in $\mathcal{C}$ have already been observed under seed s, the empirical quantity $\widehat{\Delta}_{\mathrm{seed}}^{(s)} = max_{T\in\mathcal{C}}|F_{\mathrm{real}}^{(s)}(T) - \widetilde{F}(w_T)|$ yields the same conditional inequality. This gives a seed-conditional guarantee for the observed endpoint map. Expected deployed utility is certified by the paired-repetition corollary below.

Proof. For every $T \in \mathcal{C}$,

$$F_{\mathrm{real}}(T) - F_{\mathrm{real}}(S) \le \widetilde{F}(w_T) - \widetilde{F}(w_S) + 2\Delta_{\mathrm{cpl}}. \tag{91}$$

Maximize to get (89); reverse the endpoint inequalities to get (90). □

Corollary B.16 (Finite-family paired coupling audit)**.** *Let $\mathcal{C}$ be fixed independently of the audit repetitions, with $M = |\mathcal{C}|$. For every T, suppose $D_{T,1}, \dots, D_{T,b}$ are independent, $\mathbb{E}D_{T,s} = F_{\mathrm{real}}(T) - \widetilde{F}(w_T)$, and the centered differences are τ-sub-Gaussian. Then with probability at least $1-\delta$,*

$$\Delta_{\mathrm{cpl}}(\mathcal{C}) \le \max_{T\in\mathcal{C}} \left|\bar{D}_T\right| + \tau\sqrt{\frac{2\log(2M/\delta)}{b}}. \tag{92}$$

Dependence across candidates is allowed; independence is required across the b repetitions for each candidate. Candidates added from the audit observations enter through an adaptive simultaneous class.

Proof. Apply the two-sided sub-Gaussian mean tail bound to each T and union-bound the M candidates. □

B.4 A NORM-DEPENDENT FINITE-STEP FALLBACK

Fix a training horizon K and consider an augmented state $z_{t+1} = G_t(z_t, w)$ with z_0 independent of w and $F_K(w) = -R(z_K(w))$. The following result is deliberately described as a norm-dependent fallback; it is not the affine-invariant static theorem above.

Theorem B.17 (General smooth finite-step envelope)**.** *Suppose every G_t is C^2 on an open neighborhood containing every certified weight chord and its complete state trajectories. Uniformly on that tube, assume*

$$\begin{aligned} \|D_z G_t\|_{\mathrm{op}} &\le \kappa_t, & \|D_w G_t[d]\|_2 &\le b_t \|d\|_1, \\ \left\|D^2_{zz} G_t[x,x]\right\|_2 &\le \rho_t \|x\|_2^2, & \left\|D^2_{zw} G_t[x,d]\right\|_2 &\le \upsilon_t \|x\|_2 \|d\|_1, \\ \left\|D^2_{ww} G_t[d,d]\right\|_2 &\le \omega_t \|d\|_1^2. \end{aligned} \tag{93}$$

Set $c_0 = s_0 = 0$ and

$$\begin{aligned} c_{t+1} &= \kappa_t c_t + b_t, \\ s_{t+1} &= \kappa_t s_t + \rho_t c_t^2 + 2\upsilon_t c_t + \omega_t. \end{aligned} \tag{94}$$

If $\|\nabla R(z_K)\|_2 \le B_R$ and $\left\|\nabla^2 R(z_K)\right\|_{\mathrm{op}} \le L_R$ uniformly over the terminal tube, then

$$\left|D^2 F_K(w)[d,d]\right| \le C_K^{\mathrm{dyn}} \|d\|_1^2, \qquad C_K^{\mathrm{dyn}} = L_R c_K^2 + B_R s_K. \tag{95}$$

Consequently every certified q-swap chord obeys a two-sided Taylor remainder $2C_K^{\mathrm{dyn}}(q/k)^2$.

Proof. With $J_t = Dz_t[d]$ and $V_t = D^2 z_t[d,d]$, the chain rule gives

$$\begin{aligned} J_{t+1} &= D_z G_t J_t + D_w G_t[d], \\ V_{t+1} &= D_z G_t V_t + D^2_{zz} G_t[J_t, J_t] + 2D^2_{zw} G_t[J_t, d] + D^2_{ww} G_t[d,d]. \end{aligned} \tag{96}$$

Induction gives $\|J_t\|_2 \le c_t \|d\|_1$ and $\|V_t\|_2 \le s_t \|d\|_1^2$. Since

$$D^2 F_K[d,d] = -J_K^\top \nabla^2 R(z_K) J_K - \nabla R(z_K)^\top V_K, \tag{97}$$

(95) follows. Taylor integration supplies the swap remainder. □

B.5 CERTIFY, IMPROVE, OR ABSTAIN

Let

$$\mathcal{Q}_{\mathcal{C}}(S) = \{q \in \{1, \dots, q_{\max}\} : \exists T \in \mathcal{C}, |T \setminus S| = q\}. \tag{98}$$

Use the convention $\max_{q \in \varnothing} U_q = -\infty$. For $q \in \mathcal{Q}_{\mathcal{C}}(S)$, suppose a common valid event supplies, uniformly over all candidates in $\mathcal{C}$ at radius q, the following two-sided linear-error bound and Taylor controls.

$$\begin{aligned} \left|\widehat{\ell}_T - \ell_T\right| &\le \varepsilon_q, \qquad \ell_T = D\widetilde{F}(w_S)[w_T - w_S], \\ \widetilde{F}(w_T) - \widetilde{F}(w_S) &\le \ell_T + D_q^+, \\ \left|\widetilde{F}(w_T) - \widetilde{F}(w_S) - \ell_T\right| &\le D_q^{\pm}. \end{aligned} \tag{99}$$

Let $\widehat{T}_q$ maximize $\widehat{\ell}_T$ and set $\widehat{\Gamma}_q = \widehat{\ell}_{\widehat{T}_q}$. Let $\Delta \ge \Delta_{\mathrm{cpl}}(\mathcal{C})$ be valid on the same event.

Theorem B.18 (No-false-certificate and no-false-improvement)**.** *For $q \in \mathcal{Q}_{\mathcal{C}}(S)$, define*

$$U_q = \widehat{\Gamma}_q + \varepsilon_q + D_q^+ + 2\Delta, \qquad L_q = \widehat{\Gamma}_q - \varepsilon_q - D_q^{\pm} - 2\Delta, \tag{100}$$

and treat the $q = 0$ self-comparator as exactly zero. Then

$$\mathrm{Reg}_{\mathrm{real},\mathcal{C}}(S) \le \max\{0, \max_{q \in \mathcal{Q}_{\mathcal{C}}(S)} U_q\}, \tag{101}$$

$$F_{\mathrm{real}}(\widehat{T}_q) - F_{\mathrm{real}}(S) \ge L_q. \tag{102}$$

Thus an upper bound at most ϵ safely returns CERTIFY*, any $L_q > 0$ safely returns* IMPROVE*, and all remaining cases return* ABSTAIN*.*

If a best comparator $T^\star \in \mathcal{C}$ has nonzero radius $q^\star = |T^\star \setminus S| \in \mathcal{Q}_{\mathcal{C}}(S)$, then

$$L_{q^\star} \ge \mathrm{Reg}_{\mathrm{real},\mathcal{C}}(S) - \left(D_{q^\star}^+ + D_{q^\star}^{\pm} + 2\varepsilon_{q^\star} + 4\Delta\right). \tag{103}$$

Proof. The true maximum linear gain is at most $\widehat{\Gamma}_q + \varepsilon_q$, so the first inequality in (99) and Theorem B.14 give (101). The selected candidate has true linear gain at least $\widehat{\Gamma}_q - \varepsilon_q$; the two-sided lower Taylor bound and coupling transfer give (102).

For a best comparator $T^\star$, coupling and the one-sided remainder imply $\ell_{T^\star} \ge \mathrm{Reg}_{\mathrm{real},\mathcal{C}}(S) - 2\Delta - D_{q^\star}^+$. Maximality of $\widehat{T}_{q^\star}$ loses at most $2\varepsilon_{q^\star}$ between the two true linear values. Applying the output candidate's two-sided remainder and a second coupling transfer proves (103). □

Remark B.19 (Certified-input requirements). The preceding modules execute from valid one-sided, path-uniform envelopes. Spectral terms use certified upper enclosures, and Euclidean CG residuals use a valid conversion to the H^{-1} norm. A finite paired audit certifies its declared proposal family; global decisions use envelopes over all size-k subsets. Modern-model implementation therefore centers on constructing structured, nonvacuous versions of these inputs.

Operational decision procedure. Fix S, the comparator family $\mathcal{C}$ containing S, the tolerance ϵ, and the total failure allowance δ before accessing audit evidence. The following procedure implements Theorem B.18.

1. Obtain jointly valid score, upper-remainder, lower-remainder and coupling bounds for the declared family. Allocate failure allowances to the stochastic input events so their sum is at most δ; deterministic enclosures require no failure allowance. If the inputs share one already valid joint event, use that event directly. The experiment-specific allocations appear in Appendix H.
2. For every represented radius, compute $\widehat{\Gamma}_q$ and a maximizing replacement $\widehat{T}_q$. Use Equation (100) for each supported bound. Set $U_q = +\infty$ if the score, upper remainder or coupling bound is unavailable; set $L_q = -\infty$ if the score, lower remainder or coupling bound is unavailable. Failed tube containment invalidates the shared upper remainder at that radius.

3. Set $U = \max\{0, \max_q U_q\}$ and form two separate flags, CERTIFY $= [U \le \epsilon]$ and IMPROVE $= [\max_q L_q > 0]$. Return both flags, U, and an improving $\widehat{T}_q$ with its lower bound when one exists. Return ABSTAIN only when both flags are false. The empty family of nonzero radii has regret zero and no improving replacement.

When coupling comes from Corollary B.16, each of the M family members needs b repetitions. To make its stochastic allowance at most $\eta > 0$, a sufficient count is $b \ge \lceil 2\tau^2 \log(2M/\delta_{\mathrm{cpl}})/\eta^2 \rceil$, using a valid scale τ and an allocated failure allowance δ_{cpl}. Auditing only a sample of family members controls that sampled family unless an additional uniform argument covers the unaudited members.

B.6 INDEPENDENCE OF THE TARGET EVIDENCE

Conditioning on artifacts formed before the certificate target is observed fixes the returned subset and its exchange responses. The fixed-subset bound below therefore applies under the conditional model of Appendix A. If target data also select the subset, the fixed-index event cannot simply be evaluated at that data-dependent choice: valid reuse requires simultaneous control over all choices those data can induce, including any data-dependent training responses and geometry estimates. The statistical certificate used here retains the independent-target setting.

B.7 CONDITIONAL MODEL AND FULL STATISTICAL BOUND

Use the conditional model, response sets and metric in Equations (29)–(32). Set $\gamma_{2,q} = \gamma_2(\mathcal{X}_q^{\pm}, d_\Sigma)$. Theorem A.3 gives, simultaneously over the feasible radii,

$$\sup_{x \in \mathcal{X}_q} |(\widehat{r}_S - r_S)^\top x| \le C \left[\gamma_{2,q} + \Delta_q \sqrt{\log \frac{e(q_{\max}+1)}{\delta}} \right]. \tag{104}$$

The width measures response diversity and the radius controls confidence. An estimated covariance proxy must have its own envelope on the same event.

Let m_q and β_q be the pathwise intrinsic constants in (45), and let $\widehat{\Gamma}_q(S) = \max_T \widehat{r}_S^\top x_T$ over the same q-swap class. Adding this statistical allowance to the geometric penalty combines (104) with Theorem B.2 to give

$$\mathrm{Reg}_k(S) \le \max\left\{0, \max_{1 \le q \le q_{\max}} \left[\widehat{\Gamma}_q(S) + \frac{\beta_q^2}{2m_q} \left(\frac{q}{k}\right)^2 + C \left[\gamma_{2,q} + \Delta_q \sqrt{\log \frac{e(q_{\max}+1)}{\delta}} \right] \right] \right\}. \tag{105}$$

The inner maximum is omitted when $q_{\max} = 0$. Certified envelopes $0 < \underline{m}_q \le m_q$ and $\bar{\beta}_q \ge \beta_q$ may replace the unknown pathwise quantities, using $\bar{\beta}_q^2/(2\underline{m}_q)$. In the isotropic case, the statistical line is $C\sigma n^{-1/2}[\gamma_2(\mathcal{X}_q^{\pm}, \|\cdot\|_2) + R_q\sqrt{\log(e(q_{\max}+1)/\delta)}]$, where $R_q = \sup_{x \in \mathcal{X}_q} \|x\|_2$. This event supplies the terminal target-gradient contrasts. Target-Hessian and intrinsic-geometry estimates enter the certificate through their corresponding simultaneous envelopes.

C COMPUTATIONAL BOUNDARIES AND STRUCTURED NONVACUITY

This appendix connects computational precision, verification and nonvacuity. Theorem C.5 establishes difficulty at specified additive accuracies. Theorems C.10 and C.12 give a constructive route from regional regularity to the shared constants in Assumption 4.1. Corollary C.20 then exhibits an open ridge-softmax regime where these constants yield a useful certificate. These results separate the existence of a valid bound, a procedure for computing it, and conditions under which it is informative.

The ridge-softmax existence regime requires $B^2/\lambda < 1/64$, a small target gradient in the normalization of Equation (172), and a subset that is top-k under its own scores. This is a sufficient structured regime. The general regional construction also accepts bounds computed for each instance, as in the CIFAR experiments of Appendix D.

We use the weighted-ERM notation and the quantities $E_H, \mu_H, \beta_H, m_q, \beta_q$, and Γ_q from Appendices A–B.

C.1 Auxiliary facts

Lemma C.1 (Zero-sum diameter bound)**.** *Let* $(\mathsf{X}, \|\cdot\|)$ *be a normed linear space. If* $z_1, \dots, z_N \in \mathsf{X}$ *and* $d \in \mathbb{R}^N$ *satisfy* $\mathbf{1}^\top d = 0$*, then*

$$\left\| \sum_i d_i z_i \right\| \le \frac{\|d\|_1}{2} \max_{i,j} \|z_i - z_j\| . \tag{106}$$

Proof. Write $d = d^+ - d^-$ and $c = \sum_i d_i^+ = \sum_i d_i^- = \|d\|_1 / 2$. The claim is immediate if $c = 0$. Otherwise, $c^{-1} \sum_i d_i^+ z_i$ and $c^{-1} \sum_i d_i^- z_i$ are in the convex hull of $\{z_i\}$, whose diameter is at most that of the original set in any norm. Multiplication by c proves the result. □

Lemma C.2 (Vertices of the capped simplex)**.** *The extreme points of* $\mathcal{W}_k = \{w \in \mathbb{R}^N : 0 \le w_i \le 1/k,\ \mathbf{1}^\top w = 1\}$ *are exactly* $w_S = k^{-1}\mathbf{1}_S$ *with* $|S| = k$*. Consequently,* $\mathcal{W}_k = \operatorname{conv}\{w_S : |S| = k\}$.

Proof. Put $x = kw$. If two coordinates of a feasible $x \in [0,1]^N$ are strictly fractional, a sufficiently small perturbation in direction $e_i - e_j$ writes x as the midpoint of two distinct feasible points. Because $\mathbf{1}^\top x = k$ is integral, a nonintegral feasible point cannot have exactly one fractional coordinate. Hence every extreme point is a 0–1 vector with exactly k ones. The converse is immediate, and rescaling proves the claim. □

C.2 A computational boundary

The following reductions are stated in the deterministic Turing model with rational input and rational output. They are one-query polynomial-time Turing reductions from Clique. This is enough to rule out the postulated uniform algorithms unless $\mathsf{P} = \mathsf{NP}$. The reductions target uniformly near-tight certification; sound loose certificates and structured instance families are treated by the constructive results that follow.

Theorem C.3 (Near-tight path-margin computation is Clique-hard)**.** *Given an undirected simple graph* $G = (V, E)$ *with* $|V| = n$ *and an integer* $2 \le k \le n$*, one can construct in polynomial time a weighted ERM with* $n + k$ *candidates and parameter dimension* $n + k$ *such that every inner loss is a quadratic with identity Hessian, the target is a strongly convex quadratic with Hessian condition number smaller than three, and the full-swap margin from a declared dummy baseline* S *is*

$$m_k = 2 - \frac{1}{nk} \max_{T \subseteq V :\, |T| = k} |E_G(T)|. \tag{107}$$

If a deterministic polynomial-time algorithm always returned a certified rational lower bound $\underline{m}_k$ *satisfying*

$$m_k - \frac{1}{3nk} < \underline{m}_k \le m_k, \tag{108}$$

then $\mathsf{P} = \mathsf{NP}$.

Proof. Add a set D of k dummy candidates and take $S = D$. Index coordinates by $V \cup D$, let e_i be the corresponding coordinate vectors, and set

$$\ell_i(\theta) = \frac{1}{2} \|\theta - e_i\|_2^2 . \tag{109}$$

Since $\mathbf{1}^\top w = 1$,

$$L(\theta, w) = \frac{1}{2} \|\theta\|_2^2 - \theta^\top w + \frac{1}{2}, \qquad \theta(w) = w, \qquad H = I, \qquad \mathcal{T} = 0. \tag{110}$$

All sample Hessians equal I, so $H_d = (\mathbf{1}^\top d) I = 0$ on tangent directions and $\beta_H = 0$.

Let A_G be the adjacency matrix, let $A' = A_G \oplus 0_D$, and define

$$R(\theta) = \frac{1}{2}\theta^\top \left(2I - \frac{1}{n}A'\right)\theta. \tag{111}$$

The spectrum of A_G is contained in $[-(n-1), n-1]$. Hence

$$\left(1 + \frac{1}{n}\right) I \preceq \nabla^2 R \preceq \left(3 - \frac{1}{n}\right) I, \tag{112}$$

so the claimed conditioning is uniform in the graph.

For a full swap to $T \subseteq V$, $|T| = k$, put $d = w_T - w_S$. Directly,

$$\|d\|_2^2 = \frac{2}{k}, \qquad d^\top A' d = \frac{1}{k^2}\mathbf{1}_T^\top A_G \mathbf{1}_T = \frac{2|E_G(T)|}{k^2}. \tag{113}$$

The last factor of two is the double counting of every undirected edge by the adjacency quadratic form. Moreover, $g_i(\theta) = \theta - e_i$ and $\mathbf{1}^\top d = 0$, whence

$$q_d = \sum_i d_i(\theta - e_i) = -d, \qquad u_d = -d. \tag{114}$$

Thus $H, H_d, \mathcal{T}, u_d$, and H_R—the quantities needed here to evaluate μ_H and β_H—do not vary with the chord parameter. We do not assert that every auxiliary quantity of the bilevel problem is chord-independent. Using (113),

$$\mu_H(w, d) = \frac{d^\top(2I - A'/n)d}{\|d\|_2^2} = 2 - \frac{|E_G(T)|}{nk}. \tag{115}$$

Taking the infimum over the full-swap comparators proves (107).

If G contains a k-clique, then

$$m_k = m_{\mathrm{clq}} := 2 - \frac{k-1}{2n}. \tag{116}$$

Otherwise, the maximum induced edge count is at most $k(k-1)/2 - 1$, and hence $m_k \geq m_{\mathrm{clq}} + 1/(nk)$. In the first case $\underline{m}_k \leq m_{\mathrm{clq}}$; in the second, $\underline{m}_k > m_{\mathrm{clq}} + 2/(3nk)$. Thresholding at $m_{\mathrm{clq}} + 1/(3nk)$ decides CLIQUE after one call to the postulated algorithm.

All constructed rational coefficients have $O(\log n)$-bit encodings, the instance dimension is polynomial in n, and the threshold and promise gap have polynomial bit complexity. This proves the stated deterministic Turing reduction. □

Theorem C.4 (Uniformly near-tight safe regret certification is CLIQUE-hard)**.** *For the same instance and baseline as in Theorem C.3,*

$$\mathrm{Reg}_k(S) = \frac{1}{nk^2} \max_{T \subseteq V:\, |T|=k} |E_G(T)|. \tag{117}$$

If a deterministic polynomial-time algorithm always returned a rational safe upper bound $U(S)$ satisfying

$$\mathrm{Reg}_k(S) \leq U(S) \leq \mathrm{Reg}_k(S) + \frac{1}{3nk^2}, \tag{118}$$

then $\mathsf{P} = \mathsf{NP}$.

Proof. For any size-k vertex w_T,

$$F(w_T) = -R(w_T) = -\frac{1}{k} + \frac{|E_G(T \cap V)|}{nk^2}. \tag{119}$$

The dummy baseline has value $-1/k$. A dummy in a comparator can be replaced by an unused graph vertex without decreasing the already induced edge count; since $n \geq k$, a maximizer may be chosen wholly inside V. This proves (117).

In the clique case the regret is $r_{\mathrm{clq}} = (k-1)/(2nk)$. In the non-clique case it is at most $r_{\mathrm{clq}} - 1/(nk^2)$. Under (118), the clique case has $U(S) \geq r_{\mathrm{clq}}$, whereas the non-clique case has $U(S) \leq r_{\mathrm{clq}} - 2/(3nk^2)$. Thresholding at $r_{\mathrm{clq}} - 1/(3nk^2)$ decides CLIQUE. The rational encoding and one-query Turing-model qualifications are the same as in Theorem C.3. □

Theorem C.5 (Constant-conditioned certification boundary)**.** *There is a polynomial-time rational construction from each* CLIQUE *instance to a weighted ERM whose sample Hessians are the identity and whose target-Hessian condition number is smaller than three, with the following two distinct properties.*

1. *Computing a certified lower path-margin bound with uniform additive error strictly below* $1/(3nk)$ *as in* (108) *is* CLIQUE*-hard.*
2. *Computing a safe regret upper certificate with uniform additive slack at most* $1/(3nk^2)$ *as in* (118) *is* CLIQUE*-hard.*

Thus either postulated deterministic polynomial-time algorithm would imply $\mathsf{P} = \mathsf{NP}$*. The two parts concern different computational primitives. The first is a pathwise geometric margin, whereas the second is the final decision certificate.*

Proof. Part 1 is Theorem C.3, and Part 2 is Theorem C.4. Both use the same constant-conditioned construction and the same rational Turing-model encoding. □

Proposition C.6 (Strong convexity and zero coupling do not imply a tight tangent certificate)**.** *In the preceding construction, the exact hypergradient score at the dummy baseline is* $2/k$ *on every graph candidate and zero on every selected dummy. Consequently,*

$$\Gamma_q(S) = \frac{2q}{k^2}, \qquad \max_q \Gamma_q(S) = \frac{2}{k}. \tag{120}$$

If G *contains exactly one edge, then the true regret is* $1/(nk^2)$*, so the tangent certificate is a factor* $2nk$ *larger. If* G *is empty, the true regret is zero while the tangent gap remains* $2/k > 0$*.*

Proof. At w_S, the target gradient is $r = (2I - A'/n)w_S = 2w_S$. Therefore the intrinsic score $h_i = g_i(w_S)^\top r$ equals $2/k$ for graph coordinates and zero for selected dummy coordinates. A q-swap has linear gain $q(1/k)(2/k) = 2q/k^2$. The maximum feasible radius is k, proving (120). The exact regret claims follow from (117). □

Remark C.7 (Conclusion of the reductions). The examples have smooth strongly convex inner and outer objectives, a uniformly well-conditioned target Hessian, and $\beta_H = 0$. They show that general smoothness leaves both tangent-bound tightness and uniformly near-tight certificate computation unresolved in the worst case. The trust-region constructions below obtain tractable instance-specific certificates by adding checkable structure.

C.3 A NON-ENUMERATIVE TRUST-REGION CERTIFICATE

The conditions below serve separate roles. Coercivity ensures that minimizers exist. Convexity turns a verified outward gradient on the region's boundary into containment of every minimizer. Relative third-order control propagates the base Hessian bounds through that region. With containment established, regional target and sample derivative bounds supply the path constants used by Theorem 4.2. The finite update construction in Section 5.3 uses derivatives of the declared optimizer maps under its own trajectory conditions.

Fix a returned subset S, write $w_0 = w_S$, $\theta_0 = \theta(w_0)$, and set

$$H_0 = \nabla^2_\theta L(\theta_0, w_0) \succ 0, \qquad \mathcal{U}_q(S) = \left\{ w \in \mathcal{W}_k : \|w - w_0\|_1 \le \frac{2q}{k} \right\}. \tag{121}$$

For a symmetric matrix B, use $\|B\|_{H_0,\mathrm{op}} = \left\| H_0^{-1/2} B H_0^{-1/2} \right\|_{\mathrm{op}}$, with the absolute generalized Rayleigh quotient when B is indefinite. At θ_0, define

$$\gamma_0 = \max_{i,j} \|g_i(\theta_0) - g_j(\theta_0)\|_{H_0^{-1}}, \qquad \zeta_0 = \max_{i,j} \|A_i(\theta_0) - A_j(\theta_0)\|_{H_0,\mathrm{op}}. \tag{122}$$

Lemma C.8 (Base-point gradient and Hessian perturbations)**.** *For every* $w \in \mathcal{U}_q(S)$*,*

$$\|\nabla_\theta L(\theta_0, w)\|_{H_0^{-1}} \le g_q := \frac{q}{k}\gamma_0, \tag{123}$$

$$(1 - \delta_q)H_0 \preceq \nabla^2_\theta L(\theta_0, w) \preceq (1 + \delta_q)H_0, \qquad \delta_q := \frac{q}{k}\zeta_0. \tag{124}$$

Proof. Put $a = w - w_0$. Then $\mathbf{1}^\top a = 0$ and $\|a\|_1 \le 2q/k$. Stationarity at (θ_0, w_0) and cancellation of the weight-independent regularizer give

$$\nabla_\theta L(\theta_0, w) = \sum_i a_i g_i(\theta_0), \qquad \nabla^2_\theta L(\theta_0, w) - H_0 = \sum_i a_i A_i(\theta_0). \tag{125}$$

Apply Lemma C.1 first in the H_0^{-1} vector norm and then in the H_0-relative matrix norm. The matrix-norm bound is equivalent to the Loewner sandwich in (124). □

Lemma C.9 (Relative third order implies Hessian comparison)**.** *Suppose $\delta_q < 1$ and, for every $w \in \mathcal{U}_q(S)$ and every θ with $\|\theta - \theta_0\|_{H_0} \le R$,*

$$\left|D^3_\theta L(\theta, w)[x, x, y]\right| \le \kappa\, x^\top \nabla^2_\theta L(\theta, w) x\ \|y\|_{H_0} \tag{126}$$

for all x, y. Then every such (θ, w) satisfies

$$e^{-\kappa\|\theta-\theta_0\|_{H_0}} \nabla^2_\theta L(\theta_0, w) \preceq \nabla^2_\theta L(\theta, w) \preceq e^{\kappa\|\theta-\theta_0\|_{H_0}} \nabla^2_\theta L(\theta_0, w). \tag{127}$$

Proof. Fix x, put $\Delta = \theta - \theta_0$, and define $h(t) = x^\top \nabla^2_\theta L(\theta_0 + t\Delta, w)x$. Along the tube, (126) gives

$$|h'(t)| \le \kappa\, \|\Delta\|_{H_0}\, h(t). \tag{128}$$

Whenever $\delta_q < 1$, Lemma C.8 makes $h(0) > 0$ for $x \neq 0$, and the differential inequality preserves a strict positive lower bound. Integrating the two inequalities for $e^{\pm\kappa t\|\Delta\|_{H_0}} h(t)$ gives

$$e^{-\kappa t\|\Delta\|_{H_0}} h(0) \le h(t) \le e^{\kappa t\|\Delta\|_{H_0}} h(0). \tag{129}$$

Set $t = 1$ and invoke the Loewner characterization. □

Theorem C.10 (Certified tube containment)**.** *Assume that, for every $w \in \mathcal{U}_q(S)$, $L(\cdot, w)$ is differentiable, convex, and coercive, is C^3 in the closed H_0-tube of radius R, and that* (126) *holds in the closed H_0-tube of radius R. If*

$$\delta_q < 1, \qquad g_q < R(1 - \delta_q)e^{-\kappa R}, \tag{130}$$

then every global minimizer at every $w \in \mathcal{U}_q(S)$ obeys

$$\|\theta(w) - \theta_0\|_{H_0} < R. \tag{131}$$

Proof. Coercivity gives existence. Fix $w \in \mathcal{U}_q(S)$ and a boundary displacement v with $\|v\|_{H_0} = R$. Integrating the gradient and using Lemmas C.8–C.9 yields

$$\begin{aligned}\langle \nabla L(\theta_0 + v, w), v\rangle &= \langle \nabla L(\theta_0, w), v\rangle + \int_0^1 v^\top \nabla^2 L(\theta_0 + tv, w) v\, dt \\ &\ge -g_q R + (1 - \delta_q)e^{-\kappa R} R^2 > 0.\end{aligned} \tag{132}$$

Suppose a global minimizer θ_* has $r_* := \|\theta_* - \theta_0\|_{H_0} \ge R$. Put $z = (\theta_* - \theta_0)/r_*$ and $\phi(s) = L(\theta_0 + sz, w)$. If $r_* > R$, convexity and global optimality give

$$\phi'(R) \le \frac{\phi(r_*) - \phi(R)}{r_* - R} \le 0. \tag{133}$$

If $r_* = R$, differentiability and the first-order optimality condition at the unconstrained global minimizer give $\phi'(R) = \langle \nabla L(\theta_*, w), z\rangle = 0$ directly. In either case, (132) with $v = Rz$ says $R\phi'(R) > 0$, a contradiction. Thus every minimizer is strictly inside the tube. □

Remark C.11 (Selecting a radius). Keeping the exponential integral gives the sharper sufficient condition

$$g_q < (1 - \delta_q)R\frac{1 - e^{-\kappa R}}{\kappa R}, \tag{134}$$

where the ratio is one at $\kappa = 0$. If R is selected automatically, it must be verified inside a certified regularity interval. In particular, $Re^{-\kappa R}$ is not globally monotone, so an unrestricted monotone bisection is not valid; one may isolate roots by intervals, restrict to the increasing branch, or verify a prespecified R.

Theorem C.12 (Explicit non-enumerative path envelopes)**.** *In addition to Theorem C.10, suppose throughout the tube and for every $w \in \mathcal{U}_q(S)$ that*

$$\|A_i(\theta) - A_i(\theta_0)\|_{H_0,\mathrm{op}} \le \kappa_A \|\theta - \theta_0\|_{H_0}, \tag{135}$$

$$\alpha_R H_0 \preceq H_R(\theta) \preceq L_R H_0, \qquad 0 \le \alpha_R \le L_R. \tag{136}$$

Define

$$\underline{h}_q = (1-\delta_q)e^{-\kappa R}, \qquad \overline{h}_q = (1+\delta_q)e^{\kappa R}, \tag{137}$$

$$a_0 = \|\nabla R(\theta_0)\|_{H_0^{-1}}, \qquad a_q = a_0 + L_R R, \tag{138}$$

$$\zeta_q^{\mathrm{tube}} = \zeta_0 + 2\kappa_A R. \tag{139}$$

Then the pathwise constants in Theorem B.2 satisfy

$$m_q \ge \underline{m}_q := \frac{\alpha_R}{\overline{h}_q} - \frac{\kappa a_q}{\underline{h}_q}, \tag{140}$$

$$\beta_q \le \overline{\beta}_q := \frac{a_q \zeta_q^{\mathrm{tube}}}{\underline{h}_q^{3/2}}. \tag{141}$$

No comparator, path, direction, or parameter supremum remains on the right-hand sides.

Proof. Every q-swap chord lies in $\mathcal{U}_q(S)$. Tube containment and Lemmas C.8–C.9 therefore give

$$\underline{h}_q H_0 \preceq H(\theta(w), w) \preceq \overline{h}_q H_0 \tag{142}$$

at every relevant path point.

For $\Delta = \theta - \theta_0$, target-gradient integration and (136) imply

$$\|\nabla R(\theta) - \nabla R(\theta_0)\|_{H_0^{-1}} \le L_R \|\Delta\|_{H_0} \le L_R R, \tag{143}$$

so $\|r\|_{H_0^{-1}} \le a_q$. Put $v = H^{-1}r$. The lower bound in (142) gives the covector-to-vector estimate

$$\|v\|_{H_0} \le \frac{a_q}{\underline{h}_q}. \tag{144}$$

This is an $\underline{h}_q^{-1}$ factor, rather than merely an inverse square root.

For any nonzero response u,

$$\frac{u^\top H_R u}{u^\top H u} \ge \frac{\alpha_R}{\overline{h}_q}. \tag{145}$$

Using (126) with $x = u, y = v$ and then (144),

$$\frac{|\langle v, \mathcal{T}[u,u]\rangle|}{u^\top H u} \le \kappa \|v\|_{H_0} \le \frac{\kappa a_q}{\underline{h}_q}. \tag{146}$$

The definition of μ_H proves (140).

Next, (135) gives

$$\max_{i,j} \|A_i(\theta) - A_j(\theta)\|_{H_0,\mathrm{op}} \le \zeta_0 + 2\kappa_A R = \zeta_q^{\mathrm{tube}}. \tag{147}$$

Lemma C.1 therefore yields

$$\|H_d\|_{H_0,\mathrm{op}} \le \frac{\|d\|_1}{2} \zeta_q^{\mathrm{tube}}. \tag{148}$$

Together with (144),

$$2\left|v^\top H_d u\right| \le \|d\|_1 \zeta_q^{\mathrm{tube}} \frac{a_q}{\underline{h}_q} \|u\|_{H_0}. \tag{149}$$

Finally, $\sqrt{u^\top H u} \ge \sqrt{\underline{h}_q}\|u\|_{H_0}$. Dividing by the normalization in β_H gives (141). This last conversion accounts for the additional $\underline{h}_q^{-1/2}$ and hence for the exponent $3/2$. □

Proposition C.13 (Affine invariance of the trust-region result)**.** *Under any invertible affine reparameterization $\theta = Pz + b$, all assumptions, the containment condition, and all numerical outputs in Theorems C.10–C.12 are unchanged.*

Proof. Under the pullback,

$$H_{0,z} = P^\top H_0 P, \quad H_z = P^\top H P, \quad H_{R,z} = P^\top H_R P, \quad A_{i,z} = P^\top A_i P. \tag{150}$$

For $\Delta_\theta = P\Delta_z$ and a covector $r_z = P^\top r$,

$$\|\Delta_z\|_{H_{0,z}} = \|\Delta_\theta\|_{H_0}, \qquad \|r_z\|_{H_{0,z}^{-1}} = \|r\|_{H_0^{-1}}, \qquad \left\|P^\top B P\right\|_{P^\top H_0 P,\mathrm{op}} = \|B\|_{H_0,\mathrm{op}}. \tag{151}$$

Third derivatives transform multilinearly, and Loewner inequalities are preserved by congruence. Convexity and coercivity are preserved by an invertible affine map. Thus $\gamma_0, \zeta_0, g_q, \delta_q, \kappa, \kappa_A, R, \alpha_R, L_R$ and all derived envelopes have the same numerical values. □

Corollary C.14 (Trust-region subset certificate)**.** *If Theorems C.10–C.12 hold for every $1 \le q \le q_{\max}$ and $\underline{m}_q > 0$, then*

$$\mathrm{Reg}_k(S) \le \max\left\{0, \max_{1\le q\le q_{\max}}\left[\Gamma_q(S) + \frac{\overline{\beta}_q^2}{2\underline{m}_q}\left(\frac{q}{k}\right)^2\right]\right\}. \tag{152}$$

Proof. Theorem C.12 supplies sound one-sided path envelopes. Substitution into Theorem B.2 proves the result. □

Remark C.15 (Dropping the positivity requirement). Positivity of α_R is used nowhere in the proof of Theorem C.12; it enters only through $\underline{m}_q > 0$, which Theorem B.2 needs and Theorem B.4 does not. Substituting the same envelopes together with $\bar{z}_q$ from Lemma C.17 into Theorem B.4 therefore certifies at any sign of $\underline{m}_q$, which is what a convex but singular target such as unregularized softmax cross-entropy requires, its common-shift invariance forcing $\alpha_R = 0$ on the full parameter space.

Remark C.16 (Taking the target curvature on the reachable responses). Only $u_d = H^{-1}q_d$ enters Equation (8), so the *lower* bound in (136) needs to hold only on a subspace V containing every such response, and α_R may be replaced by $\alpha_R^V = \min\{u^\top H_R u / u^\top H_0 u : u \in V \setminus \{0\}\} \ge \alpha_R$ in (140). The upper bound $H_R \preceq L_R H_0$ must be kept on the full space, because the proof of Theorem C.12 uses it to bound $\|\nabla R(\theta) - \nabla R(\theta_0)\|_{H_0^{-1}}$, and a quadratic form controlled on V does not control the H_0^{-1}-norm of $H_R V$, which may leave V. For the ridge-softmax head of Section C.3.1 every $\nabla\ell_i$ has zero class-sum and stationarity gives $W\mathbf{1}_C = 0$, so differentiating in w leaves $V = \{U : U\mathbf{1}_C = 0\}$ invariant and every response lies in it, while the common-shift directions that make the unregularized cross-entropy Hessian singular do not. The restriction is structural and holds along the whole path, but instantiating α_R^V still requires a bound uniform over the tube, and a generalized eigenvalue at θ_0 is not one; the closed forms used in Appendix D are uniform because they bound H_R and H_0 spectrally rather than at a point.

Lemma C.17 (Path-uniform gradient diameter)**.** *Assume the containment of Theorem C.10, so that for every w on a radius-q chord the point $\theta(w)$ and the segment joining it to θ_0 lie in the tube and $\|\theta(w) - \theta_0\|_{H_0} \le R$, and assume* (135)*. With $\gamma_0 = \gamma(\theta_0, H_0)$, $\zeta_0 = \max_{i,j}\|A_i(\theta_0) - A_j(\theta_0)\|_{H_0,\mathrm{op}}$ and $H(w) \succeq \underline{h}_q H_0$,*

$$\gamma(\theta(w), H(w)) \le \bar{\gamma}_q := \frac{\gamma_0 + \zeta_0 R + \kappa_A R^2}{\sqrt{\underline{h}_q}}, \tag{153}$$

so Lemma B.3 supplies $\bar{z}_q = \bar{\gamma}_q/2$ for Theorem B.4. Only the pairwise Hessian diameter enters; no bound on an individual $\|A_i\|_{H_0,\mathrm{op}}$ is required.

Proof. For symmetric M and any x, $\|Mx\|_{H_0^{-1}} = \left\|H_0^{-1/2} M H_0^{-1/2} H_0^{1/2} x\right\|_2 \le \|M\|_{H_0,\mathrm{op}} \|x\|_{H_0}$. Fix i, j, write $\theta = \theta(w)$ and $\Delta = \theta - \theta_0$, and integrate the *difference* along the segment, which lies in the tube:

$$(\nabla\ell_i - \nabla\ell_j)(\theta) - (\nabla\ell_i - \nabla\ell_j)(\theta_0) = \int_0^1 (A_i - A_j)(\theta_0 + s\Delta)\,\Delta\, ds. \tag{154}$$

By the triangle inequality and (135), $\|(A_i - A_j)(\theta_0 + s\Delta)\|_{H_0,\mathrm{op}} \le \zeta_0 + 2\kappa_A s \|\Delta\|_{H_0} \le \zeta_0 + 2\kappa_A sR$, so the integrand has $\|\cdot\|_{H_0^{-1}}$ at most $(\zeta_0 + 2\kappa_A sR)R$ and the integral at most $\zeta_0 R + \kappa_A R^2$. Hence $\|(\nabla\ell_i - \nabla\ell_j)(\theta)\|_{H_0^{-1}} \le \gamma_0 + \zeta_0 R + \kappa_A R^2$. Finally $H(w) \succeq \underline{h}_q H_0$ gives $H(w)^{-1} \preceq \underline{h}_q^{-1} H_0^{-1}$, so every $\|\cdot\|_{H_0^{-1}}$ bound becomes a $\|\cdot\|_{H(w)^{-1}}$ bound after division by $\sqrt{\underline{h}_q}$. □

Proposition C.18 (A safe tube-range baseline)**.** *Any two reachable parameters θ_a, θ_b in the tube satisfy*

$$|R(\theta_a) - R(\theta_b)| \le V_{\mathrm{tube}} := 2a_0 R + L_R R^2. \tag{155}$$

If S is top-k under its exact self-score and

$$\max_q \frac{\overline{\beta}_q^2}{2\underline{m}_q} \left(\frac{q}{k}\right)^2 < V_{\mathrm{tube}}, \tag{156}$$

then the population geometry certificate is strictly smaller than this safe range.

Proof. For $\|\Delta\|_{H_0} \le R$, Taylor integration and (136) give

$$|R(\theta_0 + \Delta) - R(\theta_0)| \le a_0 R + \frac{L_R}{2} R^2. \tag{157}$$

Apply this to both endpoints and use the triangle inequality. Exact self-top-k implies $\Gamma_q(S) \le 0$ for every q, so Corollary C.14 and (156) finish the proof. □

Certified computational cost. The direct all-pairs evaluation of the two maxima in (122) performs $O(N^2)$ pair evaluations. An exact lower bound stronger than this statement requires a specified pair-query or structural model; no such unconditional runtime lower bound is asserted here. A safe pool-linear alternative chooses anchors $\bar{g}, \bar{A}$ and uses

$$\widehat{\gamma}_0 = 2 \max_i \|g_i - \bar{g}\|_{H_0^{-1}}, \qquad \widehat{\zeta}_0 = 2 \max_i \left\|A_i - \bar{A}\right\|_{H_0,\mathrm{op}}, \tag{158}$$

which upper-bound the exact diameters by the triangle inequality and require $O(N)$ certified norm calls. Given a derivative/spectral oracle returning rigorous enclosures, the remaining work is one base ERM solve, one reusable factorization or certified preconditioner for H_0, $O(N)$ sample derivatives, one solve for $H_0^{-1}\nabla R(\theta_0)$, $O(N \log N)$ sorting for all Γ_q, and $O(q_{\max})$ scalar checks. Unenclosed Lanczos or power-iteration point estimates are not safe spectral bounds.

C.3.1 A FULLY INSTANTIATED RIDGE-SOFTMAX REGIME

Let $W \in \mathbb{R}^{d \times C}$ and suppose $\|x_i\|_2 \le B$. Consider the fixed-feature multinomial model

$$\ell_i(W) = \log \sum_{c=1}^{C} e^{x_i^\top w_c} - x_i^\top w_{y_i},$$

$$L(W, w) = \sum_i w_i \ell_i(W) + \frac{\lambda}{2} \|W\|_F^2, \qquad \lambda > 0, \tag{159}$$

$$R(W) = \frac{\nu}{2} \|W - W_\star\|_F^2, \qquad \nu > 0. \tag{160}$$

This is a multivariate nonlinear ERM whose sample Hessians vary with both the sample and W.

Lemma C.19 (Ridge-softmax derivative constants)**.** *At any base point W_0, the assumptions of Theorems C.10–C.12 hold globally with*

$$\kappa = \frac{2B}{\sqrt{\lambda}}, \qquad \kappa_A = \frac{B^3}{\lambda^{3/2}}, \tag{161}$$

$$\gamma_0 \le \frac{2\sqrt{2}B}{\sqrt{\lambda}}, \qquad \zeta_0 \le \frac{B^2}{2\lambda}, \tag{162}$$

$$\alpha_R = \frac{\nu}{\lambda + B^2/2}, \qquad L_R = \frac{\nu}{\lambda}. \tag{163}$$

Proof. Let p be the softmax probability vector for one example and $J \sim p$. For directions U, V, put $a_c = x^\top u_c$ and $b_c = x^\top v_c$. Directional derivatives of log-sum-exp are the categorical cumulants

$$D^2\ell[U,U] = \mathrm{Var}(a_J), \qquad D^3\ell[U,U,V] = \mathbb{E}[(a_J - \mathbb{E}a_J)^2(b_J - \mathbb{E}b_J)]. \tag{164}$$

Since $|b_J - \mathbb{E}b_J| \le \mathrm{range}(b) \le 2B\,\|V\|_F$,

$$\left|D^3\ell[U,U,V]\right| \le 2B\,\|V\|_F\; D^2\ell[U,U]. \tag{165}$$

After weighted averaging, the data Hessian is dominated by the full ridge Hessian. Moreover, $H_0 \succeq \lambda I$. Hence

$$\left|D^3L[U,U,V]\right| \le \frac{2B}{\sqrt{\lambda}}\,\|V\|_{H_0}\; U^\top H(W,w)U, \tag{166}$$

which proves the asserted κ.

The categorical covariance $\mathrm{Diag}(p) - pp^\top$ has operator norm at most $1/2$ because its cth Gershgorin row radius plus diagonal equals $2p_c(1-p_c) \le 1/2$. Thus

$$D^2\ell[U,U] \le \frac{B^2}{2}\,\|U\|_F^2\,. \tag{167}$$

Combining this with the preceding third-derivative estimate gives

$$\left|D^3\ell[U,U,V]\right| \le B^3\,\|U\|_F^2\,\|V\|_F\,. \tag{168}$$

Integration from W_0 to W yields

$$\|A_i(W) - A_i(W_0)\|_{\mathrm{op}} \le B^3\,\|W - W_0\|_F\,. \tag{169}$$

Using $H_0 \succeq \lambda I$ on both the matrix norm and the displacement gives

$$\|A_i(W) - A_i(W_0)\|_{H_0,\mathrm{op}} \le \frac{B^3}{\lambda^{3/2}}\,\|W - W_0\|_{H_0}\,, \tag{170}$$

which proves the asserted κ_A.

The sample gradient is $x(p - e_y)^\top$ and has Frobenius norm at most $\sqrt{2}B$. Its pairwise Euclidean diameter is therefore at most $2\sqrt{2}B$, which becomes the first bound in (162) after conversion to the H_0^{-1} norm. Each sample Hessian is positive semidefinite, and (167) gives $0 \preceq A_i \preceq B^2 I/2$. For every pair,

$$-\frac{B^2}{2}I \preceq A_i - A_j \preceq \frac{B^2}{2}I, \qquad \|A_i - A_j\|_{\mathrm{op}} \le \frac{B^2}{2}.$$

Since $H_0 \succeq \lambda I$, congruence by $H_0^{-1/2}$ bounds both spectral signs by $B^2/(2\lambda)$, proving the stated bound on ζ_0.

Finally,

$$\lambda I \preceq H_0 \preceq (\lambda + B^2/2)I, \qquad H_R = \nu I, \tag{171}$$

which is equivalent to (163). □

Corollary C.20 (An open, strictly nonvacuous ridge-softmax region)**.** *Assume $B > 0$ and define*

$$x = \frac{B^2}{\lambda}, \qquad R = 4\sqrt{2}\sqrt{x}, \qquad z = \frac{B^2 a_0}{\nu\sqrt{x}}. \tag{172}$$

If

$$0 < x < \frac{1}{64}, \qquad 0 < z < \frac{1}{20}, \tag{173}$$

and S is top-k under its exact self-score, then the following hold for every feasible q.

1. *the containment condition* (130) *holds;*
2. $\underline{m}_q \ge \nu/(3\lambda) > 0$*;*

3. the population geometry certificate satisfies

$$U_{\mathrm{geom}} := \max_q \frac{\overline{\beta}_q^2}{2\underline{m}_q}\left(\frac{q}{k}\right)^2 < \frac{1}{100}V_{\mathrm{tube}}. \tag{174}$$

Thus this nonlinear multivariate model has an open parameter region on which the certificate is strictly better than the safe tube range.

Proof. Lemma C.19 gives $\zeta_0 \le x/2$. We use the conservative relaxation $\zeta_0 \le x$ in the estimates below, retaining the stated sufficient region and constants. Uniformly over q,

$$g_q \le 2\sqrt{2}\sqrt{x}, \qquad \delta_q \le x, \qquad \kappa R = 8\sqrt{2}x. \tag{175}$$

For $x \le 1/64$, elementary monotonicity gives the conservative inequalities

$$(1-x)e^{-8\sqrt{2}x} > \frac{4}{5}, \qquad (1+x)e^{8\sqrt{2}x} < \frac{5}{4}. \tag{176}$$

Consequently,

$$R(1-\delta_q)e^{-\kappa R} > 4\sqrt{2}\sqrt{x}\frac{4}{5} > 2\sqrt{2}\sqrt{x} \ge g_q. \tag{177}$$

Containment holds and $\underline{h}_q > 4/5$, $\overline{h}_q < 5/4$.

Since $L_R = \nu x/B^2$ and $a_0 = \nu\sqrt{x}\,z/B^2$,

$$a_q \le \frac{\nu\sqrt{x}}{B^2}(z + 4\sqrt{2}x). \tag{178}$$

Equations (140), (163), and (176) imply

$$\begin{aligned}\underline{m}_q &\ge \frac{\nu}{\lambda}\left[\frac{4}{5(1+x/2)} - \frac{5}{2}(z+4\sqrt{2}x)\right] \\ &\ge \frac{\nu}{3\lambda},\end{aligned} \tag{179}$$

where the last inequality follows by evaluating the monotone worst case on the closure $x \le 1/64$, $z \le 1/20$.

Here $\kappa_A = x^{3/2}$ in the canonical Frobenius coordinates. Hence

$$\zeta_q^{\mathrm{tube}} \le x + 2x^{3/2}(4\sqrt{2}\sqrt{x}) = x + 8\sqrt{2}x^2 < \frac{6}{5}x. \tag{180}$$

The beta envelope then gives

$$\overline{\beta}_q < 1.7\,\frac{\nu x^{3/2}}{B^2}(z + 4\sqrt{2}x). \tag{181}$$

Combining this inequality with (179) and $q/k \le 1$ yields

$$U_{\mathrm{geom}} < 4.34\,\frac{\nu x^2}{B^2}(z + 4\sqrt{2}x)^2. \tag{182}$$

On the other hand,

$$V_{\mathrm{tube}} = 2a_0R + L_RR^2 = \frac{\nu x}{B^2}(8\sqrt{2}z + 32x). \tag{183}$$

Therefore

$$\frac{U_{\mathrm{geom}}}{V_{\mathrm{tube}}} < \frac{4.34x(z+4\sqrt{2}x)^2}{8\sqrt{2}z + 32x} \le \frac{4.34}{32}(z+4\sqrt{2}x)^2 < 0.01 \tag{184}$$

on (173). This proves (174). □

Remark C.21 (Canonical-coordinate softmax constants). The trust-region statements themselves are affine invariant by Proposition C.13. In contrast, the raw Euclidean/Frobenius sufficient bounds B, λ, x, z in Lemma C.19 and Corollary C.20 are defined in the declared canonical last-layer coordinates and transform with that chart. After an affine change, the intrinsic base quantities are recomputed and Theorems C.10–C.12 apply directly. The corollary treats fixed-feature smooth last-layer training; feature learning, dropout, hard clipping, optimizer state, and support-dependent loading enter through the finite-step or endpoint-coupling modules.

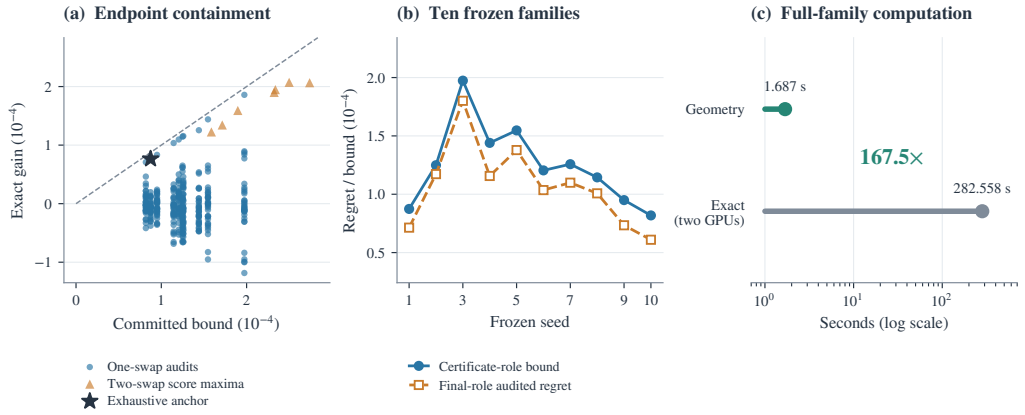

Figure 4: CIFAR-10. **(a)** Every audited endpoint falls below its committed bound, including the exhaustively retrained anchor. **(b)** Complete one-swap bounds against held-out endpoint checks across the ten frozen families. **(c)** Shared geometry against exact enumeration of the same neighbourhood.

D Prospective CIFAR-10 Large-Family Certificate

D.1 Frozen protocol and sequential roles

The prospective CIFAR-10 study (Krizhevsky, 2009) partitions the 50,000 training images by label-blind hash into candidate, proposal, development, certificate and unused roles of sizes 4,096, 4,096, 8,192, 32,768 and 848. The deduplicated 10,000 official test images form the final role; all roles are disjoint. ImageNet ResNet-18 (He et al., 2016) supplies fixed 512-dimensional average-pool features. Candidate-only PCA and unit-norm clipping precede deterministic float64 ridge-softmax training on $k = 512$ examples.

Ten seeds (74001–74010) construct class-balanced bases. Proposal labels choose one local-score exchange of radius 16 per base. Mean development cross-entropy selects $(d, \lambda) = (16, 0.3)$ from $d \in \{16, 32\}$ and $\lambda \in \{0.3, 1.0\}$; ties use accuracy, smaller dimension and larger ridge, in that order. Features, subsets and fits are frozen before certificate access.

D.2 Combinatorial certificate and untouched final audit

Target and theorem instantiation. Let C denote the fixed 32,768-image certificate role, and let W_T be the ridge-softmax ERM fit on candidate subset T. The certified target is

$$R_C(W) = \frac{1}{|C|} \sum_{(x,y)\in C} \left[\log \sum_c e^{x^\top w_c} - x^\top w_y \right] + \frac{\lambda}{2} \|W\|_F^2, \qquad \lambda = 0.3. \tag{185}$$

Thus a comparator's gain is $R_C(W_S) - R_C(W_T)$, and regret is its largest nonnegative value over the declared neighborhood. This is a fixed empirical target, not a population expectation; no target-sampling confidence term is included. Pure cross-entropy and top-1 accuracy are reported separately.

For the clipped feature bound B, both the training Hessian at the base and $\nabla^2 R_C(W)$ lie between λI and $(\lambda + B^2/2)I$. Consequently, Theorem C.12 applies with

$$\alpha_R = \frac{\lambda}{\lambda + B^2/2}, \qquad L_R = \frac{\lambda + B^2/2}{\lambda}, \qquad \kappa = \frac{2B}{\sqrt{\lambda}}, \qquad \kappa_A = \frac{B^3}{\lambda^{3/2}}. \tag{186}$$

The last two constants follow from the softmax derivative bounds in Lemma C.19; the target is the regularized empirical loss above, rather than that lemma's quadratic-target example.

For a returned subset, the comparator class contains every exchange through radius 32 for which the analytic ridge-softmax tube is valid. The sorted local maximum Γ_q is exact. Path-uniform geometry uses the fitted Hessian, exact candidate gradient diameter, relative sample-Hessian bound, and the

trust-region construction of Section 4. The implementation bounds the base Hessian diameter by $\max_i \|A_i\|_{H_0,\mathrm{op}}$. Since the sample Hessians are positive semidefinite, this quantity also bounds every relative difference $A_i - A_j$. It chooses the smallest geometric penalty among $2{,}500$ log-spaced tube radii satisfying the containment and positive-margin conditions. The stored constants give

$$R_1^{\mathrm{ub}} = \max\left\{0, \Gamma_1(S) + \frac{\overline{\beta}_1^2}{2\underline{m}_1 k^2}\right\}, \tag{187}$$

using Equations (140)–(141) with $R = R_C$ as target. These numbers evaluate the analytic exact-ERM bound in float64; certificate endpoint checks use a 10^{-10} comparison tolerance, with the exhaustive-fit residual and scalar checks reported below. The primary $q = 1$ neighborhood contains $512(4096 - 512) = 1{,}835{,}008$ non-base endpoints per seed. Before fitting any audit endpoint, each seed freezes its score-maximizing $q = 1$ exchange, its score-maximizing exchange at the largest certified radius, and 32 label-blind hash-selected $q = 1$ exchanges.

All ten seeds certify 0.02-regret over the complete $q = 1$ neighborhood. Their actual bounds are much smaller, ranging over $[8.18 \times 10^{-5}, 1.97 \times 10^{-4}]$, with mean 1.25×10^{-4}. Seven seeds certify through $q = 2$, whose cumulative family contains $839{,}936{,}835{,}585$ subsets including the base. All 337 frozen certificate endpoints satisfy their committed path bounds.

Only after the aggregate certificate lock returned GO was the final role opened. All ten selected models exceed the predeclared 0.55 top-1 floor; mean, minimum, and maximum accuracies are 0.64827, 0.60540, and 0.67700. Across the 330 frozen $q = 1$ final audits, the largest observed regret is 1.80×10^{-4}, below the predeclared 0.02 audit threshold. Here the final objective replaces C in Equation (185) by the final role, retaining the ridge penalty. This is a held-out check of the frozen audit endpoints, not a simultaneous bound over all final-role neighbors.

D.3 Full-enumeration anchor and cost

The predeclared anchor, seed 74001, is audited on the same target R_C. Two concurrent GPU shards fit all $1{,}835{,}008$ endpoints by float64 batched Newton iterations warm-started from the anchor. All residuals are at most 10^{-9} (observed maximum 9.94×10^{-10}); twenty independent scalar fits agree in objective and target gain to 4.44×10^{-16}. The largest gain is 7.6465×10^{-5} against the committed analytic bound 8.7409×10^{-5}, a ratio of 1.143. The two-device endpoint kernel takes 282.558 seconds, with 302.123 seconds including scheduling, loading, warm-up and validation. The analytic calculation takes 1.687 seconds. These are different hardware executions; the matched CPU comparison is in Appendix H.4.

D.4 The truncated envelope on the committed subsets

This retrospective comparison uses the same ten frozen subsets, features and candidate/certificate roles, with no final-role access. Both the positive-margin envelope of Equation (46) and the response-capped envelope of Theorem B.4 minimize their penalty over the same $4{,}000$ tube radii; the original prospective run used $2{,}500$. Lemma C.17 supplies $\bar{z}_q = \bar{\gamma}_q/2$. The capped penalty never exceeds the uncapped one when both apply.

For regularized cross-entropy, both bounds coincide at $q = 1$ on all ten seeds; the finer grid changes the original values by at most 9.414×10^{-7} relatively. The cap increases two-swap coverage from $7/10$ to $10/10$ and supplies bounds at $q = 4, 8$, where the positive-margin branch supplies none. Unregularized cross-entropy has negative checked margins at every tested radius and median $a_H \approx 0.32$, versus 0.05 for the regularized target. The cap still certifies all 160 cells (ten seeds, two targets, $q = 1, \ldots, 8$) at tolerance 0.02. Table 4 reports the bounds at every radius alongside a stability control.

D.5 Outward-rounded verification of the exhaustive anchor

We additionally verify seed 74001, the pre-existing exhaustive $q = 1$ anchor, with 80-bit `mpmath 1.3.0` interval arithmetic. The inputs are the frozen transformed binary64 feature rows interpreted as exact dyadic numbers, and the exact ridge coefficient $\lambda = 3/10$. The verifier uses a prespecified radius $R = 1/80$ and verifies the feature bound $B = 1.00000001$ on every candidate and certificate

row. It imports no experiment implementation. Approximate parameters and an approximate linear solve serve only as witnesses; their errors are bounded below rather than absorbed by a fixed guard.

From residuals to local scores. Let $\widehat{\theta}$ denote the fitted witness, θ_* the exact base ERM, and $\widehat{v}$ an approximate solution of the target-gradient linear system. Set $L = \lambda + B^2/2$. All gradients and Hessian–vector products at $\widehat{\theta}$ are evaluated using interval softmax probabilities, with exact dyadic input conversion and outward-rounded arithmetic. If ρ bounds the base KKT residual and ρ_v bounds $\|H(\widehat{\theta})\widehat{v} - \nabla R(\widehat{\theta})\|$, then

$$e_\theta = \rho/\lambda \;\geq\; \|\widehat{\theta} - \theta_*\|, \tag{188}$$

$$e_v = \frac{\rho_v + (8B^3\|\widehat{v}\| + L)e_\theta}{\lambda} \;\geq\; \|\widehat{v} - H(\theta_*)^{-1}\nabla R(\theta_*)\|. \tag{189}$$

The first inequality follows from strong convexity. For the second, $\|\nabla R(\theta_*) - \nabla R(\widehat{\theta})\| \leq Le_\theta$ and $\|H(\theta_*) - H(\widehat{\theta})\| \leq 8B^3 e_\theta$ bound the exact linear-system residual at θ_*. The latter constant follows by bounding the product of three centered categorical projections, each of magnitude at most twice its feature-direction norm; the ridge contributes no third derivative. Finally, $\|H(\theta_*)^{-1}\| \leq 1/\lambda$ converts that residual to e_v.

Each local sample score differs from $\nabla \ell_i(\widehat{\theta})^\top \widehat{v}$ by at most

$$e_h = \frac{B^2}{2} e_\theta \|\widehat{v}\| + \sqrt{2} B e_v, \tag{190}$$

using sample-gradient Lipschitz constant $B^2/2$ and gradient norm bound $\sqrt{2}B$. Taking the maximum interval upper endpoint outside S, subtracting the minimum interval lower endpoint inside S, and adding $2e_h$ before dividing by 512 gives a valid upper bound on the exact $\Gamma_1(S)$. Thus fitting, linear-solve and score-evaluation errors all propagate in the required direction.

From local inputs to all paths. Global ridge inequalities replace numerical eigendecompositions in this verifier. Since $0 \preceq A_i \preceq B^2 I/2$, every sample-Hessian difference lies between $-B^2 I/2$ and $B^2 I/2$. With $\lambda I \preceq H_0 \preceq LI$, valid inputs are

$$\begin{aligned} &\gamma_0 \leq 2\sqrt{2}B/\sqrt{\lambda}, && \zeta_0 \leq B^2/(2\lambda), \\ &\alpha_R = \lambda/L, && L_R = L/\lambda, \\ &a_0 \leq \big(\|\nabla R(\widehat{\theta})\| + Le_\theta\big)/\sqrt{\lambda}. \end{aligned} \tag{191}$$

Together with the softmax regularity constants in Equation (161), these inputs instantiate Theorem C.12. Interval evaluation verifies both $R\underline{h}_1 - \gamma_0/512 > 0$ and $\underline{m}_1 > 0$, so the tube contains every one-swap endpoint and supplies a valid all-path absorption bound. The final addition and division in Equation (11) are also outward rounded.

Table 3: Verified anchor inputs and resulting upper bound. Displayed upper bounds are rounded up and lower bounds down. No replacement endpoint is needed for this calculation.

Quantity	Verified bound
Base KKT norm ρ	$\leq 3.4890 \times 10^{-17}$
Base displacement e_θ	$\leq 1.1630 \times 10^{-16}$
Linear-system residual ρ_v	$\leq 8.7240 \times 10^{-17}$
Local maximum Γ_1	$\leq 8.7400 \times 10^{-5}$
Tube-containment slack	≥ 0.0018175
Absorption margin	≥ 0.0731518
Geometric penalty	$\leq 5.4775 \times 10^{-7}$
Complete one-swap regret	$\leq 8.7948 \times 10^{-5}$

The verified bound covers all $1{,}835{,}008$ one-swaps and is less than 0.62% above the original float64 analytic value. Its interval pass takes 67.21 seconds on one CPU thread, separately from the float64 timing comparison in Appendix H.4. This verifies one conservative anchor instantiation with global spectral bounds; the ten original float64 readouts retain their numerical-input scope.

D.6 A STABILITY BASELINE ON THE SAME TUBE

Theorem C.10 places every endpoint of a radius-q exchange inside the tube $\|\theta_T - \theta_S\|_{H_0} \le r$. For a convex target this alone gives a non-enumerative bound with no curvature analysis, since convexity and the dual-norm inequality yield

$$F(w_T) - F(w_S) \le -\nabla R(\theta_S)^\top (\theta_T - \theta_S) \le a_0\, r, \qquad a_0 = \|\nabla R(\theta_S)\|_{H_0^{-1}}. \tag{192}$$

Taking the smallest containment-feasible r under the constants of Appendix D.4 gives the comparison below on the same 160 cells (ten seeds, two targets, radii one to eight). The envelope is below the stability bound in every cell; its advantage narrows with the radius on the regularized target, from $3.0\times$ to $1.4\times$, as the response cap takes over from the positive-margin branch. On the unregularized target at radius eight the stability bound lies above the 0.02 tolerance for all ten subsets, with a median of 2.08×10^{-2}, while the envelope certifies every one of them with a maximum of 7.63×10^{-3}.

Table 4: Truncated envelope against the stability bound $a_0 r_{\mathrm{tube}}$ on the frozen CIFAR-10 subsets. Medians over ten seeds; the last column counts cells in which the envelope is the smaller bound.

target	q	envelope (median)	$a_0 r_{\mathrm{tube}}$ (median)	ratio	envelope tighter
ridge-regularized	1	1.23×10^{-4}	3.56×10^{-4}	2.98	10/10
ridge-regularized	2	2.40×10^{-4}	7.30×10^{-4}	3.10	10/10
ridge-regularized	3	3.80×10^{-4}	1.13×10^{-3}	3.04	10/10
ridge-regularized	4	5.60×10^{-4}	1.55×10^{-3}	2.81	10/10
ridge-regularized	5	8.14×10^{-4}	1.99×10^{-3}	2.45	10/10
ridge-regularized	6	1.19×10^{-3}	2.47×10^{-3}	2.08	10/10
ridge-regularized	7	1.75×10^{-3}	2.99×10^{-3}	1.71	10/10
ridge-regularized	8	2.59×10^{-3}	3.54×10^{-3}	1.37	10/10
unregularized CE	1	5.02×10^{-4}	2.08×10^{-3}	4.15	10/10
unregularized CE	2	1.05×10^{-3}	4.27×10^{-3}	4.06	10/10
unregularized CE	3	1.69×10^{-3}	6.59×10^{-3}	3.90	10/10
unregularized CE	4	2.42×10^{-3}	9.04×10^{-3}	3.74	10/10
unregularized CE	5	3.29×10^{-3}	1.17×10^{-2}	3.54	10/10
unregularized CE	6	4.36×10^{-3}	1.44×10^{-2}	3.31	10/10
unregularized CE	7	5.72×10^{-3}	1.75×10^{-2}	3.05	10/10
unregularized CE	8	7.49×10^{-3}	2.08×10^{-2}	2.77	10/10

D.7 EXHAUSTIVE RETRAINING ON SHRUNK INSTANCES

To distinguish search gaps from bound conservatism, three retrospective instances use 160 candidates (sixteen per class), balanced subsets of size 128, and the same frozen PCA, ridge and certificate target. Each retrains all $4{,}096$ one-swap and $4{,}031{,}488$ two-swap alternatives. At $q = 3, 4$, search covers respectively $48{,}400$ and $245{,}025$ combinations of the twelve best-scored examples on each side, plus $200{,}000$ random exchanges per radius, for $14{,}187{,}027$ endpoints in total. The search maxima at $q = 3, 4$ are lower bounds, not exhaustive regret.

Every evaluated gain is below its local prediction, by at least 8.4×10^{-6}. Table 5 retains every instance and radius. The growing penalty is conservative on the exhaustive families; the larger searched families likewise show no endpoint needing a positive correction, but do not determine the exact worst-case gap.

E PROSPECTIVE MOMENTUM OPTIMIZER-STATE CERTIFICATE

Model and declared comparison. This experiment instantiates the finite-step route with a useful classifier and an evolving optimizer state. Pinned MiniLM representations (Wang et al., 2020) from `all-MiniLM-L6-v2`[1] of AG News (Zhang et al., 2015) are projected by candidate-only PCA to 32 coordinates and subjected to a validated unit-ball guard. The classifier is $f_\theta(x) = A\tanh(Bx+b)+c$, with hidden width 16, four logits, and 596 trainable parameters. A fixed $1{,}024$-row candidate half trains the common initialization; the disjoint $1{,}024$-row half supplies selection candidates. All comparisons share this initialization and its PCA transform.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 5: Shrunk instances, $k = 128$, $n = 160$. "all" marks exhaustive families and "searched" the targeted-plus-random search. The best observed gain is exact only for exhaustive rows; for searched rows it is a lower bound on regret. "corr. share" is the correction's share of the envelope and "disp./r" the largest observed H_0-displacement over the tube radius.

seed	q	family	Γ_q	envelope	best observed	env./obs.	corr. share	disp./r
91001	1	all	6.11×10^{-4}	7.54×10^{-4}	4.32×10^{-4}	1.7	19%	0.91
91001	2	all	1.21×10^{-3}	2.74×10^{-3}	8.51×10^{-4}	3.2	56%	0.70
91001	3	searched	1.78×10^{-3}	1.05×10^{-2}	1.19×10^{-3}	8.8	83%	0.45
91001	4	searched	2.30×10^{-3}	5.85×10^{-2}	1.46×10^{-3}	40.0	96%	0.32
91002	1	all	9.62×10^{-4}	1.12×10^{-3}	7.89×10^{-4}	1.4	14%	0.89
91002	2	all	1.86×10^{-3}	3.43×10^{-3}	1.39×10^{-3}	2.5	46%	0.66
91002	3	searched	2.67×10^{-3}	1.14×10^{-2}	1.83×10^{-3}	6.2	76%	0.51
91002	4	searched	3.30×10^{-3}	5.64×10^{-2}	2.17×10^{-3}	25.9	94%	0.36
91003	1	all	9.15×10^{-4}	1.09×10^{-3}	7.39×10^{-4}	1.5	16%	0.88
91003	2	all	1.76×10^{-3}	3.51×10^{-3}	1.34×10^{-3}	2.6	50%	0.67
91003	3	searched	2.54×10^{-3}	1.23×10^{-2}	1.87×10^{-3}	6.6	79%	0.50
91003	4	searched	3.28×10^{-3}	6.78×10^{-2}	2.36×10^{-3}	28.7	95%	0.35

For a weight vector w, the eight-step map is

$$v_{t+1} = 0.9v_t + \nabla_\theta \sum_i w_i \ell_i(\theta_t) + (\theta_t - \theta_0),$$
$$\theta_{t+1} = \theta_t - 0.01 v_{t+1}, \qquad v_0 = 0, \quad t = 0, \dots, 7. \tag{193}$$

Thus the state in the finite-step theorem is $z_t = (\theta_t, v_t)$. Seeds 76001–76010 generate ten balanced size-256 bases. Proposal-role finite-step hypergradients define a label-preserving $q = 4$ exchange, with stable index-order tie breaking. Each declared family contains the base and its selected endpoint. All twenty parameter arrays and subsets are frozen before fresh target access. The endpoint utility is negative cross entropy, and the certified regret is that of selected within its declared two-endpoint family.

Fixed target population and sequential access. Excluding prior training, selection and evaluation rows in this study, together with additional confirmed uses, leaves $N = 54{,}287$ eligible AG News training rows. This fixed pool is the inferential population. All its embeddings, PCA features and unit-ball guards are computed and frozen before uniform random allocation without replacement into disjoint certificate and final samples of $n = 16{,}384$ rows each. The remaining $21{,}519$ rows are unused. The frozen feature population makes each endpoint's per-row utility independent of the realized allocation. Certificate access follows the model, protocol and row locks; final access follows the aggregate certificate decision. Both samples are training-source holdouts, distinct from the official-test audit in Appendix F.1.

Nonzero path enclosure and numerical transfer. We propagate outward-rounded scalar-path second-order jets through both parameters and momentum buffers in Equation (193). Each directed chord starts with 16 uniform cells; a single 32-cell recomputation is permitted only if its first interval width exceeds 0.016. For each cell $[a_j, b_j]$, the stored interval $[l_j, u_j]$ encloses F'' throughout that cell. With $c_j = (b_j - a_j)(1 - (a_j + b_j)/2)$, the enclosure is

$$F(1) - F(0) \in [\underline{F}'_0, \overline{F}'_0] + \sum_j c_j [l_j, u_j]. \tag{194}$$

Base-to-selected and negated selected-to-base intervals are intersected to give one interval for $\Delta = U_{\text{selected}} - U_{\text{base}}$. These intervals are committed before target endpoint diagnostics.

Separate interval trajectories at the two endpoint weights enclose the exact-real parameters initialized from the stored binary values. A uniform input-bound argument transfers the resulting utility interval to the serialized deployment parameters. For two tanh networks, define

$$r_j = \min\{2, \|\Delta B_j\|_2 + |\Delta b_j|\}, \qquad d_i = \sum_j |\Delta A_{ij}| + \sum_j |A_{\text{end},ij}| r_j + |\Delta c_i|.$$

Then every logit difference is at most $D = \max_i d_i$, and every cross-entropy difference is at most $2D$. Both orientations give valid bounds; we use their smaller outward-rounded upper bound. Applied

Table 6: Complete momentum formal ledger. Width is the larger of the two directed path widths. Certificate and final columns are finite-pool regret upper bounds; accuracy and CE use the final role. All ten return CERTIFY; none returns IMPROVE.

Seed	Path width	Certificate	Final bound	Accuracy (%)	CE
76001	0.002153	0.004080	0.002974	85.272	0.452896
76002	0.003124	0.004018	0.002430	85.278	0.452791
76003	0.005402	0.007061	0.004541	85.193	0.452947
76004	0.005732	0.006469	0.003725	85.364	0.451802
76005	0.006542	0.007469	0.004251	85.303	0.452236
76006	0.001258	0.002745	0.002062	85.315	0.452307
76007	0.001392	0.002652	0.001878	85.345	0.452094
76008	0.007470	0.008507	0.004963	85.254	0.453086
76009	0.002928	0.004778	0.003364	85.291	0.452245
76010	0.007567	0.007811	0.004105	85.358	0.452379

between ideal parameter boxes and stored endpoints, this gives allowances $\eta_{\mathrm{base}}, \eta_{\mathrm{selected}}$. Their sum expands the path interval and is at most 9.24×10^{-12} across the ten families. Applied between the two stored endpoints, the same calculation gives paired-utility support $[-2D, 2D]$ and range $R = 4D$ before target access.

Simultaneous finite-population target bound. For the current $n < N/2$, Theorem 4.3 of Bardenet & Maillard (2015) yields the one-sided radius

$$r_{\mathrm{tgt}} = \sqrt{\frac{2\overline{v}\rho_n \log(1/\delta)}{n}} + \left(\frac{7}{3} + \frac{3}{\sqrt{2}}\right)\frac{R\log(1/\delta)}{n}, \qquad \rho_n = 1 - \frac{n-1}{N}, \tag{195}$$

with failure probability at most 5δ. Here $\overline{v}$ is an outward upper bound on the sample variance. For interval-valued paired row utilities Y_i and any fixed binary midpoint c of their sample-mean interval, we use $\sum_i \sup_{y \in Y_i} |y - c|^2/(n-1)$; this also upper-bounds the n-denominator variance in that theorem. Endpoint evaluations supply the variance bound and containment diagnostics; Equation (194) supplies the certificate interval.

Each stage allocates $\alpha = 0.025$ to the two sides of ten paired comparisons, using $\delta = \alpha/(5 \cdot 20)$ rounded downward. Adding $[-r_{\mathrm{tgt}}, r_{\mathrm{tgt}}]$ to the numerical-transfer interval gives $[L, U]$ for the fixed-pool mean Δ. The outputs are

$$\overline{\mathrm{Reg}} = \max\{0, -L\}, \qquad \underline{\mathrm{Imp}} = L.$$

The marginal uniform sampling laws and a union bound give joint, unconditional two-stage failure probability at most 0.05, without requiring independence of the disjoint samples or conditional coverage after the first gate.

Frozen gates. Certificate decisions require valid ideal/deployed endpoint containment and both directed widths at most 0.016. At least eight of the ten families must receive 0.02-CERTIFY, with every completed containment audit valid, to open final. The final task-quality gate is accuracy at least 0.55 and cross entropy at most 1 in at least eight of the same ten seeds. Formal success additionally requires at least eight issued certificates with a final-stage regret bound at most 0.02 and no issued certificate contradicted by an empirical final regret above 0.02. All seeds are retained, including any numerical or resource ABSTAIN.

Results and cost. All 20 directed paths pass at 16 cells without refinement. Their nonzero widths range from 0.001255 to 0.007567; every ideal/deployed endpoint diagnostic is contained and every paired row interval lies within its support bound. All ten certificate and final regret bounds meet 0.02 (Table 6). Final mean accuracy is 85.29724% and mean cross entropy 0.45247830; all ten task-quality gates pass.

Two CPU workers complete certification in 1000.807 seconds and final audit in 131.021 seconds. Family construction takes 45.828 seconds separately; feature preparation is excluded. Independent high-precision reconstruction checks cell aggregation, target radii, numerical transfers and every decision.

F FINITE-TRAINING AND TARGET-EVIDENCE AUDITS

Table 7 maps the retained experiments to their certificate components.

Table 7: Theory–experiment map for the validated regimes. Each row names the ledger component, the experiment that exercises it directly, and the regime in which it was instantiated; no row asserts coverage of any other row's regime.

Ledger component	Direct validation	Instantiated regime
shared-geometry envelope	CIFAR-10, 1,835,008 one-swap and 8.399×10^{11} two-swap alternatives	image-domain ridge-softmax endpoint
actual nonconvex finite-step map	AG News, $50/50$ containment	trainable tanh representation, four GD steps
optimizer-state propagation	AG News momentum, $20/20$ directed paths	eight updates, two endpoints per family, finite-population target
nonzero path enclosure	HelpSteer3, $20/20$ containment	smooth deterministic convex branch
finite target evidence	HelpSteer3	declared target-sampling model
finite-step coupling	convex-head contraction stress	frozen-head gradient descent

F.1 PROSPECTIVE NONCONVEX FINITE-TRAJECTORY CERTIFICATE

This study uses four gradient steps and a different model and target population from the eight-step momentum experiment in Appendix E.

Frozen roles and actual training map. Before role allocation or nonlinear-model outcomes, the protocol hash-freezes AG News (Zhang et al., 2015) text roles using a label-blind salted hash after normalized-text deduplication. Candidate, proposal, development, certificate, and final contain respectively 2,048, 4,096, 8,192, 32,768, and 7,590 rows; the final rows come from the official test split and exclude normalized overlap with train. Pinned MiniLM embeddings (Wang et al., 2020) from `all-MiniLM-L6-v2` are reduced to eight coordinates by candidate-only PCA and clipped to unit norm. The trained model is

$$h(x) = \tanh(Bx + b), \qquad f(x) = Ah(x) + c,$$

with hidden width four and four logits. All entries of A, B, b, c are updated, so the hidden representation changes and the parameterization is nonconvex. For subset weights w, the endpoint is exactly K deterministic full-batch gradient steps on weighted cross entropy plus $\lambda\|\theta - \theta_0\|_2^2/2$.

Development gate and frozen controls. Development compares $K \in \{1, 2, 4\}$, $\eta \in \{0.01, 0.025, 0.05\}$, $\lambda \in \{0.1, 0.3, 1\}$ and $q \in \{8, 16, 32\}$ on three seeds, then propagates intervals for the three best configurations. It selects $(K, \eta, \lambda, q) = (4, 0.05, 1, 16)$; all 27 development chords contain the endpoint differences.

N1 freezes ten new class-balanced size-256 bases. Each family contains the finite-step local endpoint, the corresponding one-step GLISTER-style endpoint, a LESS-style cosine-gradient endpoint, and a matched-radius random endpoint. Direct development search among these endpoints is also recorded. Before certificate access, finite-local is the direct winner for five seeds, LESS-style for four, and GLISTER-style for one; finite-local mean development family regret is 2.03×10^{-7}. The lock fixes all 50 endpoint subsets before the certificate role is decoded.

Simultaneous certificate. For every seed we enclose base-to-finite-local and the four finite-local-to-control chords using 64 uniform cells. A scalar second-order interval jet is propagated through all four gradient steps; the full cellwise second-derivative bounds are atomically committed before exact certificate endpoints are evaluated. All 50 exact differences are contained. The median deterministic width is 4.03×10^{-8} and the maximum is 1.56×10^{-7}.

For paired row difference V_i, its deterministic support range R follows from $\|x\|_2 \le 1$, $|\tanh(\cdot)| \le 1$, and the two endpoint parameter norms. With $M = 50$, $n = 32{,}768$, and $\delta = 0.05$, each chord uses

$$r = \sqrt{\frac{2\widehat{v}\log(2M/\delta)}{n}} + \frac{7R\log(2M/\delta)}{3(n-1)}.$$

The 50 predeclared directional events use one lower bound for base-to-finite-local and four upper bounds from finite-local to its controls per seed. Applying the one-sided finite-class bound of Maurer & Pontil (2009) to these oriented functions gives joint coverage at least 0.95 for the certificate decisions. The 50 target radii have median 0.002081 and range $[0.001876, 0.002408]$, dominating path width. The ten family-regret upper bounds range from 0.002040 to 0.002415, so all ten return CERTIFY at $\epsilon = 0.02$ and none abstains. Their exact empirical regrets are at most 2.26×10^{-5}.

Untouched final audit and cost. The N1 decision authorizes a single official-test evaluation with no reselection. A descriptive check places all 50 final endpoint differences inside the stored two-ended bands, and all 50 final bands intersect them. These band checks are not 95% simultaneous two-sided coverage claims. All ten finite-local family regrets lie below 0.02 (range $[7.56 \times 10^{-6}, 3.04 \times 10^{-5}]$). The finite-local-over-base aggregate difference is -1.82×10^{-5} with a 95% one-sided lower bound of -0.004413; improvement is not confirmed. The confirmatory outcome is the regret certificate. Appendix H records the distinction between directional guarantees and historical band diagnostics.

The ten enclosure jobs take 797–867 seconds each (median 858). Endpoint construction and 50 development fits take 7.3 seconds; embedding all roles takes 14.5 GPU seconds after model loading. Independent reconstruction checks the development and certificate containments, all decisions and final comparisons.

F.2 PROSPECTIVE COMPLETE HELPSTEER3 CERTIFICATE

Data roles. We use the preference subset of HelpSteer3 (Wang et al., 2025). The composite path and 4,096-row certificate size were chosen using an opened Qwen2.5-7B (Qwen Team, 2024) design study. The confirmation then froze its protocol, ten seeds, families, comparisons and gates before accessing fresh target rows. Excluding all rows and normalized prompt groups used in eight prior builds leaves 6,136 rows in 6,001 groups. Candidate, proposal, certificate and final roles contain 768, 128, 4,096 and 896 rows, with disjoint prompt groups and one observation per prompt in the two target roles. The thirty endpoints are fixed before certificate access; final access follows the certificate decision.

Frozen family and target. Full-chat final-layer, last-nonpadding Qwen2.5-7B-Instruct representations are reduced by a candidate-only randomized PCA to 32 coordinates plus an intercept and scaled by the maximum candidate norm. Each of ten seeds fixes a size-128 base subset of the 768 candidates. Using only the 128 proposal rows, the rule constructs one natural-gradient exchange of radius 64 and one matched-radius random exchange. The gradient endpoint is selected before certificate access. The smooth training map uses labels $y_i \in \{-1, 1\}$ and projected features x_i including the intercept. Its weighted objective and population target are

$$L(\theta, w) = \frac{1}{2} \sum_i w_i (x_i^\top \theta - y_i)^2 + \frac{0.03}{2} \|\theta\|_2^2 ,$$
$$R(\theta) = \mathbb{E}_{(x,y)} \log(1 + \exp(-y x^\top \theta)). \tag{196}$$

The ridge term includes the intercept coefficient. The squared residual equals $(y_i x_i^\top \theta - 1)^2$, so both losses are smooth. The empirical target averages the logistic loss over its declared role. Projected target features are clipped at norm one; no certificate or final feature is actually clipped (maximum preclip norms 0.8748 and 0.8208).

Complete interval ledger. For each base–gradient and gradient–random chord, we apply Equation (200) on a uniform 65-node grid. The implicit derivatives use the exact ridge-squared Hessian. An analytic third-derivative norm bound supplies each cell's $h^3 M/12$ quadrature remainder, and the stored ledger also adds solve residuals and a 10^{-10} roundoff guard. All twenty exact endpoint gains fall within the resulting path intervals. A deterministic full-batch gradient descent endpoint, initialized at zero and run for 2,048 steps at learning rate 0.5, is the declared finite training operator. Its maximum final gradient norm is 4.51×10^{-16} and its maximum parameter distance from the exact endpoint is 7.55×10^{-15}; one-Lipschitz target loss converts these distances to the finite training coupling term.

For the independent certificate rows, a paired gain has absolute bound $\|\theta_a - \theta_b\|_2$ under unit feature norm. For M predeclared directional events, the one-sided empirical-Bernstein bound (Maurer &

Pontil, 2009) uses the allowance

$$r(\widehat{v}, B, n, M, \delta) = \sqrt{\frac{2\widehat{v}\log(2M/\delta)}{n}} + \frac{14B\log(2M/\delta)}{3(n-1)}, \tag{197}$$

with $M = 20$, $n = 4{,}096$, and $\delta = 0.05$. Each seed uses the gradient-over-base lower bound and the random-over-gradient upper bound, so these twenty events jointly support the decisions with probability at least 0.95 under the declared IID target model. The allowance does not assert simultaneous two-sided coverage of all twenty stored bands. For every seed, subtracting path, target, numerical, and finite-training errors leaves a positive gradient-over-base population lower bound; the ten bounds range from 0.012834 to 0.017286. The simultaneous upper bound for random over gradient is negative in every seed. Consequently the maximum of zero and the base/random regret bounds is exactly zero for all ten frozen triples. All ten decisions are therefore both IMPROVE and CERTIFY at $\epsilon = 0.02$, exceeding the predeclared $8/10$ and $8/10$ gate without seed extension.

Independent final validation and reconstruction. The final role cannot alter these decisions. It checks only whether at least eight gradient-over-base mean gains are positive and whether a 95% one-sided empirical-Bernstein lower bound for the per-row average gain over ten seeds is positive. All ten seed means are positive, ranging from 0.02340 to 0.03106, and all ten gradient endpoints have zero observed regret within their frozen triples. The aggregate gain is 0.027692, its radius is 0.024977, and its lower bound is 0.002715, so the final gate passes. Certificate and final confidence statements are stage-specific, not a joint 95% two-stage claim. Independent reconstruction checks representations, endpoints, path containments, confidence radii and final losses.

For selected-over-base chords, median total, target-sampling, and quadrature debts are 0.012190, 0.011329, and 0.000455; for selected-over-random they are 0.011175, 0.010768, and 0.000406.

The certificate covers deterministic convex heads on fixed representations over the ten declared three-member families.

F.3 CERTIFICATE DIAGNOSTICS AND REFUSAL BEHAVIOR

The following retrospective diagnostics reuse frozen endpoints and cannot change the confirmatory decisions. Figure 3(b,c) presents their target-sampling and training-discrepancy effects.

Target-statistical power and abstention. The target-sampling diagnostic uses the already opened, prompt-disjoint 896-row final role and reproduces its endpoint losses.

Five hundred fixed-seed nested permutations vary only the empirical-Bernstein target-statistical subledger for the frozen finite endpoints. For $n \in \{64, 128, 224, 448\}$, every seed abstains. At $n = 896$, all ten seeds can certify 0.02-optimality but none can certify improvement, and the joint family gate still fails. The star at $n = 4{,}096$ is the separate original confirmatory certificate role, not a final-role resample. Because path curvature is not recomputed on each subsample, this is a power diagnostic rather than a new path certificate or repeated-coverage claim. These observations give no universal minimum target size. Power depends on the gain relative to the decision threshold, the paired variance, the support range and the number of simultaneous comparisons.

Decomposing statistical power. For a paired gain bounded in absolute value by B_j, the support range is $2B_j$. The directional empirical Bernstein allowance in Appendix H separates into

$$r_j = \underbrace{\sqrt{2\widehat{v}_j\log(2M/\delta)/n}}_{r_{j,\mathrm{var}}} + \underbrace{\frac{14B_j\log(2M/\delta)}{3(n-1)}}_{r_{j,\mathrm{range}}}.$$

We decompose the existing ledgers with $M = 20$ and $\delta = 0.05$. Table 8 retains all five sample sizes and the separate confirmation role. The support term supplies a median 80.0% of the radius at $n = 896$. It decreases as $1/(n-1)$, while the variance term decreases as $1/\sqrt{n}$ when the other quantities are held fixed. Reducing variance alone leaves the support term unchanged.

Let $\overline{Y}_b$ denote the selected gain over base and $\overline{Y}_r$ the random gain over selected. Improvement requires $r_b < \overline{Y}_b$. Certification permits $r_b \le \epsilon + \overline{Y}_b$ and $r_r \le \epsilon - \overline{Y}_r$. Thus the same uncertainty

can permit certification while leaving improvement unresolved, as observed for all ten families at $n = 896$.

The comparison count enters through $\log(2M/\delta)$. Holding the statistics of the complete final role fixed gives median radii 0.024150, 0.040909 and 0.049713 for $M = 1, 20, 100$. The first concerns a single comparison and cannot replace the guarantee covering twenty directions. These are algebraic sensitivities of the existing ledger. They do not establish a universal sample threshold. The diagnostic variance term is reconstructed as the total radius minus its known support term; the confirmation records the variance directly.

Table 8: Existing ledgers for target sampling. Values are medians over both comparison types, ten families and, for diagnostic rows, 500 nested permutations. Each column is summarized separately. C/I gives the mean count of certification and improvement decisions out of ten. The 4096-row confirmation has a separate role and includes its original path and numerical allowances.

n and role	r_{var}	r_{range}	r	Range share	C/I
64, diagnostic	0.03026	0.46189	0.49282	94.0%	0/0
128, diagnostic	0.02154	0.22913	0.25105	91.5%	0/0
224, diagnostic	0.01641	0.13049	0.14714	89.0%	0/0
448, diagnostic	0.01167	0.06510	0.07690	85.0%	0/0
896, diagnostic	0.00826	0.03251	0.04091	80.0%	10/0
4096, confirmation	0.00380	0.00711	0.01091	65.4%	10/10

Finite-step coupling stress. The frozen candidate-only scaling implies $\|x\|_2 \le 1$. Hence every ridge-squared endpoint Hessian has spectrum in $[0.03, 1.03]$, and GD with step size 0.5 contracts from zero toward the exact endpoint at worst by $\rho = \max_{\lambda\in[0.03,1.03]} |1 - 0.5\lambda| = 0.985$. For an endpoint a, therefore, $\|\theta_a^{(K)} - \theta_a^\star\|_2 \le \rho^K \|\theta_a^\star\|_2$; summing the two endpoint errors gives a valid comparison-level target coupling. Reinserting this term into the original path and target ledger yields $10/10$ abstentions through 304 steps, $9/10$ ϵ-certificates at 320, $10/10$ certificates but no improvements at 336, $2/10$ and $8/10$ joint decisions at 352 and 368, and $10/10$ joint decisions from 384 steps. At $2{,}048$ steps, every saved observed parameter error lies below the spectral bound. This is a rigorous worst-case stress for the deterministic convex head.

G Comparator-wise path enclosures and related tools

G.1 Two computational uses of the path identity

The shared bound of Theorem B.4 uses one set of geometric constants for an entire neighborhood. A comparator-wise enclosure instead evaluates each path and can retain a negative integral correction. Its cost scales with the number of paths and its accuracy with the integration budget.

Proposition G.1 (Signed midpoint enclosure)**.** *Let $f \in C^4([0,1])$. Partition $[0,1]$ into cells of width h_j and midpoint c_j, and assume $|f^{(4)}(t)| \le M_{4,j}$ throughout cell j. Define*

$$J_j = (1-c_j)h_j f''(c_j) - \frac{h_j^3}{12} f'''(c_j), \qquad r_j = \frac{(1-c_j)h_j^3}{24} M_{4,j},$$
$$J = \sum_j J_j, \qquad r = \sum_j r_j. \tag{198}$$

Then

$$f(1) - f(0) - f'(0) \in [J - r, J + r]. \tag{199}$$

Consequently $f'(0) + J + r$ upper-bounds the endpoint gain. The smallest symmetric interval containing the same remainder enclosure has radius $|J|+r$, and gives the upper bound $f'(0)+|J|+r$. The latter exceeds the former by $2\max\{0, -J\}$. Symmetrizing each cell first gives the possibly larger radius $\sum_j |J_j| + r$.

Proof. On a cell, Taylor's formula gives $f''(t) = f''(c_j) + f'''(c_j)(t - c_j) + e_j(t)$ with $|e_j(t)| \le M_{4,j}(t - c_j)^2/2$. Integrating the first two terms against $1 - t$ gives J_j. The integral of the absolute

error is at most $M_{4,j}(1-c_j)h_j^3/24$; the odd term vanishes around the midpoint. Sum the cells and use Equation (6). Finally, $\max\{|J-r|,|J+r|\} = |J|+r$ for $r \geq 0$, and $|\sum_j J_j| \leq \sum_j |J_j|$. □

This comparison uses the same enclosure in both directions of error control. For a finite family, form each gain bound before maximizing over alternatives and zero; a correction from one chord cannot be assigned to another.

First-derivative integration is an alternative when those evaluations are more convenient. For $0 = t_0 < \cdots < t_J = 1$, $h_j = t_{j+1} - t_j$, $|\widehat{f}'_j - f'(t_j)| \leq \eta_j$ and a cellwise bound $|f'''| \leq M_j$, composite trapezoidal integration gives

$$\widehat{\Delta}_{\mathrm{path}} = \sum_j \frac{h_j}{2}(\widehat{f}'_j + \widehat{f}'_{j+1}),$$

$$|f(1) - f(0) - \widehat{\Delta}_{\mathrm{path}}| \leq \sum_j \frac{h_j}{2}(\eta_j + \eta_{j+1}) + \sum_j \frac{h_j^3 M_j}{12}. \tag{200}$$

The first term propagates derivative uncertainty and the second controls quadrature. Either integration form requires valid bounds across complete cells.

G.2 Relationship to existing certification tools

Table 9 distinguishes the guarantee objects. The common-Hessian case recovers the Frank–Wolfe gap (Jaggi, 2013); heterogeneous training Hessians require the additional curvature analysis of Section 4.

Table 9: Guarantee objects and inputs of related tools. Each comparison retains the assumptions of the cited result.

Tool	Existing guarantee	Role of the present analysis
Frank–Wolfe gap	Convex objective suboptimality from a linear minimization oracle	Recovers the concave-utility boundary case; heterogeneous curvature requires an additional remainder.
AMIP	Uniform finite-change approximation control for influential deletion sets	Such uniform control already supports decisions; we retain the sign of the decision-relevant remainder for fixed-budget exchanges.
Higher-order IJ	Smooth reweighted fits approximated at arbitrary order with finite-sample accuracy bounds	A strict Taylor predictor must include a valid remainder. Our numerical control does not stand for every HOIJ implementation.
Self-concordant influence diagnostics	Finite-sample and computational bounds for influence functions and AMIP	Localization is a shared tool; the present certificate explicitly combines target curvature and heterogeneous training responses.

Uniform approximation bounds already support decisions (Broderick et al., 2020; Giordano et al., 2019a; Fisher et al., 2023). The direct comparisons in Appendices H.4 and J.3 report implemented second-order and IJ bounds separately.

H Directional Confidence and Decision Diagnostics

H.1 Confidence events used by the recorded decisions

The decision depends on particular directions of comparison, fixed before target access. For paired gain Y_j with range length R_j, empirical variance $\widehat{v}_j$ and n IID target observations, let

$$r_j = \sqrt{\frac{2\widehat{v}_j \log(2M/\delta)}{n}} + \frac{7R_j \log(2M/\delta)}{3(n-1)}.$$

For M fixed signs $s_j \in \{-1, 1\}$, Theorem 4 and Corollary 5 of Maurer & Pontil (2009), applied to the normalized signed gains, imply

$$\Pr\{s_j(\mathbb{E}Y_j - \overline{Y}_j) \le r_j \text{ for all } j\} \ge 1 - \delta.$$

The sign $+1$ supplies an upper bound and -1 a lower bound. Dependence between comparisons through shared target rows is allowed; the union bound does not require independent seeds. Endpoint families and signs must be fixed independently of those target observations, and deterministic path, numerical and training allowances must also be valid.

Table 10: Events supporting the frozen decisions. Failure budgets are per stage unless a joint budget is explicitly stated. Lower/upper refer to the named gain.

Study	Required events	Failure budget
HelpSteer3	Certificate has 10 gradient-over-base lowers and 10 random-over-gradient uppers. Final has one seed-averaged gain lower.	0.05 per stage
Four-step AG News	Certificate has 10 base-to-selected lowers and 40 selected-to-control uppers. Final aggregate gain lower is separate.	0.05 per stage
Eight-step momentum	Each stage has both directions of 10 fixed-pool comparisons, with the finite-population bound of Appendix E.	0.025 per stage; 0.05 jointly

For HelpSteer3, negating the gradient-over-base lower bound supplies the base-over-gradient regret upper bound; no second event is needed. AG News assigns an event to each of its five declared directions per seed. Thus Table 10 retains the original allowances and decisions.

The listed events are directional. Simultaneous two-sided coverage of all M gains would require $\log(4M/\delta)$ instead. Stage-specific 95% bounds do not imply a joint 95% statement over all stages. Other endpoints of stored bands are descriptive diagnostics.

H.2 Tolerance sensitivity of the finite training certificates

The following readout uses only the existing ten four-step AG News and ten eight-step momentum certificate records. The full tolerance grid contains zero, every recorded regret upper bound and the original 0.02 tolerance. It is a post-outcome description of the frozen bounds, not a new choice of confirmatory threshold. Each curve shares target data across its ten families. The two studies also have different inferential targets, namely the declared IID target model for AG News and the fixed eligible population for momentum.

All four-step certificates are available by tolerance 0.00242 and all momentum certificates by 0.00851, using rounded-up maxima of the existing bounds. This describes their resolution below the original 0.02 gate without changing it. Four-step target radii dominate deterministic path widths, whereas momentum has visible path and sampling contributions. More accurate integration alone would not remove the statistical bottleneck in the former. These diagnostics do not establish positive improvement or tightness at the much smaller empirical-regret scale.

H.3 From error budgets to certification and abstention

The experiments instantiate the same decision logic using different controls on the training map and target. For a declared oriented gain g_j between two endpoints, let $[a_j, b_j]$ enclose its empirical surrogate value using the path calculation, let r_j be its target sampling allowance, and let η_j and c_j bound numerical and discrepancies from deployed training. The corresponding valid directional bounds are

$$L_j = a_j - r_j - \eta_j - c_j, \qquad U_j = b_j + r_j + \eta_j + c_j. \tag{201}$$

Each allowance is charged once; an interval already expanded for a numerical or training error requires no second charge for that error. For endpoint discrepancies Δ_a, Δ_b, one can take $c_j = \Delta_a + \Delta_b$, recovering the uniform 2Δ of Equation (13).

Certification uses upper bounds for alternatives over the returned subset, taking their maximum with zero and comparing it with ϵ. Improvement uses a positive lower bound for a declared replacement

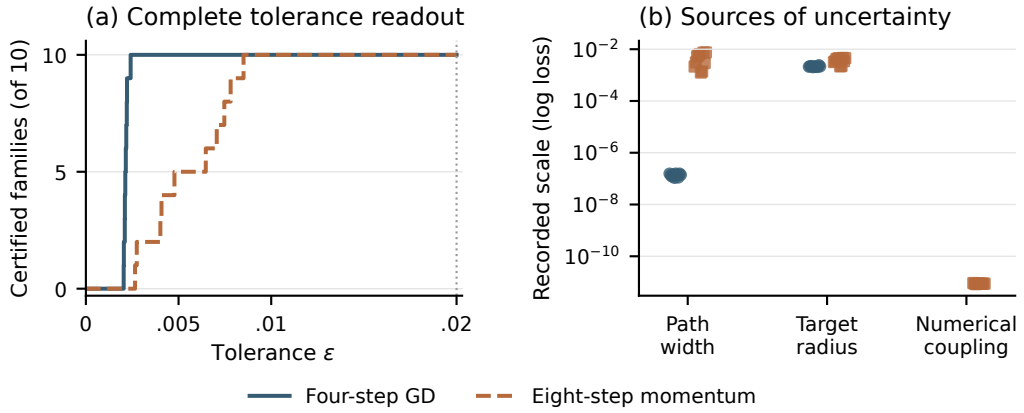

Figure 5: Existing finite-training certificates. **(a)** All tolerance breakpoints, with the original 0.02 threshold marked. **(b)** Recorded per-family scales. Four-step GD uses the largest width and target radius among its five chords; momentum uses its intersected paired interval and target radius. Widths, radii and numerical coupling are different quantities and are not added as a stacked error budget. A separate coupling allowance is not recorded for four-step GD and is not plotted as zero.

over its reference. Negating a valid lower bound supplies an upper bound for the reverse comparison; Table 10 specifies the directions available in each study. A claim whose required inequality fails remains unresolved.

For example, HelpSteer3 seed 19301 at $n = 896$ has mean gain 0.031057 and lower bound -0.014464: improvement remains unresolved but the full family regret upper bound 0.014464 permits 0.02 certification. Appendix F.3 explains the sampling terms; the finite-GD stress adds a spectral endpoint error to the same budget. The direct four-step and momentum studies instead enclose their update maps. Table 7 identifies each experiment's role.

H.4 Independent end-to-end CPU comparison

A subsequent post-outcome protocol compares the analytic and strict second-order implementations in separate CPU processes. Each of the ten original seeds is run three times per method, alternating order by seed and repetition. Both use one BLAS thread, the same fixed CPU affinity, identical candidate and certificate roles, and the complete $1{,}835{,}008$ one-swap neighborhood. Every execution independently loads and transforms the features, fits the base head from zero, constructs its required derivatives and spectral inputs, and searches the same $2{,}500$-radius grid. The second-order process additionally constructs the target Hessian, third-derivative contraction and complete quadratic-pair maximum. The analytic process does not perform those extra operations.

Table 11: Independent float64 executions. Bounds are in units of 10^{-5}; times are medians of three complete process wall times, including Python startup, loading, preprocessing, fitting, derivatives and search.

Seed	Analytic bound	Taylor2 bound	Analytic seconds	Taylor2 seconds
74001	8.7409	7.8527	1.545	1.846
74002	12.4884	11.6497	1.536	1.845
74003	19.7360	18.7993	1.688	1.996
74004	14.4024	12.7520	1.540	1.845
74005	15.4701	14.6137	1.580	1.844
74006	12.0436	11.1718	1.536	1.846
74007	12.5764	11.7600	1.538	1.850
74008	11.4449	10.5705	1.549	1.879
74009	9.4982	8.6063	1.534	1.884
74010	8.1847	7.2936	1.539	1.846

All sixty executions reproduce the bounds to absolute tolerance 10^{-10}. Analytic median wall times range from 1.534 to 1.688 seconds and second-order times from 1.844 to 1.996 seconds. The median of the ten paired time ratios is 1.2004 in favor of the analytic implementation. Second order is tighter on all ten seeds; both certify the fixed 0.02 tolerance on all ten. For the exhaustive anchor, the upper/truth ratios are 1.143 and 1.027. Thus this implementation of shared geometry saves computation at the cost of additional slack, rather than dominating second order on both axes.

The operating-system page cache is not flushed, and background contention is not excluded; the archive includes all repetitions, ranges, per-stage wall/CPU times and process affinity. Historical representation extraction is shared and excluded from both methods, whereas feature loading and transformation are charged to each. These are float64 implementation timings, separate from the 67.21-second outward-rounded anchor verification in Appendix D.5.

I Endpoint stationarity and shared matrix containment

The endpoint equation gives another way to control gains omitted by a local score. This route uses a fixed target direction and a common region containing the alternative endpoints. The theorem along data weight paths uses derivatives along retrained solution paths. We state the endpoint conditions separately and retain both routes when their respective conditions can be verified.

I.1 A bound from the endpoint equation

Fix a size-k anchor S, an exchange count q, and $\tau = q/k$. For an alternative T, write

$$D_T(\theta) = \frac{1}{q}\left(\sum_{i\in T\setminus S} \ell_i(\theta) - \sum_{i\in S\setminus T} \ell_i(\theta)\right), \qquad L_T = L_S + \tau D_T.$$

At a stationary anchor θ_0, let $H_0 = \nabla^2 L_S(\theta_0) \succ 0$, $r_0 = \nabla R(\theta_0)$ and $v = H_0^{-1} r_0$. The predicted gain is $p_T = \tau v^\top \nabla D_T(\theta_0)$. For a fixed positive definite metric G, put $\|u\|_G^2 = u^\top G u$ and $\mathcal{B}_G(\rho) = \{\theta_0 + u : \|u\|_G \le \rho\}$. The metric remains fixed throughout the following statement.

Proposition I.1 (Endpoint bound)**.** *Suppose L_S is C^3 and every D_T and R are C^2 on an open neighborhood of $\mathcal{B}_G(\rho_q)$. Assume the training rule returns a stationary endpoint $\theta_T \in \mathcal{B}_G(\rho_q)$ for every declared q-exchange. Suppose, throughout this ball and uniformly over those exchanges,*

$$\begin{aligned} \nabla^2 R(\theta) &\succeq \alpha_q G, \\ D^3 L_S(\theta)[v,h,h] &\le \kappa_q \|h\|_G^2 \quad \text{for every } h, \\ \|\nabla^2 D_T(\theta) v\|_{G^{-1}} &\le \eta_q, \qquad \eta_q \ge 0. \end{aligned} \tag{202}$$

There is no sign restriction on α_q or κ_q. Set $m_q = \alpha_q - \kappa_q$. Then

$$R(\theta_0) - R(\theta_T) \le p_T + \Psi_q, \qquad \Psi_q = \max_{0\le r\le \rho_q}\left\{-\frac{m_q r^2}{2} + \tau\eta_q r\right\}. \tag{203}$$

If $m_q > 0$, the maximizing radius is $\min\{\rho_q, \tau\eta_q/m_q\}$; if $m_q \le 0$, ρ_q is a maximizer.

Proof. Put $u = \theta_T - \theta_0$ and $r = \|u\|_G$. The two stationary equations give

$$\begin{aligned} 0 &= H_0 u + e_S(u) + \tau\nabla D_T(\theta_0) + \tau\{\nabla D_T(\theta_0+u) - \nabla D_T(\theta_0)\}, \\ e_S(u) &= \int_0^1 (1-s) D^3 L_S(\theta_0 + su)[u,u,\cdot]\, ds. \end{aligned}$$

Multiplication by $v^\top$ expresses $-r_0^\top u$ as p_T plus the two residual terms. Their upper bounds are $\kappa_q r^2/2$ and $\tau\eta_q r$, respectively. The latter follows from

$$v^\top\{\nabla D_T(\theta_0+u) - \nabla D_T(\theta_0)\} = \int_0^1 u^\top \nabla^2 D_T(\theta_0 + su) v\, ds.$$

The target expansion gives $R(\theta_0) - R(\theta_T) \le -r_0^\top u - \alpha_q r^2/2$. Combining these inequalities and maximizing over r proves the claim. The base Taylor residual has a factor $1/2$; the integral of the exchange gradient has total weight one. □

Let $s_i = v^\top \nabla \ell_i(\theta_0)$ and let Γ_q be the sum of the q largest outside scores minus the q smallest inside scores, divided by k. The regret over at most Q exchanges, including S, satisfies

$$\mathrm{Reg}_{k,\le Q}(S) \le \max\left\{0, \max_{1\le q\le Q}(\Gamma_q + \Psi_q)\right\}. \tag{204}$$

For a restricted family, the same sorted maximum is a valid relaxation. The scalar r in (203) is a parameter displacement. It differs from the response energy variable in the theorem along data weight paths. In particular, its regional constants require their own verification.

For quadratic training with a common Hessian, $\kappa_q = \eta_q = 0$ is valid. A convex target then permits $\Psi_q = 0$. If $r_0 = 0$ and the target has negative curvature, the formula still retains $\max\{0, -\alpha_q \rho_q^2/2\}$. Thus zero scores alone do not imply zero regret. When both routes are valid for the same training rule, target and family, their upper bounds may be minimized separately for each q before taking the maximum over q. No general dominance between the two routes is required.

I.2 Containment from integrated radial curvature

The stationary endpoints required above can be contained without fitting each alternative. The next argument uses convexity of the training objectives. It therefore applies to the ridge heads used below; a nonconvex training rule requires a separate containment argument.

Lemma I.2 (Radial containment)**.** *Suppose each L_T is globally convex and coercive and is C^2 near $\mathcal{B}_G(\rho)$. If, uniformly over T and $\|u\|_G = \rho$,*

$$\|\nabla L_T(\theta_0)\|_{G^{-1}} \le b_q, \qquad \int_0^1 u^\top \nabla^2 L_T(\theta_0 + su)u\, ds \ge h_q \|u\|_G^2, \qquad \rho h_q > b_q,$$

then every global minimizer lies strictly inside the ball.

Proof. The outward radial derivative at the boundary is at least $\rho^2 h_q - \rho b_q > 0$. A minimizer over the compact ball cannot lie on its boundary, since moving toward the center would decrease the objective. It is therefore an interior stationary point and a global minimizer by convexity. An exterior global minimizer would produce a segment of global minimizers crossing the boundary, where stationarity contradicts the positive radial derivative. □

The quantity h_q bounds an integrated radial Hessian. A bound along data weight paths containing an inverse Hessian still requires a pointwise lower bound. For example, logistic curvature with anchor margin zero and displacement two has radial average about 0.1904, but endpoint curvature about 0.1050. These two bounds cannot be interchanged.

I.3 Shared matrix lower bounds

Work in fixed Euclidean coordinates centered at the anchor. Suppose $A_i(\rho) \succeq 0$ is a lower bound on the integrated Hessian of example i on every radial segment in the ball. For a constant ridge Hessian $Q_0 \succ 0$, define $B = Q_0 + k^{-1}\sum_{i\in S} A_i$. The integrated Hessian of any alternative is bounded below by

$$B - \frac{1}{k}\sum_{i\in S\setminus T} A_i. \tag{205}$$

The added examples have positive semidefinite Hessians, so omitting them preserves this lower bound.

Lemma I.3 (Two matrix relaxations)**.** *Either construction below gives a common matrix K below* (205) *for every deletion set of size q.*

1. *Choose a with $C = B - aI \succ 0$ and a nonsingular W such that $WCW^\top \succeq cI$, $c > 0$. Let $\ell_i \ge \lambda_{\max}(WA_iW^\top)$ and let p be the sum of the q largest ℓ_i over S, divided by k. The matrix $K = B - (p/c)C$ is valid.*
2. *For $t \ge 0$, choose $U_i \succeq 0$ with $U_i \succeq A_i - tI$. The matrix $K = B - k^{-1}(qtI + \sum_{i\in S} U_i)$ is valid.*

Proof. For a deletion sum $D = k^{-1}\sum_{i\in S\setminus T} A_i$, the first construction gives $WDW^\top \preceq pI \preceq (p/c)WCW^\top$. Congruence by W^{-1} yields $D \preceq (p/c)C$. For the second construction, $kD \preceq qtI + \sum_{i\in S\setminus T} U_i \preceq qtI + \sum_{i\in S} U_i$. □

The first construction does not assume $p/c < 1$; the resulting K must still pass a containment test. The second permits $U_i = (A_i - tI)_+$ in exact arithmetic. With approximate spectral factors $V\Lambda V^\top$, choose

$$U_i = V\Lambda_+ V^\top + \varepsilon I, \qquad \varepsilon \geq \|A_i - tI - V\Lambda V^\top\|_{\mathrm{op}}.$$

Both required matrix inequalities hold even when the columns of V are not exactly orthogonal. The residual bound must itself be verified.

Proposition I.4 (Joint gradient and curvature containment)**.** *Let K be a common lower bound on the integrated Hessian. Suppose $h > 0$ and $Q = K - hI \succ 0$. For every exchange put $g_T = \nabla D_T(0)$. If $g_T^\top Q^{-1} g_T \leq \gamma_Q^2$ uniformly and*

$$h\rho^2 > \frac{\tau^2\gamma_Q^2}{4}, \tag{206}$$

then the convex endpoints lie inside the ball.

Proof. At $\|u\| = \rho$, the outward derivative is at least

$$u^\top K u + \tau u^\top g_T = h\rho^2 + \left\|Q^{1/2}u + \frac{\tau}{2}Q^{-1/2}g_T\right\|^2 - \frac{\tau^2}{4}g_T^\top Q^{-1}g_T > 0.$$

Apply the boundary argument of Lemma I.2. □

To compute γ_Q, verify $WQW^\top \succeq cI$, which implies $Q^{-1} \preceq W^\top W/c$. For any common center b, the sum of the top q values of $\|W\nabla\ell_i(0) - b\|$ on each exchange side, divided by q, bounds $\|Wg_T\|$. Dividing the square of this bound by c supplies γ_Q^2. This construction uses sorted example quantities and certified matrix operations; it does not enumerate deletion subsets. With $K = \lambda I$ and $h = \lambda/2$, the condition reduces to $\rho\lambda > \tau\gamma$ for a Euclidean gradient bound γ.

I.4 Constants for the implemented heads

For logistic regression, let z_i be a signed feature vector and b_i its anchor margin. Define $s(b) = \sigma(b)\sigma(-b)$ and $w_i = \rho\|z_i\|$. A radial curvature lower bound is

$$c_i = \frac{\sigma(-|b_i|) - \sigma(-|b_i| - w_i)}{w_i}, \qquad c_i = s(b_i) \text{ when } w_i = 0.$$

Indeed $s(b_i + tu^\top z_i) \geq s(|b_i| + tw_i)$ and integration gives c_i. For grouped observations, average $c_i z_i z_i^\top$ within each group to obtain the matrices used in (205). Pointwise target curvature bounds supply α_q. An upper matrix bound on the signed coefficients $s'(b_i + u^\top z_i)(z_i^\top v)$ supplies κ_q. The coefficient extrema include the critical points $\pm\log(2+\sqrt{3})$ whenever they lie in the margin interval.

For the exchange action, suppose $A_i^- \preceq \nabla^2\ell_i \preceq A_i^+$ pointwise in the ball. Put $C_i = (A_i^+ + A_i^-)/2$ and $E_i = A_i^+ - A_i^- \succeq 0$. Then

$$\|\{\nabla^2\ell_i - C_i\}v\| \leq \frac{1}{2}\sqrt{\|E_i\|_{\mathrm{op}}\, v^\top E_i v}. \tag{207}$$

To see this, express the matrix difference as $E_i^{1/2} J E_i^{1/2}$ with $\|J\|_{\mathrm{op}} \leq 1/2$ and use continuity if E_i is singular. Centered norms of $C_i v$, augmented by these errors and summed over the top q examples on each side, give η_q after division by q.

For softmax regression in fixed coordinates, let J_i map parameter displacements to logits, $A_{i0} = J_i^\top \mathrm{Cov}(p_{i0})J_i$, and $d_i = \max_{a,b}\|J_i^\top(e_a - e_b)\|$. On a radius-$\rho$ ball, probability ratios lie between $e^{-d_i\rho}$ and $e^{d_i\rho}$. The variational identity $\mathrm{Var}_p(a) = \min_c \sum_j p_j(a_j - c)^2$ transfers these factors to the Hessian in positive semidefinite order. Integration replaces $e^{-d_i\rho}$ by $(1 - e^{-d_i\rho})/(d_i\rho)$ in the radial lower bound. Pointwise derivative bounds retain the exponential factors.

Writing $a = J_i v$ and $b = J_i h$, the directional third derivative is $\mathbb{E}_p[(a - \mathbb{E}_p a)(b - \mathbb{E}_p b)^2]$. Its magnitude is bounded by $\mathrm{range}(a)\,\mathrm{Var}_p(b)$. A signed bound transports the anchor matrix for the third derivative using its fourth derivative, whose mixed cumulant is bounded by $(7/4)\,\mathrm{range}(a) d_i\,\mathrm{Var}_p(b)$ for a unit fourth direction. The centered fourth moment contributes one, and the three covariance products contribute $1/4 + 1/2$, in these units. These formulas supply regional constants without observing alternative retrained endpoints.

I.5 Residuals, numerical verification and cost

For a computed anchor $\widehat{\theta}_0$ with residual $g = \nabla L_S(\widehat{\theta}_0)$, apply the endpoint proposition to $L_T(\theta) - g^\top \theta$ for every T. The shift makes the anchor stationary and preserves all Hessians and exchange differences. If the original and shifted objectives are globally λ-strongly convex, corresponding minimizers differ by at most $\|g\|/\lambda$. A target Lipschitz constant C on all connecting segments therefore adds $2C\|g\|/\lambda$ to the gain bound. For a ridge target, C must be justified on a bounded region; it is not automatically a global constant.

This residual transfer addresses optimization error. Machine verification additionally encloses derivatives, solves, spectra, sorted sums, containment margins and the final scalar bound. The detection experiments perform these operations with 128-bit Arb for the stored features. Neural feature extraction remains outside that enclosure. The CIFAR endpoint reanalysis uses conditional float64 evaluation. Its numerical tier is stated separately in Appendix J. Shared geometry removes enumeration of alternative subsets, while derivative construction and verified matrix operations still depend on parameter dimension, data size and the declared radius grid.

I.6 Optimizing the response in the endpoint bound

The stationary equation also permits a response vector p chosen to reduce the upper bound. Its mismatch with $H_0^{-1} r_0$ then contributes a controlled linear term. Fix invertible M, write $\theta = \theta_0 + M^\top u$, and set $f(u) = L_S(\theta_0 + M^\top u)$, $\mathcal{R}(u) = R(\theta_0 + M^\top u)$ and $d_T(u) = \tau D_T(\theta_0 + M^\top u)$. All derivatives and norms in the next proposition use these fixed Euclidean coordinates. In particular, $H_0 = \nabla^2 f(0)$ is retained explicitly even when M is computed by approximate whitening.

Proposition I.5 (Endpoint bound with an arbitrary response)**.** *Suppose f is C^3, $\mathcal{R}$ and every d_T are C^2 near $\{u : \|u\| \le \rho_q\}$, and $\nabla f(0) = 0$. Every declared q-exchange returns a stationary endpoint u_T in this ball. For a fixed vector p, suppose symmetric matrices K_p, H_R and $\eta_p \ge 0$ satisfy, throughout the ball and uniformly over these exchanges,*

$$\nabla^3 f(u)[p, \cdot, \cdot] \preceq K_p, \qquad \nabla^2 \mathcal{R}(u) \succeq H_R, \qquad \|\nabla^2 d_T(u) p\| \le \tau \eta_p. \tag{208}$$

Put $r = \nabla \mathcal{R}(0)$, $a = H_0 p - r$ and $\Gamma_q(p) = \max_T p^\top \nabla d_T(0)$. For any $\mu \ge 0$, $\zeta > 0$ with $D = H_R - K_p + (2\mu - \zeta) I \succ 0$,

$$\mathcal{R}(0) - \mathcal{R}(u_T) \le B_q := \Gamma_q(p) + \mu \rho_q^2 + \frac{1}{2} a^\top D^{-1} a + \frac{(\tau \eta_p)^2}{2\zeta}. \tag{209}$$

Choosing a valid triple (p, μ, ζ) separately at each q gives $\mathrm{Reg}_{k,\le Q}(S) \le \max\{0, \max_{q \le Q} B_q\}$.

Proof. Set $u = u_T$. Expanding the training gradient at zero and using stationarity, then multiplying by $p^\top$ and expanding the target, gives the exact identity

$$\begin{aligned} \mathcal{R}(0) - \mathcal{R}(u) = p^\top \nabla d_T(0) + a^\top u \\ + \int_0^1 (1 - t) u^\top \{\nabla^3 f(tu)[p, \cdot, \cdot] - \nabla^2 \mathcal{R}(tu)\} u \, dt \\ + \int_0^1 p^\top \nabla^2 d_T(tu) u \, dt. \end{aligned}$$

The regional inequalities bound this by $\Gamma_q(p) + a^\top u - u^\top (H_R - K_p) u/2 + \tau \eta_p \|u\|$. Add the nonnegative quantity $\mu(\rho_q^2 - \|u\|^2)$ and apply $\tau \eta_p \|u\| \le \zeta \|u\|^2/2 + (\tau \eta_p)^2/(2\zeta)$. Completing the square bounds $a^\top u - u^\top D u/2$ by $a^\top D^{-1} a/2$. Maximizing over exchanges and retaining the anchor's zero gain proves the regret statement. □

Relation to the fixed response. With $p = H_0^{-1} r$, the mismatch vanishes. The inequality before completing the square gives the baseline $\Gamma_q(p) + \max_{0 \le t \le \rho_q} \{-m_q t^2/2 + \tau \eta_p t\}$, where m_q is a verified lower bound on $\lambda_{\min}(H_R - K_p)$. This retains the joint matrix difference in the scalar reduction. The arbitrary response may improve the bound by balancing the sorted support against the mismatch cost. Any verified proposal is valid, including one from an optimizer that has not reported convergence. The implemented certificate takes the smaller valid bound at each q before maximizing over q and zero.

Uniform construction and numerical transfer. For $g_i = \nabla_u \ell_i(\theta_0 + M^\top u)|_{u=0}$, the full exchange support is

$$\Gamma_q(p) = \frac{1}{k} \left(\sum_{q \text{ largest } i \notin S} g_i^\top p - \sum_{q \text{ smallest } i \in S} g_i^\top p \right).$$

The logistic coefficient bounds and Equation (207) apply with p throughout. A further action bound uses all cross pairs. If $\|x_i - C_i p\| \le e_i$ for $x_i = \nabla^2 \ell_i p$, set $b_{ij} = (\|C_i p - C_j p\| + e_i + e_j)^2$. For added and removed sets A, B of size q,

$$\left\| \sum_{i \in A} x_i - \sum_{j \in B} x_j \right\|^2 \le \sum_{i \in A, j \in B} \|x_i - x_j\|^2 \le \sum_{i \in A, j \in B} b_{ij}.$$

The first inequality drops the nonnegative squared distances within each side. Summing the q largest entries in each row, then the q largest row sums, relaxes the final maximum. The analogous column calculation is also valid. Their minimum, followed by a square root and division by q, bounds η_p without restricting the candidate sets.

Containment is verified independently of p using the preceding matrix tests. The common objective shift and $2C\|g\|/\lambda$ transfer above handle an inexact anchor. The verifier recomputes the regional constants, sorted support, positive definiteness of D, and quadratic solve with interval arithmetic. The proposition holds in any fixed invertible coordinates; the numerical search procedure has no asserted affine invariance.

J Additional evidence on coverage, prediction and cost

The experiments connect each reported outcome to its mathematical object. CIFAR and detection heads test shared bounds over exchange neighbourhoods; the direct IJ comparison evaluates fixed alternatives under the same exact training endpoints. Modern P1 tests observed gains after the declared AdamW updates, and midpoint feedback tests prediction accuracy on preselected exchanges. Table 12 records these training rules, targets and comparison families. The sampling and finite training decision experiments are mapped separately in Table 7. A fixed empirical target is the loss on its declared stored rows. Statistical claims about a population or a finite population require the corresponding sampling layer in Section 5.2 and the relevant experimental protocol.

Table 12: Objects and guarantee scopes of the additional evidence. The original validated regimes are listed in Table 7.

Setting	Training object	Comparator and target	Evidence tier
CIFAR reanalysis	Exact ridge-softmax endpoint, 160 parameters	All exchanges up to $Q \le 8$, fixed empirical CE or regularized CE	Conditional analytic bounds evaluated in float64
Detection heads	Exact ridge-logistic endpoint, 17 parameters on frozen Qwen features	All exchanges up to declared Q, fixed empirical grouped logistic loss	128-bit Arb upper enclosures for stored features
Direct IJ comparison	Same exact CIFAR endpoints	160 fixed P3 witnesses, same two targets	Conditional analytic remainders and residual bounds
Modern P1 and midpoint	Last four layers, q/v LoRA of rank eight, 64 AdamW updates	122 audited exchanges per arm; twelve fixed prediction cases; empirical audit NLL	Observed retraining gains and empirical prediction errors

J.1 Selection and retraining on modern models

We audit Qwen3 (8B base) (Qwen Team, 2025) and OLMo3 (7B, release 1025) (Team Olmo, 2025) with an implementation of LESS-style selection on a restricted pool and a fixed random selector.

Each selects 256 of 1024 instruction examples. The audited training rule initializes q/v LoRA (Hu et al., 2022) of rank eight in the last four layers, freezes all other weights, and takes 64 AdamW updates (Loshchilov & Hutter, 2019) over two epochs. The first batch's eight slots remain fixed in both epochs. This restriction was declared before observing outcomes for these models. The selector's checkpoint aggregation and Adam representations follow the pinned LESS implementation, with the pool and suffix adaptation stated explicitly in the protocol.

All four arms completed 122 counterfactual retrainings. For single exchanges, the cohort contains the 24 proposals with largest predictions and eight random proposals. Each of $q \in \{2, 8, 32\}$ has 30 random proposals. Table 13 reports every arm. Both LESS-style anchors have lower audit NLL than their random counterparts, and each arm has a positive observed improvement witness. Each arm also contains at least one underprediction by its local score. These observations support improvement on the audit target and motivate remainder control; downstream test performance is a separate measurement.

Table 13: All modern P1 arms. Gains use the same audit NLL as the anchor. Underprediction counts use tolerance 10^{-8}.

Model	Selector	Anchor NLL	Best observed gain	Underpredictions
Qwen3	LESS	0.801137	0.004491	56/122
Qwen3	Random	0.833153	0.008586	35/122
OLMo3	LESS	0.994742	0.001224	1/122
OLMo3	Random	1.027342	0.003174	7/122

Table 14: Complete modern P1 cohorts. Bootstrap intervals use 10000 paired resamples within each cohort and describe exchange variation at a fixed anchor. The mixed $q = 1$ cohort has a different sampling design from the random cohorts.

Model	Selector	q	n	Spearman	95% interval	MAE	Underpredictions
Qwen3	less	1	32	0.969	$[0.902, 0.991]$	2.69×10^{-4}	13
Qwen3	less	2	30	0.958	$[0.863, 0.992]$	9.24×10^{-5}	12
Qwen3	less	8	30	0.980	$[0.931, 0.993]$	2.08×10^{-4}	19
Qwen3	less	32	30	0.893	$[0.738, 0.958]$	6.75×10^{-4}	12
Qwen3	random	1	32	0.852	$[0.639, 0.955]$	1.40×10^{-3}	4
Qwen3	random	2	30	0.980	$[0.922, 0.995]$	2.40×10^{-4}	12
Qwen3	random	8	30	0.972	$[0.913, 0.992]$	3.59×10^{-4}	11
Qwen3	random	32	30	0.954	$[0.857, 0.987]$	1.41×10^{-3}	8
OLMo3	less	1	32	0.363	$[-0.048, 0.731]$	6.73×10^{-4}	0
OLMo3	less	2	30	0.840	$[0.560, 0.988]$	2.26×10^{-4}	1
OLMo3	less	8	30	0.940	$[0.826, 0.986]$	4.89×10^{-4}	0
OLMo3	less	32	30	0.907	$[0.779, 0.959]$	2.09×10^{-3}	0
OLMo3	random	1	32	0.832	$[0.642, 0.932]$	1.76×10^{-3}	2
OLMo3	random	2	30	0.980	$[0.918, 1.000]$	8.40×10^{-5}	3
OLMo3	random	8	30	0.834	$[0.625, 0.939]$	5.80×10^{-4}	2
OLMo3	random	32	30	0.910	$[0.781, 0.959]$	1.67×10^{-3}	0

Affine transfer under the original training rule. The fixed transfer comprises twelve cases, with one exchange selected by hash at $q = 2, 8, 32$ for each model and selector. It retains the original initialization, zero Adam moments, 64 updates and fixed first batch. Every arm freezes its three predictions before reading its retraining outcomes. Table 15 reports the same twelve cases; the method and cases were frozen before the P1 outcomes became available. The primary method propagates exact Adam updates with affine gradient feedback and evaluates the full audit loss at its predicted endpoint.

The primary MAE falls by 41.2% on Qwen3 and 84.4% on OLMo3. Only OLMo3 passes the frozen requirement that each model's MAE fall to at most 25% of its original value. The overall transfer screen therefore fails. Underprediction occurs in six of six Qwen3 cases and four of six OLMo3 cases. The two $q = 32$ cases contribute 90.5% of Qwen3's total affine absolute error. The control using a linearized target has lower MAE on Qwen3 and higher MAE on OLMo3 than the primary method. Controls do not replace the frozen primary test. These observations locate approximation sensitivity; they do not isolate the true trajectory error because the trajectories of retrained endpoints

other than the best witnesses were not saved. The completed midpoint comparison uses the same twelve cases (Appendix J.1).

Midpoint feedback on the fixed cases. The supplemental protocol freezes midpoint feedback and reuses the original twelve cases selected by hash, with $q = 2, 8, 32$ for each model and selector. The method and case lock predates availability of the corresponding P1 outcomes; each arm also freezes all three predictions before reading its retraining results. The comparison retains the original initialization, zero Adam moments, 64 updates and fixed first batch.

Let θ_t be the stored anchor trajectory, δ_t its predicted parameter displacement, and g_t^T, H_t^T the gradient and Hessian of the exchanged batch. Midpoint feedback uses

$$\widehat{g}_t^T = g_t^T(\theta_t) + H_t^T(\theta_t + \delta_t/2)\delta_t$$

inside the original AdamW update, propagating the optimizer moments together with the parameter displacement. Affine feedback evaluates this Hessian at θ_t. The primary prediction evaluates the full audit NLL at the resulting endpoint; the controls additionally evaluate a linearized target.

Table 15: All twelve fixed midpoint cases in protocol order. Values are in units of 10^{-3} audit NLL. Actual gains reuse the original P1 retraining endpoints; the remaining gain columns are predictions. Error is midpoint prediction minus actual gain.

Model	Selector	Case	Actual	Original	Affine	Midpoint	Error
Qwen3	LESS	q2_28	−1.09054	−1.15122	−1.12884	−1.09034	+0.00020
Qwen3	LESS	q8_23	−1.75107	−1.73329	−1.76096	−1.75228	−0.00121
Qwen3	LESS	q32_1	0.17916	1.09242	−0.80431	0.20869	+0.02952
Qwen3	Random	q2_13	0.07254	−0.34430	−0.00238	0.07266	+0.00013
Qwen3	Random	q8_25	1.99626	0.98587	1.81271	2.00046	+0.00420
Qwen3	Random	q32_29	−2.42732	−5.49192	−4.36088	−2.32131	+0.10601
OLMo3	LESS	q2_7	0.17941	0.24698	0.20543	0.17955	+0.00014
OLMo3	LESS	q8_15	−1.33047	−1.04703	−1.32555	−1.33131	−0.00085
OLMo3	LESS	q32_26	−6.78594	−3.41425	−7.38410	−6.78002	+0.00592
OLMo3	Random	q2_2	0.84271	0.97044	0.82769	0.84277	+0.00006
OLMo3	Random	q8_26	−0.44629	0.19489	−0.53195	−0.48391	−0.03762
OLMo3	Random	q32_3	0.41325	3.76955	−0.08042	0.35002	−0.06323

Across six cases per model, midpoint MAE is 2.3545×10^{-5} on Qwen3 and 1.7968×10^{-5} on OLMo3, giving reductions of 95.6% and 91.2% relative to affine feedback. Both meet the frozen requirement that midpoint MAE be at most 25% of affine MAE, and all twelve predicted gains have the correct sign. Tables 15 and 16 retain every case and declared control. These are paired empirical comparisons on the original training seed for each model.

Table 16: Midpoint comparison and all declared controls, with six cases per model. MAE and maximum absolute error are in units of 10^{-3} audit NLL. AT, MT and FT use affine, midpoint and frozen feedback with a linearized target. Under counts predictions below the observed gain by more than 10^{-8}.

Model	Method	MAE	Max. error	Under	Correct sign
Qwen3	Original	0.91392	3.06459	4/6	5/6
Qwen3	Affine	0.53728	1.93356	6/6	4/6
Qwen3	Midpoint	0.02354	0.10601	1/6	6/6
Qwen3	Frozen	2.62448	5.53123	6/6	4/6
Qwen3	AT	0.1836	0.7232	4/6	5/6
Qwen3	MT	0.49702	1.68055	0/6	6/6
Qwen3	FT	1.76900	3.89899	5/6	5/6
OLMo3	Original	1.30799	3.37169	0/6	5/6
OLMo3	Affine	0.20391	0.59816	4/6	5/6
OLMo3	Midpoint	0.01797	0.06323	3/6	6/6
OLMo3	Frozen	1.58431	5.24780	5/6	4/6
OLMo3	AT	0.4146	1.3519	5/6	5/6
OLMo3	MT	0.33551	1.10502	6/6	5/6
OLMo3	FT	1.61640	5.63615	5/6	4/6

J.2 COVERAGE AT LARGER RADII AND OBSERVED GAPS

P3 searches 30 fixed instances, consisting of ten detection heads and ten CIFAR anchors with two targets, over $q = 1, \ldots, 8$. The fixed search combines adaptive greedy proposals, local search and random proposals. It performs 22808 distinct optimization solves with no failure. Each feasible witness retains the removed and added IDs and an optimization residual correction. These witnesses are conservative float64 estimates; they are not outward-rounded lower enclosures or known neighborhood maxima.

All 80 detection upper bounds are enclosed with 128-bit Arb. The 160 historical CIFAR bounds and the directional endpoint reanalysis retain their conditional float64 status. Table 17 gives the common scale of the bounds and witnesses. At $Q = 8$, the detection witness reaches a median of only 9.59% of the upper bound, leaving a substantial unresolved gap. For CIFAR CE, the witness reaches between 95.57% and 96.05% of the directional endpoint bound across all ten anchors. The latter is a small observed gap for this target, subject to the stated numerical tier and the incomplete search.

Table 17: P3 gaps for at most Q exchanges. Every row summarizes ten anchors with $k = 512$; $Q = 4$ and 8 replace 0.78125% and 1.5625% of the subset. Medians in different columns are computed separately.

Target	Q	Upper method	Median upper	Median witness	Median gap	Witness/upper
Detection	4	Arb endpoint	0.005518	0.002333	0.003185	38.71%
Detection	8	Arb endpoint	0.029269	0.003989	0.025280	9.59%
CIFAR reg.	4	Truncated	0.000560	0.000371	0.000203	64.74%
CIFAR reg.	4	Endpoint	0.000465	0.000371	0.000105	79.29%
CIFAR reg.	8	Truncated	0.002593	0.000627	0.001987	23.73%
CIFAR reg.	8	Endpoint	0.000896	0.000627	0.000297	68.46%
CIFAR CE	4	Truncated	0.002419	0.001797	0.000622	74.32%
CIFAR CE	4	Endpoint	0.001830	0.001797	0.000033	98.22%
CIFAR CE	8	Truncated	0.007487	0.003350	0.004135	44.76%
CIFAR CE	8	Endpoint	0.003500	0.003350	0.000148	95.75%

Table 18: Every detection anchor at $Q = 8$. The report retains anchors with large gaps and those degraded by prior selection. Each upper value is the maximum over all required exchange counts, including the anchor's zero gain.

Anchor	Upper	Witness estimate	Gap	Witness/upper
degrade_16	0.837483	0.055999	0.781484	6.69%
degrade_4	0.089111	0.015336	0.073776	17.21%
hash_existing	0.032977	0.005185	0.027792	15.72%
hash_fresh	0.038946	0.006092	0.032854	15.64%
improve_1	0.025561	0.002794	0.022768	10.93%
improve_16	0.017804	0.000584	0.017220	3.28%
improve_2	0.023035	0.001901	0.021134	8.25%
improve_4	0.015033	0.001077	0.013956	7.16%
improve_8	0.012185	0.000599	0.011586	4.92%
onestep_top	0.050687	0.009592	0.041095	18.92%

What contributes to the gap. The decomposition ledger retains the linear maximum, geometric price, parameter radius, target and training curvature, exchange action, active maximization branch and optimization residual transfer for all 240 cells. Error from target sampling is zero for these fixed empirical targets. The geometric parameters are coupled. Matched ablations examine their joint effect on the final bound. On CIFAR at $q = 8$, the median matched path bounds are 0.00089654 and 0.00362072 for the regularized and CE targets, respectively. Integrated radial containment leaves both values unchanged on the fixed grid. The directional endpoint values are 0.00089622 and 0.00349996. Most improvement over the historical implementation comes from finer regional constants. The endpoint route gives a further small improvement under matched information, especially for CE. The historical and new searches use different radius grids, which prevents interpreting their total difference as a pure formula ablation.

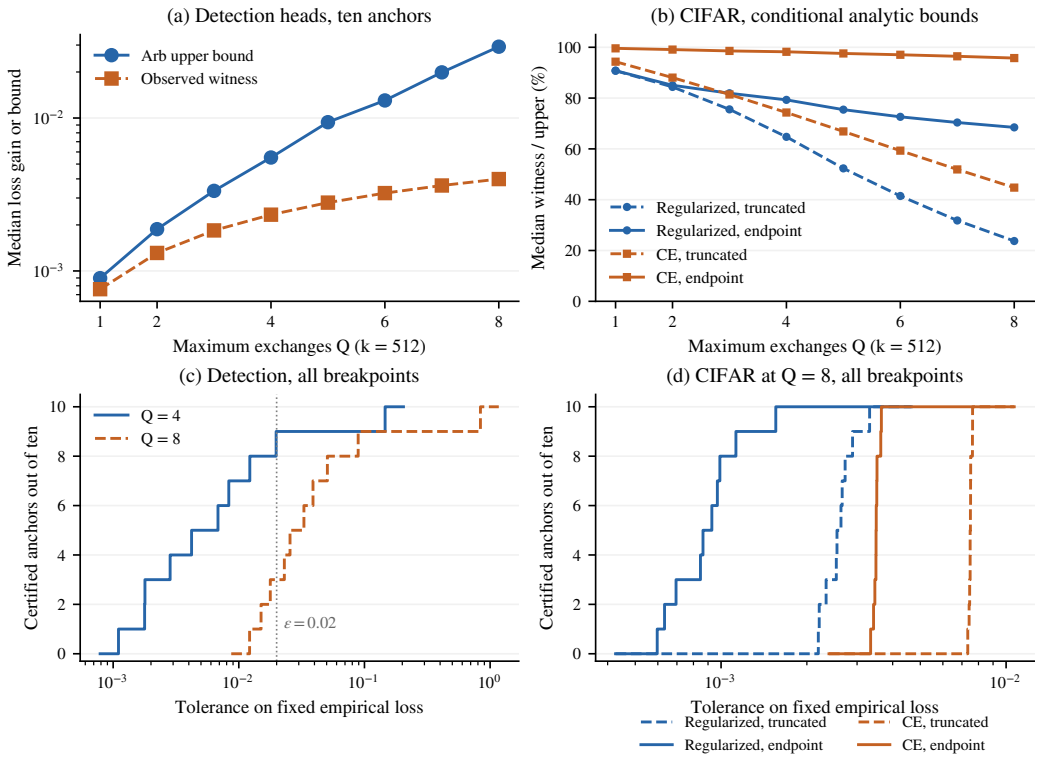

Figure 6: Complete tolerance curves and observed witnesses. **(a)** Detection upper bounds and witness estimates, with medians over all ten anchors. **(b)** CIFAR witness ratios for both targets and both historical upper methods. **(c,d)** Certification counts at every breakpoint of the upper bound for the displayed neighborhoods. The underlying ledger includes all $Q = 1, \ldots, 8$. Detection curves use Arb enclosures; CIFAR curves use conditional analytic values.

A frozen confirmation of the matrix route. A separate detection experiment froze the method and a new anchor of 512 examples selected by a fixed hash before using 145 previously unused target answers. The candidate pool and feature projection were reused. All six declared exchange counts completed Arb verification. At four exchanges the signed endpoint upper bound is 0.014218, compared with 0.041013 for the symmetric bound and 0.045850 for the simple convex target bound. Under the frozen tolerance 0.02, the three methods cover up to four, two and one exchanges, respectively. Five and six exchanges abstain for the signed endpoint method. This confirms useful certification on a fresh target and anchor within the fixed representation, with the scope of Proposition I.1. It does not certify fine-tuning of the whole model or a population target.

J.3 Direct comparison with AMIP-style and higher-order IJ predictions

AMIP studies a scalar linearization under data deletion (Broderick et al., 2020). HOIJ expands the implicit parameter solution in data weights (Giordano et al., 2019a). For the exchanges at fixed budget here, we give an explicit chord specialization and report its remainders separately from the point predictions. All 160 exact-q CIFAR witnesses from P3 are retained, with the same training objective, two targets and original data roles. The outcomes were already known when this comparison was designed, making it a retrospective method comparison.

Write $L_t = L_S + tD$, with $D = k^{-1}(\sum_{T\setminus S} \ell_i - \sum_{S\setminus T} \ell_i)$. Every L_t, $t \in [0,1]$, is globally λ-strongly convex. For $H = L_S''(\theta_0)$, $f = D'(\theta_0)$ and $E = D''(\theta_0)$, the first two implicit derivatives are

$$u = -H^{-1} f, \qquad v_2 = -H^{-1}\{L_S'''(\theta_0)[u,u] + 2Eu\}. \tag{210}$$

IJ1 evaluates the target at $\theta_0 + u$; HOIJ2 evaluates it at $\theta_0 + u + v_2/2$. The scalar second-order gain prediction is $-r_0^\top u - (u^\top R''(\theta_0)u + r_0^\top v_2)/2$. It is a different approximation from HOIJ2. The AMIP-style exchange prediction is $-r_0^\top u$, equal to the local score. The original guarantee for deletion weights is not directly applied to additions.

For completeness, let B, C_3, C_4, D_3 bound $\|D''\|$, $\|L_t'''\|$, $\|L_t''''\|$, $\|D'''\|$ along the chord region. Strong convexity contains the shifted solutions within $\rho = \|f\|/\lambda$ of the anchor. Differentiating the stationary equation gives

$$\begin{aligned} a_1 &= (\|f\| + B\rho)/\lambda, \\ a_2 &= (C_3 a_1^2 + 2Ba_1)/\lambda, \\ a_3 &= (C_4 a_1^3 + 3C_3 a_1 a_2 + 3D_3 a_1^2 + 3Ba_2)/\lambda. \end{aligned}$$

These bound the first three solution derivatives. Taylor integration therefore bounds the IJ1 and HOIJ2 parameter errors by $a_2/2$ and $a_3/6$, respectively. For softmax examples with $z_i = \|x_i\|$, valid global constants are

$$\begin{aligned} B &= \frac{\max\{\sum_{S\setminus T} z_i^2, \sum_{T\setminus S} z_i^2\}}{2k}, \\ C_3 &= \frac{\max\{\mathrm{mean}_S\, z_i^3, \mathrm{mean}_T\, z_i^3\}}{\sqrt{2}}, \qquad C_4 = \frac{7}{4}\max\{\mathrm{mean}_S\, z_i^4, \mathrm{mean}_T\, z_i^4\}, \\ D_3 &= \frac{\sum_{S\setminus T} z_i^3 + \sum_{T\setminus S} z_i^3}{k\sqrt{2}}. \end{aligned}$$

The softmax moment bounds in Appendix I justify these constants. The third solution derivative follows by differentiating stationarity three times,

$$H_t\theta''' + L_t''''[\theta', \theta', \theta'] + 3L_t'''[\theta', \theta''] + 3D'''[\theta', \theta'] + 3D''\theta'' = 0.$$

Applying the bounds on the inverse Hessian and multilinear operators gives a_3. The first two differentiations give a_1 and a_2.

For a computed anchor, shift all training objectives by $-g_0^\top\theta$, where $g_0 = \nabla L_S(\theta_0)$. Put $e_0 = \|g_0\|/\lambda$ and $A_0 = \|r_0\|e_0 + L_R e_0^2/2$. For an IJ parameter approximation z, a symmetric bound on loss error is

$$A_0 + \|\nabla R(z)\|e + L_R e^2/2, \qquad e = \begin{cases} a_2/2 + e_0, & \text{IJ1}, \\ a_3/6 + e_0, & \text{HOIJ2}. \end{cases}$$

Here $L_R = \frac{1}{2}\lambda_{\max}(\mathrm{mean}_i\, x_i x_i^\top) + \lambda_{\text{target}}$ is a global target smoothness constant. Adding $L_R\|u\|^2/2$ to the IJ1 loss error bounds the scalar linear prediction. For scalar Taylor2, a valid remainder is

$$\frac{M_{R,3}a_1^3 + 3L_R a_1 a_2 + (\|r_0\| + L_R\rho)a_3}{6} + A_0 + (\|r_0\| + L_R\rho)e_0 + \frac{L_R e_0^2}{2},$$

where $M_{R,3} = \mathrm{mean}_i\, \|x_i\|^3/\sqrt{2}$ over target examples. The first term bounds the third derivative of $R(\theta(t))$ and its Taylor integral; the remaining terms transfer the two shifted endpoints.

A separate residual certificate uses $e = \|\nabla L_T(z)\|/\lambda$ for a predicted parameter z. If R has gradient Lipschitz constant L_R, then

$$|R(z) - R(\theta_T)| \le \|\nabla R(z)\|e + L_R e^2/2.$$

This follows from strong convexity followed by a target Taylor bound. The same argument accounts for the anchor's optimization residual. This added certificate is our explicit wrapper, with its gradient cost charged to the corresponding query.

HOIJ2 gives a mean absolute prediction error of 2.95×10^{-8}. Its residual certificate has mean remainder 7.94×10^{-8} and a median query time of 7.57 ms, compared with 11.48 ms for warm-start retraining. Each method and retraining control is measured three times per case. Shared preparation has median cost 1.56 seconds per instance, including data preparation, anchor fitting, factorization and target constants. The suite peaks at 426 MiB of process RSS; this is a measurement for the entire suite. The host also runs the existing experiments.

Table 19: Direct comparison on the same 160 declared witnesses. Underpredictions refer to the point prediction at tolerance 10^{-12}; every version with a remainder has zero checked upper violations. Values are conditional float64 evaluations. Query time includes the displayed remainder or residual certificate.

Method	Point MAE	Underpredictions	Mean remainder	Query ms	Retrain/query
AMIP-style exchange	5.79×10^{-5}	0	2.70×10^{-4}	7.26	1.58
IJ1 + chord remainder	3.25×10^{-6}	39	1.76×10^{-4}	7.25	1.58
HOIJ2 + chord remainder	2.95×10^{-8}	0	2.68×10^{-5}	7.49	1.53
Scalar Taylor2 + remainder	7.61×10^{-8}	43	4.63×10^{-5}	4.38	2.63
IJ1 + residual	3.25×10^{-6}	39	9.49×10^{-6}	7.38	1.56
HOIJ2 + residual	2.95×10^{-8}	0	7.94×10^{-8}	7.57	1.52

The reported certificate curves maximize bounds over the same fixed witness family, with at most eight alternatives per anchor and target. The shared endpoint and truncated certificates cover every exchange in their declared neighborhoods without enumerating those alternatives. An IJ guarantee over the full neighborhood would additionally require valid maximization of its predictions and remainders over that family. The comparison establishes point accuracy and the cost of an explicit certificate on declared alternatives; it does not establish equal cost for these different coverage contracts.

J.4 Post hoc certification across returned detection subsets

Frozen objects and role of the experiment. E4 tests whether the same certificate gives different decisions across returned subsets under one training rule, budget and target. The protocol was frozen before the ten subsets or their bounds were computed. Qwen2.5-7B-Instruct layer-26 features and a frozen supervised projection define a ridge-logistic head with 16 coordinates and an intercept, ridge 0.001, and subset size 512 from $1{,}024$ candidate answers.

The target is the fixed empirical grouped logistic loss on 512 held-out answers containing $5{,}007$ entities. Its values also guide the construction of some returned subsets, so the claim concerns these stored target rows. Population generalization would require additional control of this target reuse. The frozen features and projection are shared across all ten subsets.

All ten returned subsets. Two anchors use stored hash orderings. Starting from the existing anchor, five subsets take $1, 2, 4, 8, 16$ rounds of eight score guided exchanges, refitting and recomputing scores at each round; two reverse this rule for four or sixteen rounds. The final subset takes the highest scored 512 candidates under the initial anchor's scores. These rules create varied inputs to the post hoc audit; the certificate accepts each returned subset as fixed.

Exhaustive check and numerical scope. For every subset, E4 retrains all $512 \times 512 = 262{,}144$ single exchanges with a Newton gradient threshold of 10^{-11} and records zero failed solves. Table 20 reports their largest gains, computed in float64, and the subsequent P3 Arb upper enclosures for the same ten anchors. The 80 upper enclosures cover $Q = 1, \ldots, 8$; the exhaustive gains retain their original numerical scope. At one exchange, the upper/exhaustive ratios range from 1.130 to 1.352.

Radius and tolerance. The original protocol reports all 61 logarithmically spaced tolerances from 10^{-4} to 10^{-1} and radii through six; P3 extends the same anchors through eight. Figure 7 evaluates the stored Arb bounds on this tolerance grid and retains every subset and radius. At the illustrative readout $Q = 4$, $\epsilon = 0.005$, all five iterative subsets are certified and the other five remain uncertified. The adjacent frozen grid point $0.005011872\ldots$ gives the same decisions. The displayed 0.005 readout is descriptive and introduces no new confirmatory threshold.

At $Q = 4$, the smallest symmetric and convex tangent bounds are 0.02171 and 0.00908, respectively, so neither baseline certifies this tolerance. An upper bound above tolerance leaves optimality unresolved; it does not establish poor subset quality or imply the full decision rule's ABSTAIN outcome when improvement evidence is also available. E4's searches at larger radii supply witnesses, with the extended P3 searches reported in Appendix J.2; exhaustive regret is available only at one exchange.

Table 20: All ten E4 subsets in protocol order. r_1 is the exhaustive float64 one-swap gain. U_Q is the P3 Arb regret upper bound over all exchanges through Q; the symmetric and convex tangent columns retain the original E4 float64 evaluations. C records only whether $U_4 \leq 0.005$; "no" means certification is withheld. All entries concern the same fixed empirical target.

Subset	Loss	Exhaustive r_1	U_1	U_1/r_1	U_4	Sym. U_4	Tangent U_4	C
hash_existing	0.5796	1.09×10^{-3}	1.25×10^{-3}	1.155	6.83×10^{-3}	3.47×10^{-2}	2.85×10^{-2}	no
hash_fresh	0.5828	1.23×10^{-3}	1.45×10^{-3}	1.175	8.32×10^{-3}	3.67×10^{-2}	3.19×10^{-2}	no
improve_1	0.5745	4.34×10^{-4}	5.41×10^{-4}	1.244	4.21×10^{-3}	3.18×10^{-2}	2.21×10^{-2}	yes
improve_2	0.5719	3.68×10^{-4}	4.44×10^{-4}	1.206	2.84×10^{-3}	2.25×10^{-2}	1.62×10^{-2}	yes
improve_4	0.5689	1.58×10^{-4}	1.79×10^{-4}	1.130	1.79×10^{-3}	2.20×10^{-2}	1.35×10^{-2}	yes
improve_8	0.5657	8.58×10^{-5}	1.04×10^{-4}	1.213	1.10×10^{-3}	2.17×10^{-2}	1.04×10^{-2}	yes
improve_16	0.5642	1.76×10^{-4}	2.38×10^{-4}	1.352	1.78×10^{-3}	2.29×10^{-2}	9.08×10^{-3}	yes
degrade_4	0.6106	2.64×10^{-3}	3.07×10^{-3}	1.162	1.98×10^{-2}	5.17×10^{-2}	5.99×10^{-2}	no
degrade_16	0.7588	9.35×10^{-3}	1.17×10^{-2}	1.249	1.46×10^{-1}	3.00×10^{-1}	2.44×10^{-1}	no
onestep_top	0.5859	1.64×10^{-3}	1.96×10^{-3}	1.195	1.22×10^{-2}	3.43×10^{-2}	3.62×10^{-2}	no

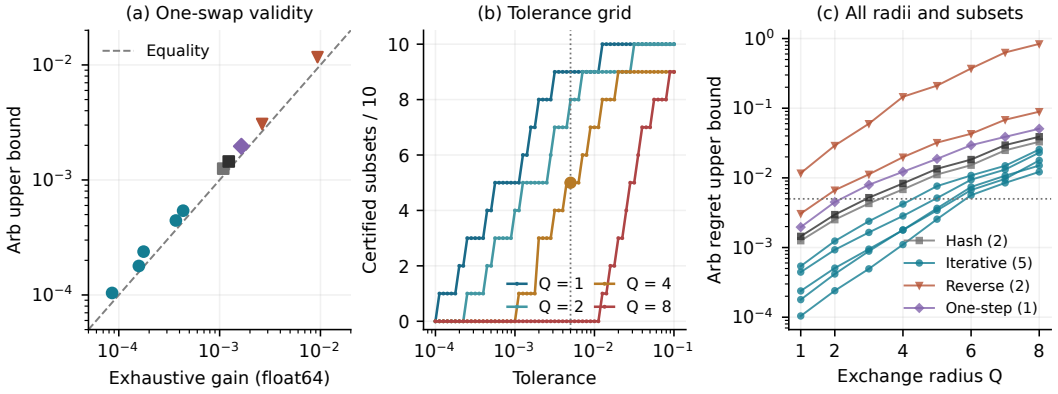

Figure 7: E4 inputs with P3 upper verification. **(a)** Every one-swap upper bound against its exhaustive float64 gain. **(b)** Counts of certified subsets over the complete frozen tolerance grid at four displayed radii. **(c)** All ten upper bound curves through $Q = 8$; group sizes are shown in parentheses. Dotted lines mark the descriptive tolerance 0.005.

No false certifications are observed for the original rule using only local scores on its frozen grid, using exhaustive gains at $Q = 1$ and witnesses at larger radii. Necessity of remainder control follows from the constructions in Appendix B.

K Tightness and construction cost over complete exchange neighborhoods

K.1 Setting and comparison

We evaluate the endpoint extension of Proposition I.5 on the detection head of Appendix J.4, using its stored candidate features, projection and empirical target.

Evaluation metrics. Let U_Q be the verified upper bound on $\mathrm{Reg}_{k,\leq Q}(S)$ and let $L_Q > 0$ be a verified lower bound from a feasible retrained exchange of radius at most Q. We report U_Q/L_Q, which bounds from above $U_Q/\mathrm{Reg}_{k,\leq Q}(S)$. Both methods use the same witnesses and lower bounds. The ratio is a conservative measure of tightness; the exact maximum gain remains unknown at $Q = 4$. Construction time covers all $q = 1, \ldots, 4$.

Datasets and models. The logistic head with ridge regularization has 16 projected coordinates and an intercept, $\lambda = 0.001$, $N = 1024$ candidate answers and $k = 512$ selected answers. The target is the fixed grouped logistic loss on the stored 512 target answers. Every selected answer may be removed and every outside answer may be added. For each of four random seeds, we retain rounds $0, 4, 16$ of the same selection rule, replacing the eight lowest scored selected answers by the eight highest

scored outside answers and refitting at each round. Seeds 2026092001 and 2026092002 supply six development anchors. The code and protocol were frozen before generating six confirmation anchors from seeds 2026092106 and 2026092107. All twelve anchors share the same data and target; confirmation varies the starting subset and its trajectory.

Baseline methods. The baseline fixes $p = H_0^{-1} r$ and uses the joint matrix difference described in Appendix I.6, with shared matrix containment and action bounds over all cross pairs. It searches a grid of 180 radii per exchange count. The extension optimizes (p, μ, ζ) at each baseline's certified radius, using the frozen SLSQP search with at most 180 iterations. Every proposed bound is checked with 128-bit Arb; a failed proposal leaves the valid baseline available. Upper verification includes containment and anchor residual transfer. Witness search is separate, and witness gains include interval evaluation and optimization residual allowances. These guarantees concern exact retraining of the head on stored features and the fixed empirical target.

K.2 Tightness across all retained anchors

Table 21 retains all twelve anchors. At $Q = 4$, the extension reduces the conservative ratio on every anchor, from baseline values between 1.504 and 2.273 to values between 1.421 and 1.564. Nine of the twelve ratios are below 1.5. The confirmation anchors attain 1.421 to 1.564, with the original data, target and search settings held fixed. The largest improvement occurs at round 16 of development seed 2026092002, where the ratio falls from 2.273 to 1.448. All 48 proposed bounds, covering four exact exchange counts at every anchor, pass interval verification. The optimizer reports convergence for 16 of these 48 proposals; validity follows from the verified inequalities for each proposal.

Table 21: Complete detection neighborhoods with $N = 1024$, $k = 512$. Ratios U_Q/L_Q are rounded upward to three decimals. Development and confirmation each contain two seeds with three retained rounds. All methods use the same strict witness lower bounds.

Seed	Round	New U_1/L_1	New U_2/L_2	Base U_4/L_4	New U_4/L_4
2026092001	0	1.081	1.145	1.599	1.480
2026092001	4	1.098	1.193	1.611	1.563
2026092001	16	1.056	1.138	2.227	1.467
2026092002	0	1.091	1.172	1.583	1.430
2026092002	4	1.078	1.173	1.578	1.554
2026092002	16	1.052	1.145	2.273	1.448
2026092106	0	1.043	1.112	1.504	1.421
2026092106	4	1.079	1.189	1.686	1.564
2026092106	16	1.045	1.133	2.144	1.478
2026092107	0	1.064	1.146	1.558	1.440
2026092107	4	1.073	1.182	1.728	1.494
2026092107	16	1.049	1.129	2.075	1.430

K.3 Cost of constructing and querying the certificate

We repeat construction three times per anchor in 36 fresh processes on a shared Intel Xeon Platinum 8558 host. Six workers use distinct physical cores and one BLAS thread each; case order rotates across repetitions. The environment uses Python 3.12.11, NumPy 1.26.4, SciPy 1.16.0 and python-flint 0.8.0. Every process rebuilds the baseline and the extension. The extension's total time includes all baseline work. Input staging, imports, anchor fitting and witness search are excluded; data reads and output writes inside construction are included. The operating system page cache is retained.

Table 22 reports every anchor's median and range. Across anchor medians, baseline construction takes 114.90 seconds and the combined method takes 144.35 seconds. The median paired increase is 25.3%, with anchor medians ranging from 20.8% to 41.8%. Table 23 separates the construction stages. The extension adds median optimization and verification times of 9.72 and 19.63 seconds. Stage medians are computed separately and need not sum to the median total. All 144 repeated proposals pass interval verification.

A cached query aggregates the stored bounds through $Q = 4$ and compares them with eight tolerances. Five batches of 2000 queries per method give median latency 2.69 microseconds for both methods,

Table 22: Construction seconds for all four exchange counts together. Entries give the median [minimum, maximum] of three repetitions. Increase is the median of paired relative increases within each anchor.

Seed	Round	Baseline	Baseline + dual	Increase
2026092001	0	113.9 [113.5, 116.6]	139.8 [139.3, 142.3]	22.7%
2026092001	4	115.1 [113.7, 116.5]	163.3 [162.4, 165.0]	41.8%
2026092001	16	114.9 [114.3, 116.8]	151.1 [150.7, 153.0]	31.5%
2026092002	0	113.8 [113.6, 115.1]	141.5 [141.3, 144.1]	24.4%
2026092002	4	114.4 [113.0, 115.9]	142.9 [141.5, 144.7]	24.8%
2026092002	16	114.9 [114.8, 116.1]	142.0 [141.0, 146.2]	22.8%
2026092106	0	114.9 [114.5, 116.2]	140.7 [140.5, 142.0]	22.5%
2026092106	4	115.1 [114.1, 115.4]	157.5 [157.0, 158.7]	37.5%
2026092106	16	114.6 [114.2, 119.0]	145.9 [144.8, 150.5]	26.8%
2026092107	0	113.6 [113.4, 114.1]	137.2 [137.0, 140.0]	20.8%
2026092107	4	117.8 [115.2, 122.8]	148.0 [145.4, 154.5]	25.8%
2026092107	16	115.2 [114.1, 116.5]	151.3 [149.1, 151.6]	30.6%

Table 23: Construction stages in seconds. The median and range summarize the medians for individual anchors over all twelve anchors.

Stage	Median	Range
Shared geometry and radius grid	34.71	$[33.41, 36.06]$
Containment search	59.88	$[59.57, 60.16]$
Baseline verification	19.63	$[19.29, 20.60]$
Additional dual optimization	9.72	$[3.99, 29.37]$
Additional dual verification	19.63	$[19.10, 20.48]$
Baseline total	114.90	$[113.61, 117.78]$
Baseline + dual total	144.35	$[137.23, 163.27]$

excluding file reads and deserialization. Median peak process memory rises from 268.7 to 287.5 MiB. The measured tradeoff is a tighter verified upper bound for additional construction cost. Complete enumeration at $q = 4$ was not timed, so these measurements provide no empirical speedup factor over exhaustive retraining.