WHEN HARD-EXAMPLE SELECTION FAILS: GAUSSIAN THRESHOLDS AND EMPIRICAL DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting the highest-loss examples can fall tens of accuracy points below random sampling at small retained fractions. We explain when this happens in a Gaussian model and test its predictions on image classification. For a fixed aligned score, we identify a threshold $p^*(r)$, given by one scalar equation in the signal-to-noise ratio r, and prove that with high probability in finite samples the selected class means invert below it, so that a nearest-mean classifier converges to the sign-flipped Bayes rule. The same population threshold holds for ridge-logistic regression. For a misaligned score we derive the selected means and the exact clean risk, so the sign flip becomes a continuous degradation with a computable above-chance boundary, and closed-form extensions cover independent score noise and a uniform-first random quota. On CIFAR-10, loss-ranked selectors do not invert real class geometry, but they displace class means four to six times farther than random selection does, consistently toward the nearest competing class. Two interventions on EL2N at a 5% fraction recover between a third and two thirds of its lost accuracy: scoring after 20 probe epochs instead of 1 raises final accuracy from 14% to 32%, and, in a separate paired experiment, replacing half of the hard quota with random draws under matched initialization raises it from 15% to 45%, against 62% for random selection. Only the second intervention moves the fixed-feature margin in step with accuracy. A geometry diagnostic that explains one successful intervention need not explain another, so such diagnostics should be validated across interventions before they are used to explain selection quality.

1 INTRODUCTION

Data selection aims to preserve useful training information with fewer examples. Methods range from geometric prototype matching (Welling, 2009) to training-dynamics and gradient scores (Toneva et al., 2019; Paul et al., 2021), gradient-matching objectives (Mirzasoleiman et al., 2020; Killamsetty et al., 2021), and distribution-aware rules (Xia et al., 2023; Zheng et al., 2023). Under shared training recipes, random sampling remains competitive (Guo et al., 2022), while some loss-ranked selections fall tens of percentage points below it at low fractions. We focus on this latter failure mode: when can selecting difficult examples systematically distort the distribution used to fit a classifier?

Existing theory already establishes that pruning quality depends on data abundance, that score-based selection can be biased, and that imperfect surrogates can change which pruning rule works best (Sorscher et al., 2022; Ayed & Hayou, 2023; Kolossov et al., 2024). Our aim is a calculable example with an explicit failure boundary. In a balanced Gaussian mixture with signal-to-noise ratio r, top-loss selection against a fixed aligned reference inverts the selected class means below $p^*(r) = \Phi(z^*)$, where $\lambda(z^*) = r$ for the inverse Mills ratio λ. A nearest-mean classifier then approaches the sign-flipped Bayes rule. Theorem 1 controls the empirical cutoff and class counts to establish this direction change with high probability. No label noise is required.

We next vary the scoring direction. Theorem 2 gives the selected population means and the exact clean risk of the nearest-mean learner as functions of the retained fraction and alignment. This extends the aligned sign flip to continuous degradation and shows how the above-chance boundary

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

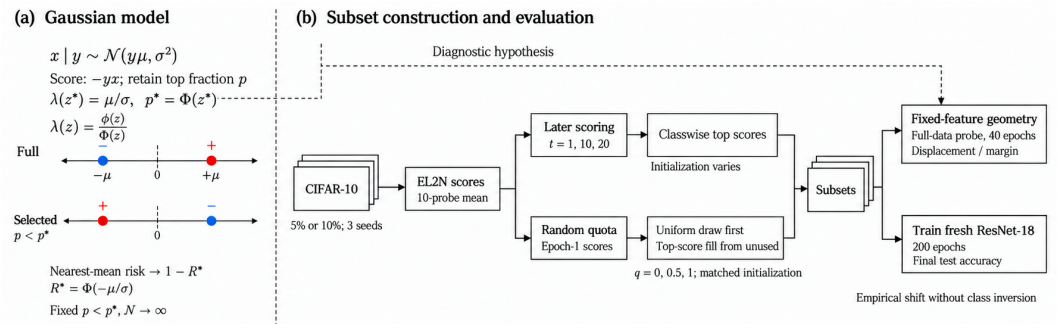

Figure 1: Overview. (a) Gaussian model of Section 4: top-loss selection inverts the selected class means below p^* (signs shown, not magnitudes). (b) CIFAR-10 experiments of Section 5: later scoring and a uniform-first random quota change the EL2N subsets, read out by fixed-feature geometry and by the final accuracy of a fresh ResNet-18. The dashed arrow marks the model prediction that the geometry readout tests.

moves. The choice of learner remains material: the aligned threshold also holds for population ridge-logistic regression, but the mismatched-direction risk formula does not generally transfer to logistic regression. These results concern a fixed score and a specified learner, not trained deep selectors in general.

The image experiments test specific implications of the model rather than assuming that the Gaussian mechanism transfers (Figure 1). The directional result predicts that, for a positively aligned score, a hard-example selector displaces each selected class mean toward the competing class by an amount that scales with an inverse Mills ratio of the retained fraction. We measure this displacement in a fixed feature space on CIFAR-10, where it separates the loss-ranked selectors from random and distribution-matching ones, and we test its predicted fraction dependence. We then compare two ways to repair low-fraction EL2N, whose benchmark protocol scores after a single probe epoch: score later in probe training, or replace part of its quota with random examples under matched downstream initialization. Both recover a substantial part of the lost accuracy. Only the second raises the measured margin along with it. A diagnostic that tracks one successful intervention can thus miss another, and this separates the usefulness of an intervention from the adequacy of a proposed geometric explanation.

Our contributions are:

- An explicit failure threshold for aligned top-loss selection with a finite-sample proof, an exact population risk curve under scoring-direction mismatch, and closed-form extensions to independent score noise and to the uniform-first random quota of our intervention.
- Fixed-feature diagnostics on CIFAR-10 showing that loss-ranked selectors displace class means boundary-ward by an amount that separates them from random and distribution-matching selection, together with a descriptive retained-fraction check of the predicted inverse-Mills scaling.
- Two interventions on EL2N, later scoring and a paired random-quota replacement, that each recover a substantial part of the lost accuracy but produce opposite margin responses, showing that a geometry diagnostic must be validated across interventions before it is used to explain selection quality.

Overlap, model-and-recipe reuse, per-class scarcity, and standard sampling guarantees are auxiliary analyses in the appendix.

2 Related Work

Data selection methods. Our experiments cover several standard families: prototype and geometry matching (Welling, 2009), forgetting and gradient statistics (Toneva et al., 2019; Paul et al., 2021), submodular and gradient-matching optimization (Mirzasoleiman et al., 2020; Killamsetty et al., 2021; Minoux, 1978), and distribution-aware scoring (Xia et al., 2023). Original CCS (Zheng

et al., 2023) is not among the evaluated implementations. DeepCore (Guo et al., 2022) standardizes implementations and already documents how thin the margins over random sampling are under a unified recipe, and Okanovic et al. (2024) show that repeatedly drawing fresh random subsets beats one-shot selectors in time-to-accuracy.

Theory of score-based pruning. Sorscher et al. (2022) analyze data pruning through statistical mechanics in a teacher-student perceptron and show that the optimal pruning policy depends on data abundance, with hard examples harmful when data is scarce. Ayed & Hayou (2023) identify the truncated distribution induced by score-based selection, prove general inconsistency and failure results, and propose random exploration with reweighting to correct the resulting bias. Kolossov et al. (2024) analyze surrogate-based selection in generalized linear models and show that surrogate accuracy need not order selection quality, with regimes where keeping hard examples is optimal and regimes where it is harmful. Distribution shift under score-based selection, high-compression failure, imperfect-proxy effects, and a distorted selected class geometry are thus established phenomena. Our contribution is not the existence of such a distortion but its explicit boundary and learner-dependent consequences in a specified model: for a fixed label-dependent high-loss rule in a Gaussian mixture, where exactly does learning turn into anti-learning, when does this happen reliably in finite samples, how does directional mismatch change the clean risk, and do different learners share the boundary? The aligned threshold with its finite-sample cutoff and class-count control, the closed-form risk under mismatch, and the noise and quota extensions are the additions.

Hard-example emphasis elsewhere. Paul et al. (2021) flag the instability of early-score pruning at low fractions. Moderate (Xia et al., 2023) and CCS (Zheng et al., 2023) already respond to this failure by avoiding the hardest examples or stratifying the selection across scores, and random exploration with reweighting (Ayed & Hayou, 2023) adds random examples to correct the bias. Theorem 1 gives the failure a precise form: below an explicit fraction, difficulty ranking against a reference predictor inverts the selected class means rather than merely emphasizing the boundary. Our unweighted random-quota intervention is a diagnostic rather than a method, and we use it to ask what adding random examples changes in the selected geometry.

3 Setup and Data

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training set with K classes and N_k examples in class k. A selection rule maps $\mathcal{D}$ to a subset $S \subseteq \mathcal{D}$ with $|S| = m = pN$, and its class-balanced variant takes $m_k = pN_k$ examples per class (rounding in implementations is addressed in Appendix A). Throughout, Random denotes the rule drawing S uniformly without replacement and Uniform its class-balanced counterpart drawing m_k examples uniformly within each class. The named selectors are Herding (Welling, 2009), Forgetting (Toneva et al., 2019), GraNd and EL2N (Paul et al., 2021), Moderate (Xia et al., 2023), a class-balanced greedy k-center baseline (kCenter), Craig (Mirzasoleiman et al., 2020), GradMatch (Killamsetty et al., 2021), and GraphCut Submodular selection (Wei et al., 2015) with lazy greedy maximization (Minoux, 1978), eleven rules with Random and Uniform. The kCenter implementation starts at the embedding nearest each class centroid and repeatedly adds the farthest point from the current selected set. Its historical artifact key is `CCS`, but it is not the importance-stratified CCS method of Zheng et al. (2023), which is not evaluated here. Craig, GradMatch, and Submodular are infeasible at ImageNet scale and excluded there. Overlap between subsets is Jaccard similarity $J(S, S') = |S \cap S'|/|S \cup S'|$.

Empirical claims are tested on controlled image classification experiments on CIFAR-10 (Krizhevsky, 2009), Tiny ImageNet (Le & Yang, 2015), ImageNet-1K (Russakovsky et al., 2015), and CUB-200 (Wah et al., 2011), trained with a ResNet-18 (He et al., 2016) under one fixed recipe per dataset with three seeds and index files to be released upon publication (Appendix E), so every geometry and overlap number can be recomputed from the artifacts. Our primary empirical evidence comes from two CIFAR-10 studies at retained fractions 0.05 and 0.1, each with three seeds and final test accuracy as the endpoint: a scoring-time sensitivity study at probe epochs 1, 10, and 20, and a paired uniform-first composition intervention. The broader multi-dataset grid supplies the geometry diagnostics and auxiliary context and reports peak test accuracy. Its epoch-1 GraNd and EL2N scoring is this benchmark's protocol choice rather than the later-scoring schedule of Paul et al. (2021).

Further Tiny ImageNet results, CUB-200, cross-seed consistency, and the model-and-recipe reuse experiments with ViT-Tiny (Dosovitskiy et al., 2021; Touvron et al., 2021) appear in the appendix.

4 HARD-EXAMPLE SCORING INVERTS CLASS GEOMETRY IN A GAUSSIAN MODEL

Random sampling comes with standard representativeness guarantees (Appendix I), but these do not explain why some selectors lose by tens of points at low fractions. This section gives the mechanism in the simplest model that exhibits it, with a critical fraction specified by one scalar equation.

Model. Labels $y \in \{\pm 1\}$ are equiprobable and $x \mid y \sim \mathcal{N}(y\mu, \sigma^2)$ with signal-to-noise ratio $r = \mu/\sigma$. The Bayes classifier is $h^*(x) = \operatorname{sign}(x)$ with risk $R^* = \Phi(-r)$. A hard-example selector scores each point by its margin loss against the reference, $s_i = -y_i x_i$, and keeps the top $m = pN$. Any strictly decreasing loss of the signed margin $y_i x_i$, including the logistic loss, induces the same top-loss ranking as this score. A plug-in learner then fits the selected set: $\hat{w} = \hat{\mu}_+ - \hat{\mu}_-$, $\hat{\theta} = (\hat{\mu}_+ + \hat{\mu}_-)/2$, predicting $\operatorname{sign}(\hat{w}(x - \hat{\theta}))$. Let $\lambda(z) = \phi(z)/\Phi(z)$ denote the inverse Mills ratio, which is strictly decreasing.

Theorem 1 (Phase transition for top-loss selection)**.** *Assume $\mu, \sigma > 0$, $0 < p < 1$, $m = pN \in \mathbb{N}$, and use the fixed classifier $h \equiv +1$ if either selected class is empty. Let z^* be the unique solution of $\lambda(z) = r$ and set $p^*(r) = \Phi(z^*)$.*

- *(a)* ***Inversion.*** *For $0 < \epsilon < 2\min(p, 1-p)$ and $0 < \delta < 1$, there is a constant C_r such that $m \geq C_r \epsilon^{-2} \log(C_r/\delta)$ guarantees, with probability at least $1 - \delta$, $\hat{w} < 0$ when $p \leq p^*(r) - \epsilon$ and $\hat{w} > 0$ when $p \geq p^*(r) + \epsilon$. For fixed $p \neq p^*(r)$ and $N \to \infty$, the risk is $1 - R^* + O_p(m^{-1})$ below the threshold and $R^* + O_p(m^{-1})$ above it.*
- *(b)* ***Random baseline.*** *For a universal constant C, if $m \geq C \max\{1, r^{-2}\} \log(8/\delta)$, training the same learner on m uniformly random examples gives $R_{\mathrm{rand}} - R^* \leq C \log(8/\delta)/m$ with probability at least $1 - \delta$ over the data and selection.*

The population threshold follows from one identity. Appendix B handles the empirical selection cutoff and class counts. Because $y_i x_i \sim \mathcal{N}(\mu, \sigma^2)$ independently of y_i, the top-loss region is $\{yx \leq a_p\}$ with $a_p = \mu + \sigma\Phi^{-1}(p)$, and the selected positive class has conditional mean $\mu - \sigma\,\lambda(\Phi^{-1}(p))$. The sign of that mean, hence of the population fitted direction, is decided by comparing $\lambda(\Phi^{-1}(p))$ with r, and monotonicity of λ makes the crossing unique. Below p^* the selected positives concentrate on the negative side of the origin in the mean, the learner recovers the geometry of the selection rather than the geometry of the classes, and its risk converges to $1 - R^*$, the mirror image of Bayes. The same population threshold also holds for ridge-regularized logistic regression with a fitted intercept (Lemma 1). Neither result establishes a corresponding claim for deep networks.

Remark 1 (Noise and dimensions)**.** *The population threshold increases strictly with the independent label-flip rate $0 \leq \eta < 1/2$ when scoring uses the observed labels ($p^*_{0.1} \simeq 0.560$ at $r = 1$). The same population threshold holds for isotropic Gaussian mixtures scored along their signal direction. The one-dimensional finite-sample theorem applies to a learner restricted to that direction, and unrestricted higher-dimensional fitting adds estimation error (Appendix B).*

Simulation matches the threshold. The simulated flip point of $\hat{w}$'s sign coincides with $p^*(r)$ at every tested signal-to-noise ratio (Table 1). All four match to within 5×10^{-4}, deep below threshold the empirical risk equals $1 - R^*$ to three decimals, and far above threshold top-loss selection is as good as random. Figure 2 shows the full error curves with the anti-learning region marked, alongside EL2N underperformance at low retained fractions on CIFAR-10 and Tiny ImageNet. The image panel does not establish class inversion.

A misaligned scoring direction turns inversion into continuous degradation. The aligned model isolates a sharp sign change. Allowing the reference predictor to rotate gives a continuum of population risks, so poor performance need not require the selected means to point exactly opposite to the original means.

Table 1: Theoretical critical fraction against simulated flip points ($N = 10^5$, 20 trials, grid resolution 0.005, flip point estimated per trial as the midpoint of the sign change and averaged).

$r = \mu/\sigma$	z^*	$p^*(r)$	simulated flip
0.5	0.5179	0.6977	0.6977 ± 0.0026
1.0	-0.3026	0.3811	0.3806 ± 0.0025
1.5	-0.9685	0.1664	0.1659 ± 0.0023
2.0	-1.5719	0.0580	0.0584 ± 0.0019

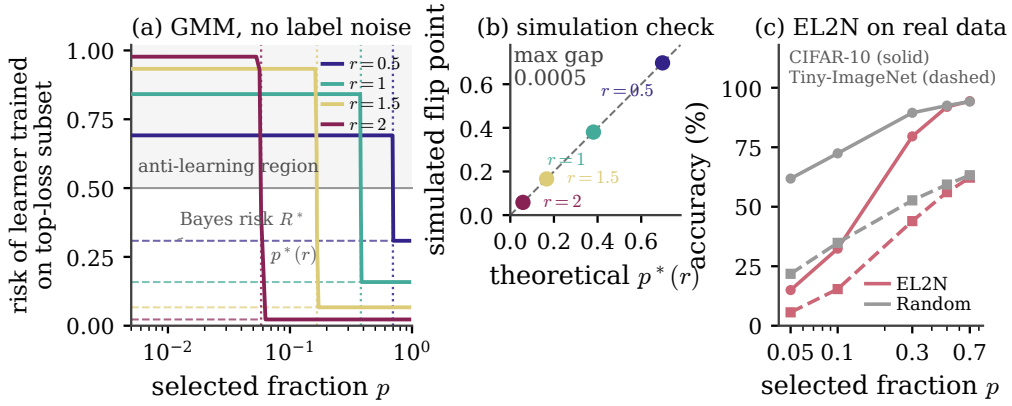

Figure 2: Left: risk of the plug-in learner on top-loss selections against fraction p for four signal-to-noise ratios, with $p^*(r)$ marked and the anti-learning region shaded. Middle: simulated flip points against the theoretical threshold. Right: EL2N against Random accuracy by fraction on CIFAR-10 and Tiny ImageNet, illustrating low-fraction underperformance rather than observed class inversion.

Theorem 2 (Population risk under proxy-direction mismatch)**.** *Let y be uniform on $\{-1, 1\}$ and $x \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, \sigma^2 I_d)$, with $\boldsymbol{\mu} \neq 0$ and $\sigma > 0$. Fix a unit scoring direction v, independent of the training sample, and retain the population fraction $p \in (0, 1)$ with largest score $-yv^\top x$. Write $r = \|\boldsymbol{\mu}\|/\sigma$, $c = v^\top \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, and $\beta_p = \lambda(\Phi^{-1}(p))$. The selected class means are $\pm a_p$, where*

$$a_p = \boldsymbol{\mu} - \sigma\beta_p v. \tag{1}$$

For $a_p \neq 0$, the nearest-mean classifier has clean test risk

$$R_p(v) = \Phi\left(-\frac{r^2 - rc\beta_p}{\sqrt{r^2 + \beta_p^2 - 2rc\beta_p}}\right). \tag{2}$$

It is worse than chance exactly when $c > 0$ and $\beta_p > r/c$. For $|c| < 1$, its risk increases strictly as p decreases, from $\Phi(-r)$ as $p \uparrow 1$ to $\Phi(rc)$ as $p \downarrow 0$.

For $c = 1$, this recovers the original threshold $p^*(r)$ and the two risks R^* and $1 - R^*$. At the unique degeneracy $c = 1$, $\beta_p = r$, both centroids are zero, and a fixed-label tie rule has risk $1/2$. For $0 < c < 1$, the above-chance boundary is $p^*(r/c)$ (Figure 3, right), smaller than the aligned boundary. Thus stronger alignment can enlarge the anti-learning region for this particular high-loss rule. This is not a claim that less accurate probes improve arbitrary selection methods. For $c \leq 0$ the risk never exceeds chance, although imperfect alignment still raises it above Bayes risk at every $p < 1$ when $|c| < 1$. This is a population nearest-mean result, and learner dependence under mismatch is substantive rather than a proof limitation. For $\sigma = 1$, $\boldsymbol{\mu} = (1, 0)$, $v = (0.8, 0.6)$, and $p = 0.3$, the nearest-mean clean risk is approximately 0.459, whereas population ridge-logistic regression with penalty 0.1 on the same selected population has risk approximately 0.592, so the two learners lie on opposite sides of chance. The logistic equivalence in Lemma 1 is thus specific to the aligned model. Appendix C gives the proof, the endpoint cases, and the numerical validation of this example.

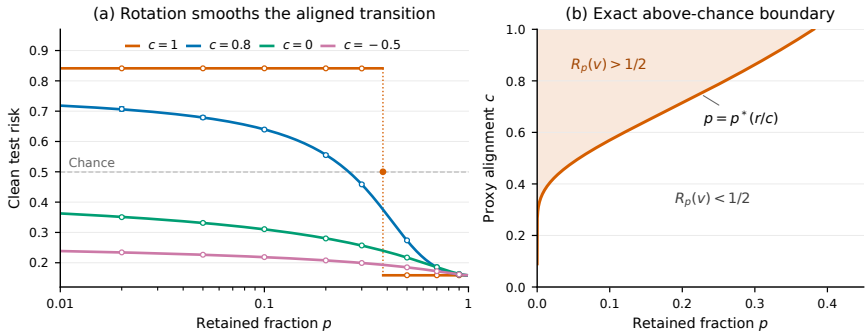

Figure 3: Proxy direction changes clean risk ($r = 1$). Left: exact population risks (lines) and empirical top-m nearest-mean fits (points, mean $\pm$ one standard deviation over 32 independent samples with $N = 40{,}000$). Clean risk is evaluated analytically. Right: the above-chance boundary $p = p^*(r/c)$ for $c > 0$. The points check the derived Gaussian risk curve.

5 Empirical diagnostics and subset interventions

5.1 Does real class geometry show the predicted distortion?

What the model predicts for real data. Equation (1) supplies a quantitative signature: at a fixed scoring direction, a hard-example selector displaces each selected class mean from its full-data mean along the score direction, by an amount proportional to $\beta_p = \lambda(\Phi^{-1}(p))$. Under directional mismatch the displacement need not be collinear with the signal, and it does not guarantee a positive projection onto it, so the absence of inversion on real data neither confirms nor refutes the model. What can be tested is whether the displacement exists, whether it points toward the competing class, whether it grows with compression at the predicted rate, and whether it moves together with accuracy when the subset is changed. We measure displacement and margin in one fixed probe feature space (Appendix D) and take these questions in turn. The class-balanced quota does not by itself prevent inversion: in the Gaussian model the score distribution is independent of the class label even though the score uses the label.

Class means shift but do not invert in the probe feature space, and the families differ in how far they move. On CIFAR-10 the sign inversion of Theorem 1 is absent at the class-mean level, and what separates the loss-ranked selectors from the rest is the size of a residual displacement. We compare the class means of every historical subset at $f{=}0.05$ with the full-data class means. Every selected class mean keeps the direction of its full-data class. Even the lowest single-class cosine over all methods and seeds, 0.987, stays close to one, and nearest-centroid identification is perfect. With a probe at full training accuracy, whose linear head classifies every training point correctly, a selected class mean is expected to stay on its own side of that head, so the absence of inversion is anticipated and the informative quantities are the size and direction of the shift. The magnitude of the shift is where the families part. EL2N, GraNd, and Forgetting move their class means roughly 4 to 6 times farther from the full-data centroids than Random and Uniform do, Herding stays near the random level, and kCenter lands between the groups (displacement norm in Table 2, cosines in Appendix D). A hard-example selector thus distorts the class geometry of real features without inverting it.

The displacement is boundary-ward, and direction alone does not separate the families. The loss-ranked selectors move their class means toward the competing class, and it is the size of that move rather than its direction that distinguishes them from Herding. A displacement magnitude does not fix its direction, so we also measure the signed margin from each selected class mean to the nearest competing full-data centroid (Appendix D, Table 2). Forgetting, EL2N, and GraNd give up 6 to 11% of the full-data margin, boundary-ward in 88 of 90 class-and-seed cases, whereas Random and Uniform keep the margin essentially unchanged. Herding moves boundary-ward in every one of its cases as well, yet gives up only a small part of the margin. The direction is shared across selectors and the magnitude is not. The cross-method comparison thus shows an association between displacement magnitude and low-fraction failure. Whether the same diagnostic tracks changes within one selector is the question of Section 5.2.

Table 2: Direction of the selected-class-mean displacement on CIFAR-10 at f=0.05, against each class's nearest competing full-data centroid, over 3 seeds and 10 classes. Margin retention below 1 and a negative along-u component both mean boundary-ward. Along-u component and displacement norm are fractions of the distance between the two full-data centroids. The last column counts the 30 class-and-seed cases whose displacement points at the competing class.

Method	Margin retention	Along-u component	Displacement norm	Boundary-ward cases
Forgetting	0.886	-0.057	0.078	30/30
EL2N	0.930	-0.035	0.078	30/30
GraNd	0.940	-0.030	0.058	28/30
Herding	0.973	-0.013	0.021	30/30
kCenter	0.986	-0.007	0.035	22/30
Moderate	0.995	-0.003	0.016	18/30
Random	0.999	-0.001	0.015	18/30
Uniform	1.002	$+0.001$	0.014	12/30

The predicted fraction scaling partly holds for early scores and fails for late ones. The directional model predicts that the boundary-ward displacement scales with β_p as the fraction changes, and the early-score curve partly follows this form while the late-score curve does not. We test it by reselecting seven class-balanced fractions from each saved scoring vector and fitting a single coefficient multiplying β_p to the mean boundary-ward displacement (Appendix F). The centered R^2 of this through-origin fit is 0.830 at scoring epoch 1 and 0.581 at epoch 10, but -0.030 at epoch 20, where the proportional form does worse than a constant. A single proportional scale thus does not carry over from early to late scores. Allowing a free intercept raises the epoch-20 mean-curve R^2 to 0.984, so the restriction that fails is proportionality through the origin, not dependence on the inverse-Mills term. These are fits to fixed-feature geometry, rather than calibrated predictions of downstream risk.

5.2 Two interventions that recover accuracy with opposite margin responses

Later scoring improves accuracy without monotone margin recovery. Scoring EL2N later in probe training recovers a substantial part of its low-fraction accuracy, and the fixed-feature margin does not recover with it. The historical benchmark scores after one probe epoch, earlier than the epoch-10 and epoch-20 settings of Paul et al. (2021), so we rerun EL2N at scoring epochs 1, 10, and 20 under the unchanged downstream recipe (Appendix G). Accuracy rises with the scoring epoch at both fractions. At $f = 0.05$, final accuracy more than doubles, from 14.12% at epoch 1 to 32.13% at epoch 20 (Table 3). Even at epoch 20, EL2N remains 29.82 and 26.32 points below Random at the two fractions, and every epoch-20 run stays below its seed-matched Random and Uniform controls, so the unusually early benchmark scoring adds to the severity of the failure but does not create it. The margin measured on the same subsets in the same fixed feature space does not follow. It dips at epoch 10 and ends at epoch 20 below its epoch-1 value at both fractions, with retention 0.927, 0.886, and 0.900 at $f = 0.05$ for the rerun subsets, against 0.930 for the historical subsets in Table 2 (Figure 4, Table 4). Because scoring consumes different random streams before downstream initialization, this is a protocol sensitivity study rather than a paired intervention. Even so, the margin diagnostic misses an improvement that accuracy registers clearly, so it is not a sufficient explanation of recovery under changed scoring time.

Table 3: CIFAR-10 scoring-epoch sensitivity (Appendix G): mean $\pm$ sample standard deviation over three seeds, accuracy in percent. Final accuracy is the primary endpoint and peak test accuracy a secondary, optimistically selected one. Controls are rerun in the same batch.

	$f = 0.05$		$f = 0.1$	
Selection	Final	Peak	Final	Peak
EL2N, $t = 1$	14.12 ± 1.78	14.41 ± 1.55	28.80 ± 7.29	29.05 ± 7.38
EL2N, $t = 10$	25.07 ± 0.16	25.72 ± 0.46	38.23 ± 2.11	38.66 ± 1.83
EL2N, $t = 20$	32.13 ± 1.66	32.57 ± 1.75	46.44 ± 1.90	46.88 ± 2.05
Random	61.96 ± 1.10	62.11 ± 1.03	72.76 ± 0.94	72.82 ± 0.95
Uniform	60.72 ± 3.31	60.92 ± 3.33	72.66 ± 1.84	72.81 ± 1.85

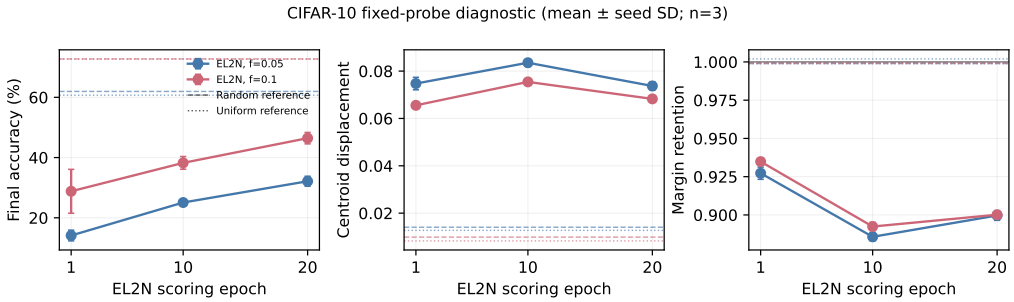

Figure 4: Changing EL2N scoring time: final accuracy improves, whereas displacement and margin do not improve monotonically. Lines show three-seed means, error bars one sample standard deviation. Dashed and dotted horizontal lines are fresh Random and Uniform means. Geometry is measured in one fixed 40-epoch ResNet-18 feature space and is not a downstream margin.

Later scoring is not a change in score noise in the model. Lemma 2 constrains how the directional model can account for this improvement. At a fixed direction, independent Gaussian score noise divides the mean shift by $\sqrt{1+\tau^2}$, and the nearest-mean risk increases strictly with the shift for $|c| < 1$, whatever the value of τ. Later scoring, however, increases the boundary-ward shift, as the margin retentions above show, while accuracy rises. No change in score-noise variance, in either direction, reproduces this pattern within the model, so the improvement must involve a change in the score direction or structure, which the present experiments do not identify.

A half-random quota closes about 65% of the accuracy gap. Diluting the epoch-1 EL2N subset with random examples, at a fixed subset size and under identical downstream initialization, closes about 65% of the gap to random selection at both fractions and improves on pure EL2N in every seed. The intervention allocates a share $q \in \{0, 0.5, 1\}$ of each class quota to uniform random draws and fills the remainder with the highest-scoring unused examples, holding the fraction, class quotas, training recipe, and initial weights fixed (Appendix H). At $f = 0.05$, mean final accuracy rises from 15.02% for pure EL2N to 45.23% at the half share, against 61.67% for class-balanced random selection. The ordering is the same at $f = 0.1$ (Table 4). Half of the subset drawn at random thus removes about two thirds of the damage, and the remaining score-selected half still costs 14 to 16 points relative to the random endpoint.

Table 4: The two interventions side by side: three-seed mean final accuracy and fixed-feature margin retention on CIFAR-10. The scoring-time study and the paired mixing study are separate batches with their own EL2N baselines, so their starting points are not one paired experiment. Standard deviations are in Table 3 and Appendix H.

Intervention	f	Final accuracy (%)	Margin retention
Scoring epoch $1 \to 10 \to 20$	0.05	$14.12 \to 25.07 \to 32.13$	$0.927 \to 0.886 \to 0.900$
Scoring epoch $1 \to 10 \to 20$	0.1	$28.80 \to 38.23 \to 46.44$	$0.935 \to 0.892 \to 0.900$
Random share $q = 0 \to 0.5 \to 1$	0.05	$15.02 \to 45.23 \to 61.67$	$0.927 \to 0.962 \to 0.997$
Random share $q = 0 \to 0.5 \to 1$	0.1	$32.36 \to 58.45 \to 72.68$	$0.935 \to 0.963 \to 0.996$

Random mixing raises the margin in step with accuracy, unlike later scoring. The composition intervention moves accuracy and the fixed-feature margin in the same direction, which the scoring-epoch change did not. Margin retention rises with the random share at both fractions, from 0.927 for pure EL2N to 0.997 for the random endpoint at $f = 0.05$, with the half share at 0.962 (Table 4 and Figure 5). The nearly linear mean geometry is descriptive. Lemma 2 predicts a different curve, because drawing random points first and excluding duplicates changes the hard component's effective tail, whereas intermediate shares $q = 0.25, 0.75$ measured in the same feature space fall on the line between the endpoints (Appendix F). The two interventions together separate an effective subset change from an adequate explanation of it. Mixing changes many properties correlated with the score at once, so the matching trends here do not by themselves identify a causal effect of the measured margin, and the scoring-epoch experiment shows that the margin can stay low while accuracy recovers.

Auxiliary analyses in Appendices I and J complement these CIFAR-10 results on the other datasets: at low fractions, methods that match Random in accuracy on ImageNet and Tiny ImageNet select subsets that overlap with one another about as much as independent draws under the same class quotas would, Herding's low-fraction edge often survives reuse of its subsets under a different model and recipe whereas Forgetting's edge at $f = 0.3$ inverts, shrinking the per-class count on Tiny ImageNet reproduces the CUB-200 method ordering, and the selectors that lose most to Random are among the most self-consistent across seeds, so repeatability is not evidence of subset quality (Appendix I.2).

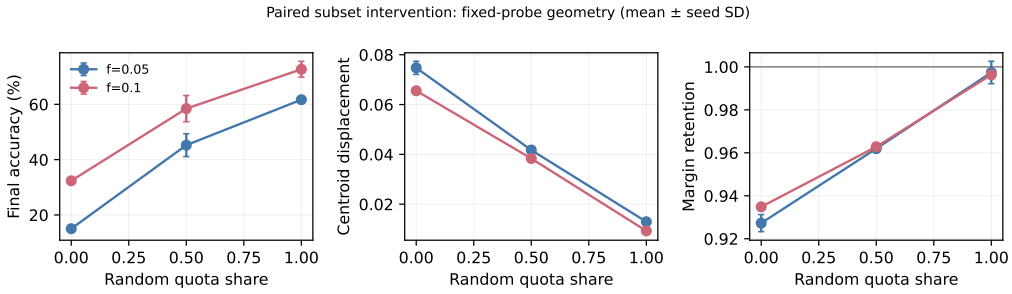

Figure 5: Paired random-mixing intervention with fixed subset size and class quotas. Increasing the random share improves final accuracy and moves the fixed-probe geometry toward the random level. Points are three-seed means with one sample standard deviation. Initial weights and seeded random streams are matched across shares within each fraction and seed, with nondeterministic GPU kernels retained. Geometry is a readout of the intervention, not its independently randomized component.

6 Discussion and Conclusion

The Gaussian analysis and the image experiments answer different questions, and the conclusions depend on keeping them apart. The theory says exactly when a fixed high-loss rule with a specified learner inverts or degrades the class geometry it selects. The experiments ask whether the signature this predicts for a positively aligned score, a boundary-ward shift of selected class means, appears on real features, and whether it moves together with accuracy when the subset is changed. It appears, it separates loss-ranked selectors from random and distribution-matching ones, and its fraction dependence partly follows the predicted proportional form for early scores but not for late ones. It does not move together with accuracy under every intervention: later scoring improves accuracy while leaving the margin lower, whereas random mixing improves both.

Two further points follow. The mixing result locates the damage: giving half of each class quota to uniform draws recovers about 65% of the gap to random selection at both fractions, and the score-selected half that remains still costs 14 to 16 points (Table 4), so the loss is carried by the hard examples themselves rather than by the reduced subset size. For analysts, a fixed-feature geometry diagnostic that tracks one successful intervention can miss another, so agreement on a single intervention is not evidence that the diagnostic captures the mechanism.

Limitations. The theory covers Gaussian mixtures with a fixed score and nearest-mean or population logistic learners, and it makes no claim about deep networks. The interventions use EL2N on CIFAR-10 at two fractions, the scoring-time study is a protocol sensitivity study rather than a paired one, and the geometry is read in one fixed probe feature space rather than in the downstream network's own, so no mediation by the margin is established.

Hard-example selection can distort the class geometry a learner sees, and the Gaussian model makes the distortion explicit: an aligned score inverts the selected means below a threshold given by one scalar equation, a misaligned score degrades risk continuously with a computable above-chance boundary, and score noise and a uniform-first random quota change the selected means in closed form. On CIFAR-10, two interventions on EL2N recover between a third and two thirds of its lost accuracy. A geometry diagnostic should be validated across distinct interventions before it is used to explain selection quality.

Reproducibility statement

Full proofs are in Appendices A, B, and C, and Appendix E specifies the training recipes and probe configurations. Code, selected-index files, and per-run result tables will be released upon publication: the simulation code for Theorem 1 with fixed seeds, the numerical validation of Theorem 2, the index files and result tables behind the geometry, overlap, reuse, and intervention analyses, and the scripts that recompute every figure from them.

AI use statement

In this work, we used generative AI tools to implement standard methods, to assist with data analysis, numerical checks and figure drawing, and to draft and edit parts of the paper. We have reviewed all AI-assisted work: the proofs were checked by the authors and selected identities verified numerically, the generated code was tested, and all drafted and edited text was read and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Fadhel Ayed and Soufiane Hayou. Data pruning and neural scaling laws: Fundamental limitations of score-based algorithms. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2302.06960.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. DeepCore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications*, volume 13426 of *Lecture Notes in Computer Science*, pp. 181–195, 2022. doi: 10.1007/978-3-031-12423-5_14. URL https://link.springer.com/chapter/10.1007/978-3-031-12423-5_14.

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. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Germain Kolossov, Andrea Montanari, and Pulkit Tandon. Towards a statistical theory of data selection under weak supervision. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.14563.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. *CS231N Course Report, Stanford University*, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, pp. 234–243. Springer, 1978. doi: 10.1007/BFb0006528. URL https://link.springer.com/chapter/10.1007/BFb0006528.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Patrik Okanovic, Roger Waleffe, Vasilis Mageirakos, Konstantinos E. Nikolakakis, Amin Karbasi, Dionysis Kalogerias, Nezihe Merve Gürel, and Theodoros Rekatsinas. Repeated random sampling for minimizing the time-to-accuracy of learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.18424.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/ac56f8fe9eea3e4a365f29f0f1957c55-Abstract.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019. URL https://www.cambridge.org/core/books/understanding-machine-learning/076FB5C9121E120E7B206E1774DC114D.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://epubs.siam.org/doi/10.1137/1116025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1954–1963, 2015. URL https://proceedings.mlr.press/v37/wei15.html.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 1121–1128, 2009. doi: 10.1145/1553374.1553517. URL https://dl.acm.org/doi/10.1145/1553374.1553517.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7D5EECbOaf9.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QwKvL6wC8Yi.

A PROOF OF THEOREM 3

For a hypothesis class $\mathcal{H}$ and loss ℓ bounded in $[0,1]$, write $\hat{L}_S(h) = \frac{1}{m}\sum_{i\in S}\ell(h(x_i), y_i)$. Call S ϵ-representative for $\mathcal{H}$ if $\sup_{h\in\mathcal{H}}|\hat{L}_S(h) - \hat{L}_{\mathcal{D}}(h)| \le \epsilon$, and let $\mathcal{F}(\epsilon, \mathcal{H})$ be the family of ϵ-representative size-m subsets. The theorem below is standard background for the empirical comparisons in Appendix I, using two-sided uniform convergence for sampling without replacement and hypergeometric concentration for overlap baselines.

Theorem 3 (Uniform-sampling reference)**.** *Fix $\mathcal{D}$ and $0 < \epsilon, \delta < 1$. Let $d \ge 1$ bound the VC-subgraph dimension of the induced loss class $\{(x,y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ (VC dimension for binary losses), and set $M(d, \epsilon, \delta) = C\epsilon^{-2}[d\log(2d/\epsilon) + \log(2/\delta)]$ for a universal C.*

(i) ***Uniform sampling and counting.*** *If $m \ge M(d, \epsilon, \delta)$, a uniform size-$m$ subset is ϵ-representative with probability at least $1-\delta$, hence $|\mathcal{F}| \ge (1-\delta)\binom{N}{m}$. For proportional class-balanced sampling, assume integer quotas $m_k = pN_k$. If every $m_k \ge M(d, \epsilon, \delta/K)$, the corresponding bound is $|\mathcal{F} \cap \Omega_{\mathrm{bal}}| \ge (1-\delta)\prod_k \binom{N_k}{m_k}$, where Ω_{bal} is the quota-constrained sample space. Independent draws in either space have $\mathbb{E}J \to p/(2-p)$ as $N \to \infty$ at fixed $p \in (0,1)$.*

(ii) ***Simultaneous guarantee.*** *For a finite menu $\mathcal{H}_1, \ldots, \mathcal{H}_J$, use the largest induced loss-class dimension and replace δ by δ/J in the corresponding condition in (i). The same draw is then representative for all J classes with probability at least $1-\delta$.*

(iii) ***Per-class approximation.*** *Let $P_{\mathcal{D},k}$ be the empirical distribution of class k in $\mathcal{D}$, with support size s_k and $N_k > 1$. For a uniform size-m_k class subset, $\mathbb{E}\,\mathrm{TV}(\hat{P}_k, P_{\mathcal{D},k}) \le \frac{1}{2}\sqrt{\frac{(s_k-1)(N_k-m_k)}{m_k(N_k-1)}}$. If all class examples are distinct, the exact value for every subset is $\mathrm{TV}(\hat{P}_k, P_{\mathcal{D},k}) = 1 - m_k/N_k$, so at a fixed fraction this statistic cannot distinguish selection rules or per-class sample counts.*

Uniform convergence for a fixed finite population. The randomness here is only the draw of a subset of the fixed dataset $\mathcal{D}$. The target is its empirical loss, not an unknown population loss. Write $\mathcal{G} = \{\ell \circ h : h \in \mathcal{H}\}$ and $Z_S = \sup_{g\in\mathcal{G}}|m^{-1}\sum_{i\in S} g(x_i, y_i) - P_{\mathcal{D}}g|$. Hoeffding's convex comparison for sampling without replacement applies to finite vectors of loss values. In particular, the exponential of the supremum of the absolute centered coordinates is convex, so

$$\mathbb{E}e^{tZ_S} \le \mathbb{E}e^{tZ_{\mathrm{iid}}}, \qquad t > 0,$$

where the right side uses m independent draws from $P_{\mathcal{D}}$. For completeness, couple the draws as follows: take m iid indices $I_1, \ldots, I_m$ and uniformly augment their distinct indices to an m-element set S. This S is uniform by permutation symmetry. Conditional on S, each of its elements has expected multiplicity one among the iid draws, so $\mathbb{E}[\sum_j v(I_j) \mid S] = \sum_{i\in S} v(i)$ for any fixed vectors $v(i)$. Jensen's inequality gives the displayed comparison. One can first use a finite set of loss vectors and pass to the supremum. On the fixed finite dataset its closure lies in the compact cube $[0,1]^N$. For iid draws the standard bounded VC-subgraph empirical-process bound gives $\mathbb{E}Z_{\mathrm{iid}} \le C_0\sqrt{d/m}$, and bounded differences gives $\mathbb{E}e^{t(Z_{\mathrm{iid}} - \mathbb{E}Z_{\mathrm{iid}})} \le e^{t^2/(8m)}$. Hence Chernoff's inequality yields

$$\Pr\Big(Z_S > C_0\sqrt{d/m} + \sqrt{\log(1/\delta)/(2m)}\Big) \le \delta.$$

The stated M is a sufficient, deliberately loose threshold. These are standard uniform-convergence facts, not a new capacity bound (Vapnik & Chervonenkis, 1971; Shalev-Shwartz & Ben-David, 2014; Hoeffding, 1963). If the sufficient threshold exceeds the dataset size, it gives no nontrivial guarantee for a proper subset. Taking the whole dataset is always exact.

Counting and stratification. Every ordinary size-m subset has mass $\binom{N}{m}^{-1}$, so the high-probability statement immediately gives the count in (i). For the class-balanced case draw each S_k independently and uniformly from class k and use confidence δ/K in the preceding bound. With probability at least $1-\delta$, every class-conditional loss deviation is at most ϵ. Proportional quotas

give $m_k/m = N_k/N$, and hence

$$\sup_h |\hat{L}_S(h) - \hat{L}_{\mathcal{D}}(h)| \le \sum_k \frac{N_k}{N} \sup_h |\hat{L}_{S_k}(h) - \hat{L}_{\mathcal{D}_k}(h)| \le \epsilon.$$

The uniform measure is now on Ω_{bal}, whose size is $\prod_k \binom{N_k}{m_k}$, not on all size-m subsets. For non-proportional or rounded quotas the same proof adds the deterministic term $\sum_k |m_k/m - N_k/N|$ to the deviation bound. The exact proportional theorem should not silently be applied after rounding.

Overlap is a sampling reference, not an algorithm prediction. For two independent ordinary draws, $I = |S \cap S'| \sim \text{Hypergeom}(N, m, m)$, with $\mathbb{E}I = m^2/N$ and $\text{Var}(I) \le m$. For proportional stratified draws, $I = \sum_k I_k$ is a sum of independent per-class hypergeometrics with $\mathbb{E}I = pm$ and again $\text{Var}(I) \le m$. Thus $I/m \to p$ in probability as $m \to \infty$, and $J = (I/m)/(2 - I/m) \to p/(2-p)$. Since $0 \le J \le 1$, the expectations converge as well. The finite-sample expectation is not in general the ratio of the expectations. None of this assumes that learned selectors draw independently or uniformly from $\mathcal{F}$.

Simultaneous approximation. Apply (i) to each induced loss class at confidence δ/J and take a union bound. In the stratified case this means per-class confidence $\delta/(JK)$. This proves (ii). It gives neither an accuracy comparison between trained architectures nor a guarantee for nonuniform selection rules.

Finite-support approximation. For class k, write $q_j = P_{\mathcal{D},k}(j)$ on its s_k support points. Sampling without replacement gives

$$\text{Var}(\hat{q}_j) = \frac{q_j(1-q_j)}{m_k} \frac{N_k - m_k}{N_k - 1}.$$

Consequently, Jensen's and Cauchy–Schwarz inequalities give

$$\begin{aligned}\mathbb{E}\,\text{TV}(\hat{P}_k, P_{\mathcal{D},k}) &\le \frac{1}{2} \sum_j \sqrt{\text{Var}(\hat{q}_j)} \\ &\le \frac{1}{2} \sqrt{\frac{N_k - m_k}{m_k(N_k - 1)}} \sqrt{s_k \left(1 - \sum_j q_j^2\right)} \le \frac{1}{2} \sqrt{\frac{(s_k - 1)(N_k - m_k)}{m_k(N_k - 1)}}.\end{aligned}$$

For $N_k = m_k = 1$ the deviation is zero. This is a comparison with the full empirical class distribution. Comparison with an unknown population would additionally require controlling how that full dataset approximates the population. In particular, a CDF bound for a one-dimensional marginal is not a total-variation bound for a continuous distribution. For distinct class examples, direct summation gives

$$\text{TV}(\hat{P}_k, P_{\mathcal{D},k}) = \tfrac{1}{2}\left[m_k(1/m_k - 1/N_k) + (N_k - m_k)/N_k\right] = 1 - m_k/N_k.$$

The bound then reduces to $\frac{1}{2}\sqrt{(1-p)/p}$ for $m_k = pN_k$. Neither quantity changes with N_k at fixed p, and the exact TV is identical for every selection rule. Thus (iii) cannot explain the fixed-fraction scarcity experiment.

B Proof of Theorem 1

Write $A_i = y_i x_i$. The A_i are independent $\mathcal{N}(\mu, \sigma^2)$ variables independent of the labels. Top-loss selection retains the m smallest A_i. Let $T = A_{(m)}$, $z_q = \Phi^{-1}(q)$, and $b(t) = \mathbb{E}[A \mid A < t] = \mu - \sigma\lambda((t-\mu)/\sigma)$.

Truncated samples and the order-statistic boundary. Conditional on $T = t$, the $m - 1$ observations strictly below t, in a uniformly random ordering, are independent draws from $A \mid A < t$. The remaining selected observation equals t. All selected labels remain independent uniform signs, independent of these values. This standard order-statistic factorization follows directly by factoring

the joint density below and above t. It avoids treating all m selected observations as independent truncated draws.

For $Z \sim \mathcal{N}(0,1)$ conditioned on $Z < z$, the log moment-generating function has second derivative $\mathrm{Var}(Z \mid Z < z - u) \le 1$ at tilt u, since

$$\mathrm{Var}(Z \mid Z < v) = 1 - \lambda(v)\{v + \lambda(v)\} \le 1.$$

Here $v + \lambda(v) > 0$ because the conditional mean $-\lambda(v)$ lies strictly below v. Thus the centered truncated variable is 1-sub-Gaussian, and each centered truncated A is σ-sub-Gaussian, uniformly over the cutoff.

Cutoff and class-count events. Put $q_\pm = p \pm \epsilon/2$ and $a_q = \mu + \sigma z_q$. The Dvoretzky–Kiefer–Wolfowitz inequality and $F_N(T) = m/N = p$ give

$$\Pr\{T \notin [a_{q_-}, a_{q_+}]\} \le 2e^{-N\epsilon^2/2}.$$

Among the $m-1$ observations below T, let n_+ and n_- count the two labels. Hoeffding's inequality gives

$$\Pr\{\min(n_+, n_-) < (m-1)/4\} \le 2e^{-(m-1)/8}.$$

On the complementary event, $n_+, n_- \ge m/8$ for $m \ge 2$.

The finite-sample direction. The population selected mean is $b(a_p) = \sigma\{r - \lambda(z_p)\}$. The identities

$$\lambda'(z) = -\lambda(z)\{z + \lambda(z)\} < 0, \qquad \frac{d}{dq}\lambda(\Phi^{-1}(q)) = -\frac{z_q + \lambda(z_q)}{q}$$

and the endpoint limits of λ establish the unique threshold $p^*(r)$. Below the threshold define $g = \lambda(z_{q_+}) - r > 0$ and $B = z_{q_+} + \lambda(z_{q_+}) > 0$. Conditional on T and the labels, both averages of the observations strictly below T differ from $b(T)$ by at most $\sigma g/4$, except with probability at most $4e^{-mg^2/256}$ on the class-count event. The boundary observation contributes only to its own class average. Its displacement from the conditional mean satisfies

$$0 \le T - b(T) \le \sigma B,$$

because $z + \lambda(z)$ is increasing: its derivative is the positive truncated-normal variance. Consequently the boundary changes that class's average of A by at most $8\sigma B/m$, in addition to the stated concentration error. If $m \ge 32B/g$, both selected class averages of A are at most $-\sigma g/2$. Since $\hat{w}$ is the sum of these two averages, $\hat{w} < 0$.

Above the threshold take $g = r - \lambda(z_{q_-}) > 0$. The same concentration calculation applies, and adding the boundary observation can only increase a class average relative to $b(T)$. Hence both selected class averages of A are positive, giving $\hat{w} > 0$, without the boundary-bias condition. In either case the failure probability is at most

$$2e^{-N\epsilon^2/2} + 2e^{-(m-1)/8} + 4e^{-mg^2/256}, \tag{*}$$

subject to the boundary-bias condition below the threshold.

To recover the stated sample-size condition, note that $g \ge c_r\epsilon$. Indeed, below the threshold compare q_+ with $p^* - \epsilon/2$, and above it compare q_- with $p^* + \epsilon/2$. The displayed derivative has magnitude bounded below on the compact interval $[p^*/2, (1+p^*)/2]$. Below the threshold, $B \le z^* + r$. Thus $m \ge C_r\epsilon^{-2}\log(C_r/\delta)$ absorbs both $m \ge 32B/g$ and all three tails in (*), using $N \ge m$ and increasing C_r if necessary. In particular, missing classes are included in the failure event.

Risk limits. For fixed $p \in (0,1)$, $T \to a_p$ in probability. The two selected averages of A share the conditional mean $b(T)$, so this common random term cancels from $\hat{\theta}$, which is half their difference. The conditional concentration above and the single boundary contribution imply $\hat{\theta}/\sigma = O_p(m^{-1/2})$. For negative fitted direction, with $t = \hat{\theta}/\sigma$, the clean test risk is

$$R_-(t) = \tfrac{1}{2}\{\Phi(r-t) + \Phi(r+t)\} = \Phi(r) - \tfrac{1}{2}r\phi(r)t^2 + O_r(t^4).$$

For positive fitted direction,

$$R_+(t) = \tfrac{1}{2}\{\Phi(t-r) + \Phi(-t-r)\} = \Phi(-r) + \tfrac{1}{2}r\phi(r)t^2 + O_r(t^4).$$

For fixed $p \ne p^*$, the probability of the incorrect direction tends to zero exponentially by (*). These expansions give the stated $O_p(m^{-1})$ rates. The theorem makes no assertion at $p = p^*$.

Random baseline. A uniform subset of an independent sample has, unconditionally over the data and selection, the law of m independent draws. With probability at least $1 - 2e^{-m/8}$, both class counts are at least $m/4$. Put $L = \log(8/\delta)$ and $a = \sqrt{8L/m}$. Conditional Gaussian concentration then places both class-mean errors within σa, except with probability at most $4e^{-ma^2/8} = \delta/2$. If $m \geq C \max\{1, r^{-2}\}L$, these events jointly have probability at least $1 - \delta$, $a \leq r/2$, $\hat{w} > 0$, and $|\hat{\theta}|/\sigma \leq a$. Since $R'_+(0) = 0$ and $|R''_+(t)| \leq 1/\sqrt{2\pi e}$, Taylor's theorem yields

$$0 \leq R_{\mathrm{rand}} - R^* \leq \frac{a^2}{2\sqrt{2\pi e}} \leq C\frac{\log(8/\delta)}{m}.$$

The sample-size condition also controls the missing-class event for the prescribed fallback.

Lemma 1 (Population logistic regression has the same threshold)**.** *Let the training distribution be the Gaussian model conditioned on $A = yx \leq a_p$. For any $\tau > 0$, the unique minimizer (w_τ, b_τ) of*

$$\mathbb{E}[\log(1 + e^{-y(wx+b)}) \mid A \leq a_p] + \frac{\tau}{2}(w^2 + b^2)$$

has $b_\tau = 0$ and $\mathrm{sign}(w_\tau) = \mathrm{sign}\{r - \lambda(z_p)\}$. Hence, away from the threshold, its clean test risk is $1 - R^*$ below p^* and R^* above it.

Proof. Ridge regularization makes the objective coercive and strictly convex. The selected distribution is invariant under $(x, y) \mapsto (-x, -y)$, so the objective is even in b and its unique minimizer has $b = 0$. At $w = 0$ the derivative of this one-dimensional strictly convex objective is $-\mathbb{E}[A \mid A \leq a_p]/2$. Its minimizer has the sign of that conditional mean, proving the claim. □

Symmetric label noise. For an independent sign B with $\Pr(B = -1) = \eta < 1/2$, let the observed label be $\widetilde{y} = By$, and score by $-\widetilde{y}x$. Then $\widetilde{A} = \widetilde{y}x$ is independent of $\widetilde{y}$ and has distribution $(1 - \eta)\mathcal{N}(\mu, \sigma^2) + \eta\mathcal{N}(-\mu, \sigma^2)$. For a normalized cutoff t, define

$$F_\eta(t) = (1 - \eta)\Phi(t - r) + \eta\Phi(t + r),$$
$$G_\eta(t) = (1 - \eta)\{r\Phi(t - r) - \phi(t - r)\} + \eta\{-r\Phi(t + r) - \phi(t + r)\}.$$

The selected positive observed-class mean is $\sigma G_\eta(t)/F_\eta(t)$. Since $G'_\eta(t) = tF'_\eta(t)$, $G_\eta(0) < 0$, and $G_\eta(\infty) = (1 - 2\eta)r > 0$, it has a unique finite zero $t^*_\eta > 0$. The population inversion threshold is $p^*_\eta = F_\eta(t^*_\eta)$. Increasing η strictly increases the CDF at every finite cutoff, hence strictly decreases each quantile and the lower-tail mean $p^{-1}\int_0^p F_\eta^{-1}(u)\,du$. Hence the zero crossing in p increases strictly with η. For example, solving this scalar equation gives $p^*_{0.1} \simeq 0.5597$ at $r = 1$. This is a population statement under the specified independent-noise and observed-label scoring model.

Higher dimensions. For $x \mid y \sim \mathcal{N}(y\mu e_1, \sigma^2 I_d)$ scored by $-yx_1$, the population selected class means are $\pm b(a_p)e_1$: the orthogonal coordinates are independent of the score and have zero conditional mean. Thus the population plug-in classifier has the same threshold. Restricting the learner to x_1 also preserves the one-dimensional finite-sample theorem. An unrestricted empirical plug-in learner estimates additional orthogonal mean coordinates. These fluctuate at scale $\sigma\sqrt{d/m}$, so the one-dimensional finite-sample guarantee does not apply unchanged. For fixed d and $p \neq p^*$, the estimated means nevertheless converge to their population values and the clean risk converges to the same limit.

C PROXY-DIRECTION MISMATCH: PROOF AND SCOPE

Proof of Theorem 2. Put $Z = yx$. Then $Z \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$ independently of y, and

$$Z = \boldsymbol{\mu} + \sigma(Tv + U), \qquad T \sim \mathcal{N}(0, 1), \quad U \sim \mathcal{N}(0, I_d - vv^\top), \quad T \perp U.$$

The score is $-v^\top Z$ and the selection event is $T \leq z_p = \Phi^{-1}(p)$. Hence selection is independent of the label, and the selected classes remain balanced. Since $\mathbb{E}[T \mid T \leq z_p] = -\phi(z_p)/\Phi(z_p) = -\beta_p$ and $\mathbb{E}[U \mid T \leq z_p] = 0$, the selected class means are $\pm a_p$ as in equation 1. Nearest-mean classification with these symmetric means is $h_p(x) = \mathrm{sign}(a_p^\top x)$ when $a_p \neq 0$. On a clean test observation,

$$ya_p^\top x \sim \mathcal{N}(a_p^\top \boldsymbol{\mu}, \sigma^2\|a_p\|^2).$$

Here $a_p^\top \boldsymbol{\mu} = \sigma^2(r^2 - rc\beta_p)$ and $\|a_p\|^2 = \sigma^2(r^2 + \beta_p^2 - 2rc\beta_p)$, proving equation 2. Because $r > 0$ and $\beta_p > 0$, the numerator is negative exactly when $c > 0$ and $\beta_p > r/c$. Hence monotonicity of the inverse Mills ratio gives the unique above-chance boundary $p^*(r/c)$ for $c > 0$.

For $|c| < 1$, put $q(\beta) = (r^2 - rc\beta)/(r^2 + \beta^2 - 2rc\beta)^{1/2}$. Its denominator is nonzero, and direct differentiation yields

$$q'(\beta) = -\frac{r^2\beta(1-c^2)}{(r^2+\beta^2-2rc\beta)^{3/2}} < 0 \quad (\beta > 0).$$

Since β_p decreases strictly from ∞ to 0 as p increases from 0 to 1, $R_p(v) = \Phi(-q(\beta_p))$ increases strictly under compression. The endpoint limits $q(0) = r$ and $q(\beta) \to -rc$ as $\beta \to \infty$ give the stated risks. The only possible zero of a_p requires $v = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and $\beta_p = r$. For $c = 1$ away from that point, the classifier has the Bayes direction if $\beta_p < r$ and its negative if $\beta_p > r$. For $c = -1$, $a_p = (1 + \beta_p/r)\boldsymbol{\mu}$, so the risk equals $\Phi(-r)$ for every $p \in (0, 1)$. □

Endpoints and ties. The theorem assumes $\boldsymbol{\mu} \neq 0$. If $\boldsymbol{\mu} = 0$, c is undefined and should not be used in equation 2. Every classifier evaluated on an independent clean observation has risk $1/2$, because its label is independent of its input. For $p = 1$, there is no selection and the population means are $\pm\boldsymbol{\mu}$, giving Bayes risk. For $p = 0$ there are no selected observations. Only the one-sided population limit is stated. The tie rule at $a_p = 0$ must be specified: always predicting one fixed label (or an independent fair coin) gives risk $1/2$. An arbitrary input-dependent tie breaker need not do so. Likewise, the finite-sample limit at the aligned critical point is not covered by the theorem.

Empirical top-m selection. For fixed d, $p \in (0, 1)$, and a fixed scoring direction v, the empirical selected means converge almost surely to $\pm a_p$ as $N \to \infty$ with $m/N \to p$. To see this, note that the empirical score cutoff converges to its continuous population quantile. Sandwich the selected sum between cutoffs $t_p - \varepsilon$ and $t_p + \varepsilon$: the norm of its difference from the fixed-cutoff sum is bounded by the empirical average of $\|x\|$ in that score strip. The strong law applies because Gaussian x has finite first moment. Then let $\varepsilon \downarrow 0$. Apply the same argument separately to each label and divide by its selected frequency, which converges to $p/2 > 0$. For $a_p \neq 0$, continuity of the clean Gaussian risk in the estimated means then gives convergence to equation 2. This argument supplies no dimension-uniform rate and no guarantee for a direction fitted on the same data. The numerical check uses the actual top-m operation and estimates both centroids and their midpoint.

Why the logistic conclusion is narrower. The selected distribution is invariant under $(x, y) \mapsto (-x, -y)$, so a population logistic objective with positive ridge penalty has a unique optimum with zero intercept. In the aligned case, rotational symmetry also restricts its coefficient to the signal direction, and the derivative at zero fixes its sign as in Lemma 1. Under mismatch, symmetry only restricts the coefficient to $\operatorname{span}\{\boldsymbol{\mu}, v\}$. The gradient at zero is $-a_p/2$, but in multiple dimensions a gradient at one point does not fix the minimizer's direction. Consequently neither the nearest-mean risk formula nor its above-chance boundary has been asserted for unrestricted logistic regression. Gaussian quadrature in the validation script checks an explicit case where their directions and risks differ. Specifically, for $\sigma = 1$, $\boldsymbol{\mu} = (1, 0)$, $v = (0.8, 0.6)$, $p = 0.3$, and ridge coefficient $\tau = 0.1$, deterministic quadrature and convex optimization give $w \simeq (-0.253316, -1.060421)$. Its clean risk is approximately 0.591865, whereas the nearest-mean risk is approximately 0.458526. The gradient residual is below 2×10^{-12}, and increasing both quadrature orders changes each fitted coefficient by less than 10^{-6}. This is a numerical example, not a separately proved logistic phase-boundary theorem.

C.1 Independent score noise and uniform-first mixing

Lemma 2 (Noisy scoring and the uniform-first quota)**.** *Under the Gaussian assumptions and notation of Theorem 2, score each observation by $S = -yv^\top x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau^2\sigma^2)$ independently, with $\tau \geq 0$. For a retained fraction $p \in (0, 1)$ in each class, allocate a share $q \in [0, 1]$ of its final quota to uniform draws, then fill the quota with the highest-scoring unused observations. In the population limit, the selected class means are $\pm(\boldsymbol{\mu} - \sigma b v)$, where*

$$a = \frac{p(1-q)}{1-pq}, \qquad b = \frac{(1-q)\lambda(\Phi^{-1}(a))}{\sqrt{1+\tau^2}} \quad (q < 1), \qquad b = 0 \quad (q = 1).$$

For nonzero selected centroid, the clean nearest-mean risk is

$$R = \Phi\left(-\frac{r^2 - rcb}{\sqrt{r^2 + b^2 - 2rcb}}\right).$$

It exceeds chance exactly when $c > 0$ and $b > r/c$. For $q = 0$ the critical fraction is $p^*(r\sqrt{1+\tau^2}/c)$ when $c > 0$. At fixed p and finite τ, b decreases strictly with q, so for $|c| < 1$ the risk decreases strictly with q. At fixed $q < 1$, reducing τ increases the risk for $|c| < 1$.

Proof. Write $Z = yx = \boldsymbol{\mu} + \sigma(Tv + U)$ as in the proof of Theorem 2. The standardized variable $W = (T - \epsilon/\sigma)/\sqrt{1+\tau^2}$ is standard normal and $\mathbb{E}[T \mid W] = W/\sqrt{1+\tau^2}$. Thus retaining the upper score fraction a shifts the mean of Z by $-\sigma\lambda(\Phi^{-1}(a))v/\sqrt{1+\tau^2}$.

For $q < 1$, let P denote the original within-class law of (Z, ϵ). The uniform stage takes fraction pq of each original class. The hard stage takes fraction $p(1-q)$ from the remaining fraction $1 - pq$, whose limiting distribution is unchanged by the uniform draw. Hence its tail fraction is a, and the selected population law is $qP + (1-q)P(\cdot \mid S$ in its upper a tail$)$. This gives b. The case $q = 1$ follows directly from uniform selection. The risk and chance condition follow by substitution into Theorem 2. At $c = 1, b = r$ both centroids vanish, and the stipulated fixed-label tie rule gives risk $1/2$.

For monotonicity, put $z = \Phi^{-1}(a)$ and rearrange:

$$(1-q)\lambda(z) = \frac{1-p}{p}\frac{\phi(z)}{1-\Phi(z)}.$$

The upper Gaussian hazard is strictly increasing in z, while z decreases strictly with q. The expression tends to zero as $q \uparrow 1$. Theorem 2's derivative then proves the risk claims, since b increases when τ decreases. □

For $0 < q < 1$, the effective tail a is smaller than p, so the uniform-first intervention leaves a larger shift than the fixed-tail mixture formula $(1-q)\beta_p/\sqrt{1+\tau^2}$. The lemma concerns a population limit. It does not assert that finite without-replacement samples are independent mixture draws, nor that risk is concave in q.

D RESIDUAL GEOMETRY MEASUREMENT

The probe is a ResNet-18 trained on full CIFAR-10 for 40 epochs with seed 0 (100% train accuracy, 85.78% test accuracy), and features are the 512-dimensional penultimate activations. For each method and seed at f=0.05 we form per-class means of the selected examples and compare them with the full-data class means after centering by the global mean. Table 5 reports per-method averages over 3 seeds and 10 classes. Two caveats bound what the measurement can see. Every method except Random selects exactly 250 examples per class under the class-balanced quota, so between-class counts carry no signal and the measurement describes within-class displacement rather than changes in class proportions. The measurements use only this highly fitted feature space. Full training accuracy does not establish how sensitively it measures rotation or displacement, and robustness to an independent feature representation is not assessed. Worst cases stay far from inversion: the lowest single-class cosine across all methods and seeds is 0.987 (EL2N, class deer), and the largest per-seed displacement is 0.076 of the median inter-centroid distance.

Direction of the displacement. Neither the cosine nor the displacement magnitude says which way a selected class mean moves, so we measure the signed margin directly. For class k let $j(k)$ be the nearest full-data centroid, $u = (\mu_k - \mu_j)/\|\mu_k - \mu_j\|$ the discriminant direction of the k-against-j nearest-centroid rule, and place the decision point at the midpoint of the two full-data centroids. The full-data margin is $M_k = \|\mu_k - \mu_j\|/2$ and the selected margin is $\hat{M}_k = (\hat{\mu}_k - (\mu_k + \mu_j)/2)^\top u$, so a retention $\hat{M}_k/M_k$ below one means the selected mean has moved toward the competing class. Table 2 reports the retention, the signed component of the displacement along u as a fraction of the k-to-j centroid distance, and the share of 30 class-and-seed cases whose displacement is boundary-ward. The loss-ranked family gives up 6 to 11% of the margin and does so in 88 of 90 cases, against a

margin that Random and Uniform keep to within 0.2%. Herding shifts in the same direction in every case but gives up only 2.7%, which is the same separation by magnitude that Table 5 shows. Three caveats apply. The baseline for every column is the measured Random and Uniform row rather than an isotropic null, because the probe features have low effective rank and even a random subset's displacement concentrates along u well above the isotropic level. The readout is a nearest-centroid boundary in probe feature space, not the decision surface of the downstream network. And the 90 cases are 10 classes crossed with 3 seeds on one dataset, so they share both the class geometry and the probe and are not 90 independent replicates, which is why we report them as a direction count rather than as a test. The measurement script will be released as `run_margin_direction.py`.

Table 5: Selected-class-mean geometry on CIFAR-10 at $f{=}0.05$ in probe feature space, mean over 3 seeds and classes: cosine between centered selected and full-data class means, and displacement $\|\hat{\mu}_k - \mu_k\|$ relative to the median inter-centroid distance.

	EL2N	Forgetting	GraNd	kCenter	Herding	Moderate	Random	Uniform
Cosine	0.9968	0.9983	0.9976	0.9992	0.9998	0.9998	0.9998	0.9998
Displacement	0.074	0.074	0.055	0.033	0.020	0.015	0.014	0.013

E Experimental details

All native experiments train a ResNet-18 from scratch on the selected subset with SGD (momentum 0.9, Nesterov, weight decay 5×10^{-4}) under a cosine-annealed learning rate, with one fixed recipe per dataset: CIFAR-10 (batch 256, initial lr 0.1, 200 epochs), Tiny ImageNet (batch 128, lr 0.1, 200 epochs), CUB-200 (batch 64, lr 0.01, 200 epochs), and ImageNet-1K (batch 256, lr 0.1, 90 epochs). Train-time augmentation is one fixed policy per dataset, a random crop with reflection padding and a horizontal flip on CIFAR-10 and Tiny ImageNet and a random resized crop to 224 with a horizontal flip on CUB-200 and ImageNet-1K, applied identically to every method and fraction, so a recipe difference cannot account for a gap between methods, though a shared policy does not rule out an interaction between augmentation and the selected distribution, which we do not audit. Except for the scoring-epoch and paired-composition studies, which use final accuracy as the primary endpoint, reported accuracy is each run's highest value over its test evaluations rather than its last, and the experiments carry no held-out validation split. Every comparison we report is between methods at a matched epoch count, so the number of evaluations is the same on both sides of each difference. Selection signals come from a from-scratch ResNet-18 probe trained on the same split for 40 epochs (10 on ImageNet-1K). GraNd and EL2N score after a single probe epoch, averaging 10 independently initialized probes. Paul et al. (2021) average over ten initializations in the same way but take their scores later, at epoch 10 and epoch 20 of a sweep that starts at initialization, so the single-epoch probe is this benchmark's protocol choice rather than their specification. Reuse targets train ViT-Tiny from scratch for 300 epochs under a DeiT-style recipe (AdamW with warmup and cosine schedule, RandAugment, label smoothing, stochastic depth). The m_k intervention subsamples Tiny ImageNet to the stated per-class counts before selection at $f{=}0.1$ and keeps the unified recipe. All experiments use seeds $\{0, 1, 2\}$.

F Quantitative checks of the Gaussian analogy

Fraction dependence. For each scoring epoch $t \in \{1, 10, 20\}$ and seed, we use the saved ten-probe mean score vector from the $f = 0.05$ rerun, hold that vector fixed, and select within-class top fractions $p \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$. This is a geometry-only reselection: it does not add trained accuracy measurements at the new fractions. Independently executed fraction folders contain different score vectors, so the reselected $p = 0.1$ sets are not asserted to equal the original $f = 0.1$ training sets. The fixed feature cache and nearest-competitor direction are those in Appendix D.

Write $d_k(p) = -(\hat{\mu}_k(p) - \mu_k)^\top u_k / \|\mu_k - \mu_{j(k)}\|$. Equation (1) motivates the one-scale description $\bar{d}(p) = A\beta_p$ when the score direction and effective geometry are fixed. We fit A by least squares without an intercept, using the seven class-and-seed-averaged displacements. The centered $R^2 = 1 - \sum_p (\bar{d}(p) - A\beta_p)^2 / \sum_p (\bar{d}(p) - \bar{\bar{d}})^2$ is $0.830/0.581/-0.030$ at epochs $1/10/20$ (Figure 6). A

negative value means this fit performs worse than the constant sample mean on these points. Pooling individual class-and-seed values instead gives $0.233/0.265/-0.007$. These are different summaries, because class-specific displacement scales vary. Both fits and all residuals will be released upon publication. A free-intercept fit gives higher mean-curve R^2 $(0.963/0.991/0.984)$, but changes the proposed proportional scaling and is reported only descriptively. None of these in-sample fits establishes isotropic Gaussian features or a downstream risk calibration.

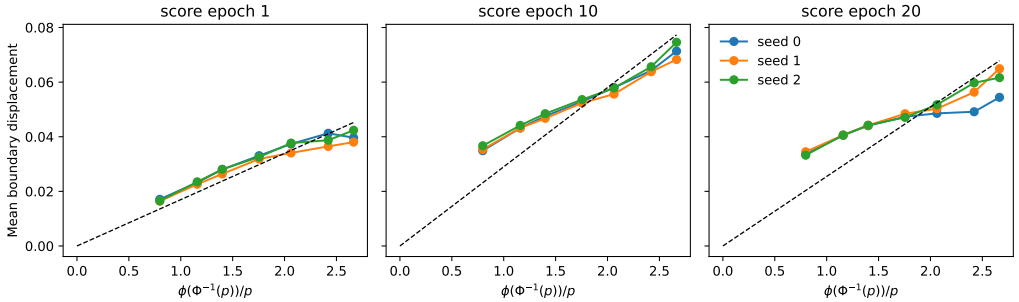

Figure 6: Retained-fraction check in the fixed feature space. Each seed curve averages ten class-wise normalized boundary displacements. Dashed lines are through-origin inverse-Mills fits. Seven reselected fractions share each underlying score vector. The early-score mean trend partly agrees, whereas later-score curves have systematic residuals.

Composition dependence. We add $q = 0.25, 0.75$ geometry measurements using exactly the uniform-first, no-duplicate algorithm of Appendix H. All 18 reconstructed sets at $q = 0, 0.5, 1$ match the saved training subsets. At the three interior shares, the absolute deviation of the class-and-seed mean normalized boundary displacement from its observed endpoint line is at most 0.0014 (Figure 7). Averaging absolute classwise deviations instead gives 0.0026–0.0043. The mean curve is approximately affine, with cancellation of class-specific deviations. This is descriptive agreement with endpoint interpolation, not the exact prediction of Lemma 2: the implemented rule changes the hard tail to $a = p(1-q)/(1-pq)$. Substituting its Gaussian attenuation into the endpoint comparison does not consistently improve agreement.

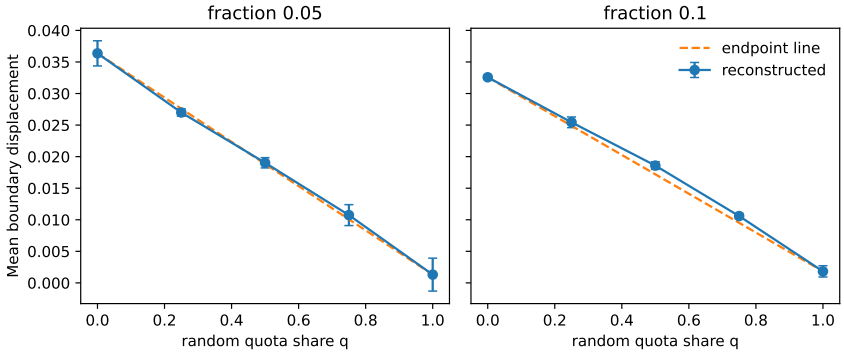

Figure 7: Geometry-only intermediate mixing shares. Points show class-averaged boundary displacement, with one sample standard deviation across three seeds. Dashed lines interpolate the two observed endpoints. They are not a fitted risk model. No additional downstream training was run at $q = 0.25, 0.75$.

Why no fitted noisy-score risk curve is reported. The saved score arrays average ten probes and omit the constituent score vectors and probabilities, so they do not supply across-probe correlation measurements. Even with those measurements, converting correlation to τ would require a common latent signal and independent additive noise, and estimating the risk formula would additionally require validating the representation and learner assumptions. We do not choose r, c, τ to match the

observed accuracy ordering. The scripts to be released upon publication retain the ensemble-mean repeatability statistics as descriptive quantities only.

G SENSITIVITY TO EL2N SCORING EPOCH

This experiment tests whether the low-fraction EL2N failure is confined to the benchmark's unusually early scoring point. We rerun CIFAR-10 with scoring epochs $t \in \{1, 10, 20\}$, fractions $f \in \{0.05, 0.1\}$, and seeds $\{0, 1, 2\}$, together with fresh Random and Uniform controls: 18 EL2N runs and 12 control runs. Each EL2N selection averages scores from ten independently initialized ResNet-18 probes and retains the highest-scoring examples within each class. We preserve the original scoring optimizer and all downstream hyperparameters: a fresh ResNet-18, 200 epochs, batch size 256, SGD with learning rate 0.1, momentum 0.9, Nesterov momentum, weight decay 5×10^{-4}, and cosine decay to 10^{-4}, with the same crop-and-flip augmentation. Scoring runs are separate at each epoch setting. The original random-number flow is retained rather than resetting the downstream random state after selection, so downstream initial weights are not held identical across epoch settings. This is a protocol sensitivity study, not a paired intervention on subsets with identical downstream randomness.

Later scoring improves EL2N substantially, but even epoch 20 remains 29.82 and 26.32 percentage points below Random in mean final accuracy at the two fractions, respectively. The mean improvement from epoch 1 to 20 is 18.02 and 17.64 points. Every epoch-20 run remains below its seed-matched Random and Uniform controls. The ranking also holds for peak accuracy. The one-epoch reruns reproduce severe underperformance qualitatively, not the exact historical accuracies, and are kept separate from the original benchmark tables.

Selection consistency also changes: mean pairwise cross-seed Jaccard decreases from 0.492 to 0.336 to 0.151 at $f = 0.05$, and from 0.520 to 0.387 to 0.209 at $f = 0.1$, across scoring epochs 1, 10, and 20. The corresponding rerun Random values are 0.023 and 0.052. Thus the accuracy improvement accompanies reduced agreement, but these observations do not establish that agreement causes poor accuracy or that the Gaussian sign-inversion mechanism explains the real-data gap. Figure 4 recomputes the fixed-feature geometry for these same subsets: its nonmonotone response does not track their monotone accuracy improvement. Margin retention is 0.927, 0.886, and 0.900 at $f = 0.05$ and 0.935, 0.892, and 0.900 at $f = 0.1$ for scoring epochs 1, 10, and 20. The epoch-1 value 0.927 is from this rerun batch, and the 0.930 in Table 2 uses the historical subsets in the same feature space. The experiment supports a restricted conclusion: early scoring affects the severity of EL2N failure, while the tested later scoring points still leave a large low-fraction gap under this recipe. No run clocks from this experiment are used as cost evidence.

H PAIRED INTERVENTION ON HARD-VERSUS-RANDOM COMPOSITION

We reuse the saved ten-probe epoch-1 EL2N scores for each fraction and seed from Appendix G. No scoring model is retrained. For class k with quota m_k, draw $\mathrm{round}(qm_k)$ examples uniformly without replacement from the entire class, then fill the remaining quota with the highest-scoring examples not already drawn. The three fixed shares are $q = 0, 0.5, 1$. This controls the quota allocated to random draws, not the exact number of changed identities, because random draws may include original top-scoring examples. We sort indices within each class before downstream training. At $q = 0$ we verify exact set equality with the epoch-1 rerun selection of Appendix G, and $q = 1$ is a fresh class-balanced random selection.

All 18 runs use the same 200-epoch ResNet-18 recipe as the scoring-epoch study. Within each fraction and seed, we reset the downstream random generators immediately before network construction. Initial-weight and post-initialization RNG hashes agree across all three shares. Equal subset sizes yield equal minibatch and update counts, and data loading uses the same seeded sequence by minibatch position. We retain the original cuDNN benchmark settings, so this is not a claim of bitwise deterministic optimization. Every run records 110 test evaluations, uses the final evaluation as its primary endpoint, and passes independent reconstruction of the saved selected indices. These are new paired runs. Their pure-EL2N values need not equal the earlier sweep, whose downstream initialization was not reset after scoring.

Table 6: Paired composition intervention: mean ± sample standard deviation of final accuracy over three seeds. Every cell uses the same per-fraction training-update budget.

Random quota share	$f = 0.05$	$f = 0.1$
$q = 0$ (EL2N)	15.02 ± 1.07	32.36 ± 0.57
$q = 0.5$	45.23 ± 4.13	58.45 ± 4.73
$q = 1$ (class-balanced random)	61.67 ± 1.04	72.68 ± 2.88

The half share improves on pure EL2N in every seed, by 21.30 to 33.41 percentage points, and remains below the $q = 1$ endpoint. The fixed 40-epoch feature cache is reused without fitting it to these outcomes. Class centroids, nearest competing centroids, normalized displacement, and margin retention follow Appendix D. Margin retention across $q = 0, 0.5, 1$ is 0.927, 0.962, and 0.997 at $f = 0.05$ and 0.935, 0.963, and 0.996 at $f = 0.1$. The intervention changes score distribution, examples, and other unmeasured subset properties jointly. The geometry is a readout rather than the randomized treatment. Thus the accuracy change supports the value of changing subset composition, but not mediation by any single diagnostic. All ordinary execution clocks are excluded from paper cost analyses.

I Uniform sampling as a statistical reference

Standard uniform convergence gives sufficient sample-size conditions under which most uniform subsets approximate a bounded loss class, and a union bound extends this to a finite menu of classes (Theorem 3, Appendix A). These are background results, not guarantees about the learned selectors tested here. Their capacity bounds are vacuous at our deep-network scales. Independent random subsets supply an overlap reference, but a large combinatorial family does not establish that two algorithms draw independently from it. Likewise, simultaneous approximation does not predict which method transfers best.

Unrestricted total variation is unsuitable for our scarcity comparison: when all empirical examples are distinct, every size-m_k subset has exactly $\mathrm{TV}(\hat{P}_k, P_{\mathcal{D},k}) = 1 - m_k/N_k$. At fixed fraction this is constant across methods and class counts, so we treat scarcity, overlap, and transfer as empirical diagnostics, without deriving them from that bound.

I.1 Overlap relative to chance and subset quality

For independent uniform subsets, pairwise Jaccard approaches $p/(2-p)$. We compare algorithm outputs to this reference without assuming independent selection. We quantify deviation with the excess ratio $\rho(p) = J_{\mathrm{obs}}/J_{\mathrm{rand}}$. On ImageNet the cross-method mean gives $\rho = 1.38$ at $p = 0.05$, 1.04 at $p = 0.3$, and 1.02 at $p = 0.7$, and Tiny ImageNet and CUB-200 reproduce the shape (Figure 8). Most of the low-fraction excess comes from a single pair of near-twin loss-ranked scorers. Without the EL2N-GraNd pair, the ImageNet mean at $p = 0.05$ drops to 1.04. The hypergeometric null band is very tight at these N, so the excess is unambiguous: methods share some preferred examples, whose value is not established by overlap alone. The ratio is constrained by $\rho \le (2-p)/p$, whose upper limit tends to one as $p \to 1$. Hence its decreasing shape alone does not establish increasingly independent selection. Moreover, this average includes unsuccessful selectors. It cannot by itself establish that different subsets are equally useful.

I.2 Consistency does not certify quality

The selectors that lose most to Random are among the most self-consistent across seeds, so repeatability is not evidence of a good subset. At f=0.05 on ImageNet, GraNd and EL2N have cross-seed Jaccard 0.43 and 0.34, an order of magnitude above the Random baseline of 0.026, while sitting at the bottom of the accuracy ranking. Figure 9 plots the accuracy edge over Random (peak test accuracy, Appendix E) against normalized cross-seed consistency for every method and dataset at this fraction. The harmful quadrant, high consistency with negative edge, is populated by the loss-ranked scorers and by kCenter. The latter uses geometric farthest-first traversal rather than loss scores, so this quadrant is not specific to a difficulty-ranking mechanism. The converse fails, and the plot shows it: GraphCut Submodular selection is fully deterministic, $J = 1$ across seeds, with

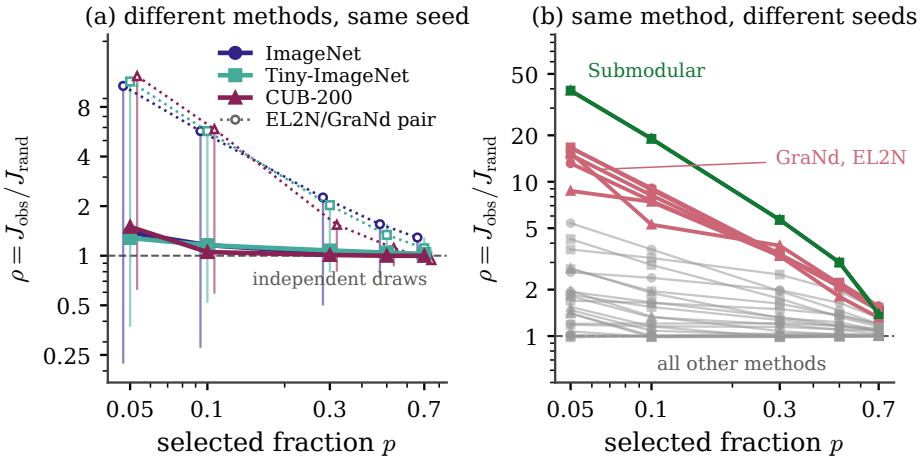

Figure 8: Excess overlap ratio $\rho(p) = J_{\text{obs}}/J_{\text{rand}}$ against fraction, with $J_{\text{rand}} = p/(2-p)$. Left: mean over all unordered method pairs per dataset (28 pairs on ImageNet, 55 on the others, seed-matched Jaccards averaged over 3 seeds) approaches one, partly constrained by $\rho \leq (2-p)/p$. This panel includes low-accuracy subsets. Whiskers span the min-to-max pair range, and the EL2N-GraNd pair, included in the mean, is also drawn alone as the outlier. Right: cross-seed overlap, one line per method and dataset averaged over the 3 seed pairs, separates loss-ranked selectors from the rest, as discussed in Appendix I.2.

mid-table accuracy. High agreement indicates repeatability, not necessarily a deterministic rule, and fixed-score determinism does not model independently trained scoring probes.

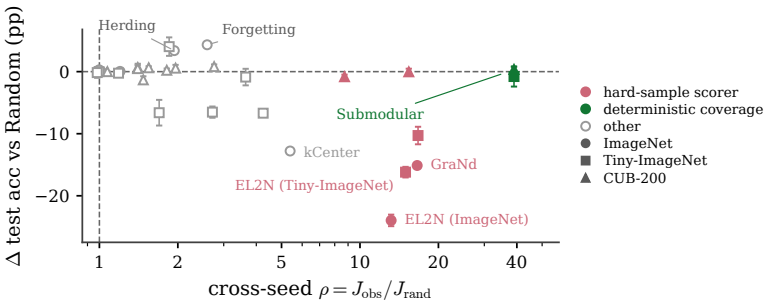

Figure 9: Accuracy edge over Random against normalized cross-seed consistency J/J_{rand} at f=0.05, one point per method and dataset. Loss-ranked scorers populate the high-consistency low-accuracy corner. Deterministic coverage selection (Submodular, J=1) is mid-table, separating determinism from harm.

I.3 Quality-conditioned overlap

Qualification uses three-seed mean peak test accuracy because complete final endpoints are unavailable for the historical grid. We retain methods within one percentage point below Random and report zero- and two-point sensitivity. This is a retrospective diagnostic, not a validation-selected method recommendation. For each seed-matched pair A, B with class counts $m_{A,k}, m_{B,k}$, define

$$E_0 I = \sum_k \frac{m_{A,k} m_{B,k}}{N_k}, \qquad O_{\text{adj}} = \frac{|A \cap B| - E_0 I}{\min(|A|, |B|) - E_0 I}.$$

The chance expectation assumes independent uniform draws within the observed class quotas. The denominator uses the smaller subset size. It need not equal the conditional maximum possible intersection when quotas differ. We validate the ImageFolder class layout against the class-concatenated Uniform index arrays. The artifact records the reconstruction constraints and checks

against 58,087 independently labeled cached ImageNet points. Seed-block means, rather than the dependent method pairs, define the uncertainty unit.

At $f = 0.05$ and 0.1, mean adjusted overlap is 0.0005 ± 0.0004 and 0.0017 ± 0.0003 on ImageNet, and -0.0005 ± 0.0004 and -0.0002 ± 0.0012 on Tiny ImageNet (sample standard deviation over three seed blocks, Figure 10). The five qualifying ImageNet methods yield 10 pairs, and the six Tiny ImageNet methods yield 15 pairs at each of these fractions. Zero- and two-point tolerance checks preserve the low-fraction near-chance conclusion. However, the ImageNet value reaches 0.0249 ± 0.0008 at $f = 0.7$, so excess intersection is not uniformly negligible. CUB-200 at $f = 0.05$ admits ten methods, including poorly performing loss-ranked selectors, because Random is itself near chance. Its adjusted mean is 0.0095 ± 0.0044 with a large EL2N-GraNd pair. Being close to Random here is not evidence of good absolute accuracy.

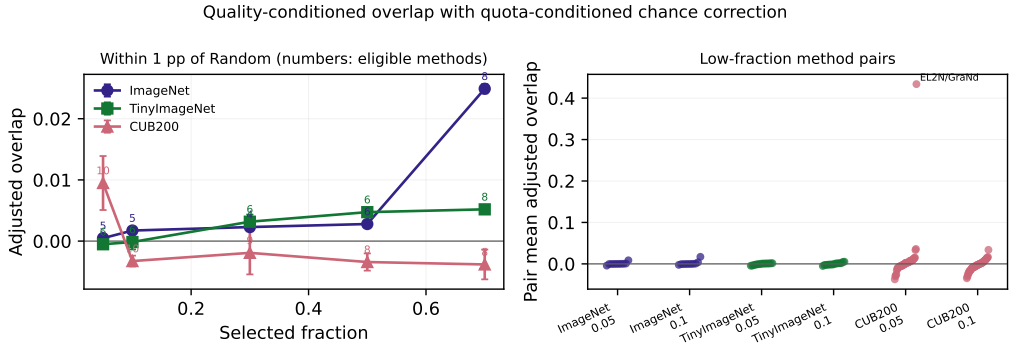

Figure 10: Performance-conditioned overlap beyond class-quota chance. The primary filter keeps methods whose mean peak accuracy is no more than one percentage point below Random, and zero- and two-point filters give sensitivity checks. Uncertainty averages method pairs within each seed before summarizing the three seed blocks. The low-fraction CUB-200 outlier and the positive high-fraction ImageNet endpoint delimit the near-chance finding.

I.4 Changing the downstream model and recipe changes method advantages

The reuse experiment holds subsets fixed and jointly changes the architecture and training recipe: indices selected with a ResNet-18 probe are reused to train ViT-Tiny from scratch. ResNet-18 uses SGD for 200 epochs, whereas ViT-Tiny uses AdamW for 300 epochs with DeiT-style augmentation and regularization. This comparison does not isolate architecture from optimization or augmentation. Figure 11 plots each method's accuracy edge over Random natively against the same edge after reuse. Herding's low-fraction edge survives with attenuation (Tiny ImageNet f=0.05: +4.0 pp native, +1.5 pp reused), Uniform and Moderate stay in the top cluster, and one clear empirical reversal is that Forgetting, the best of the six reuse methods under its native probe at f=0.3 with +2.5 pp, crosses zero and finishes 2.3 pp below Random after reuse. The flip is seed-consistent: Forgetting's edge is positive in all three seeds natively, negative in all three after reuse, and a paired per-seed test rejects no change ($t(2)$=14.8, p=0.005). Herding's reused edge stays positive in every seed at $f \in \{0.05, 0.1\}$, significant at f=0.1 (p=0.013) and underpowered at f=0.05. EL2N's collapse persists across these reuse settings. These are observations about transfer to a different model and recipe, not isolated architecture effects or corollaries of the uniform-sampling theorem.

J Per-class scarcity and method differences

We examine per-class sample scarcity as a separate empirical question. Most methods converge toward full-data accuracy at high fractions on the natural-image datasets, while CUB-200, whose $p = 0.1$ slice keeps three examples per class, separates methods at every fraction. The overlap data agree: at f=0.05 on CUB-200, 32 of 36 sophisticated method pairs overlap above the hypergeometric baseline (median excess +0.008) while the typical pair on Tiny ImageNet and ImageNet sits at or below it. But cross-dataset comparisons confound per-class count with resolution, class similarity,

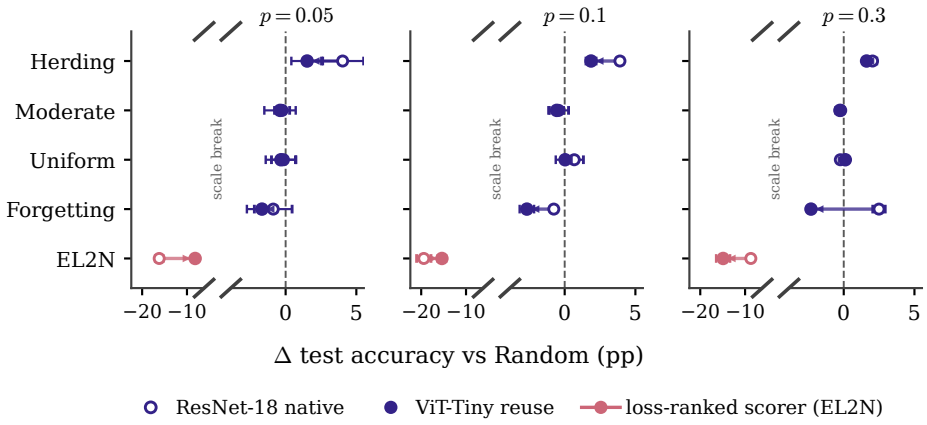

Figure 11: Accuracy edge over Random, native ResNet-18 training against reuse in ViT-Tiny training, per fraction on Tiny ImageNet with 3-seed error bars. Lines crossing zero mark advantages that invert after changing both model and recipe. CIFAR-10 counterpart in Figure 13.

and dataset size, so we intervene directly: subsample Tiny ImageNet to $N_k \in \{30, 100\}$ per class, keep the full dataset as the N_k=500 leg, and select at $f = 0.1$, giving $m_k \in \{3, 10, 50\}$ under one fixed distribution and recipe. This moves N_k and m_k together, so it isolates per-class count rather than m_k alone. The absolute accuracy spread among Uniform, Herding, and Moderate is 4.34, 2.66, and 0.97 percentage points at $m_k = 50, 10, 3$, respectively. Its ratio to the group mean instead increases, from 12% at m_k=50 to 25% at both smaller counts, and Herding's seed-matched edge over Uniform grows from 9% to 29% (Figure 12). At m_k=3 the full six-method ordering, Forgetting on top and EL2N at the bottom, reproduces the CUB-200 ordering at the same per-class count method for method. This ordering agreement is descriptive. Persistent method strengths across datasets make a uniform-permutation null inappropriate.

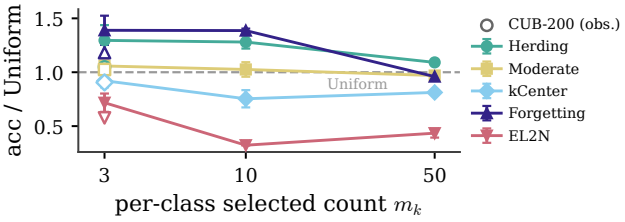

Figure 12: Accuracy relative to seed-matched Uniform against per-class selected count m_k, dataset family held fixed by per-class subsampling of Tiny ImageNet at f=0.1, error bars over 3 seed-paired ratios, open markers the CUB-200 observational point at its own m_k=3. Distribution-preserving selectors fan out as m_k falls, Forgetting tops the table at $m_k \leq 10$, and the m_k=3 ordering matches CUB-200 method for method.

K Additional figures

Figure 13 is the CIFAR-10 counterpart of Figure 11. The pattern matches Tiny ImageNet with one instructive difference: Herding has no native edge at low fractions on CIFAR-10 (−0.3 pp at f=0.05), compatible with the per-class-scarcity hypothesis.

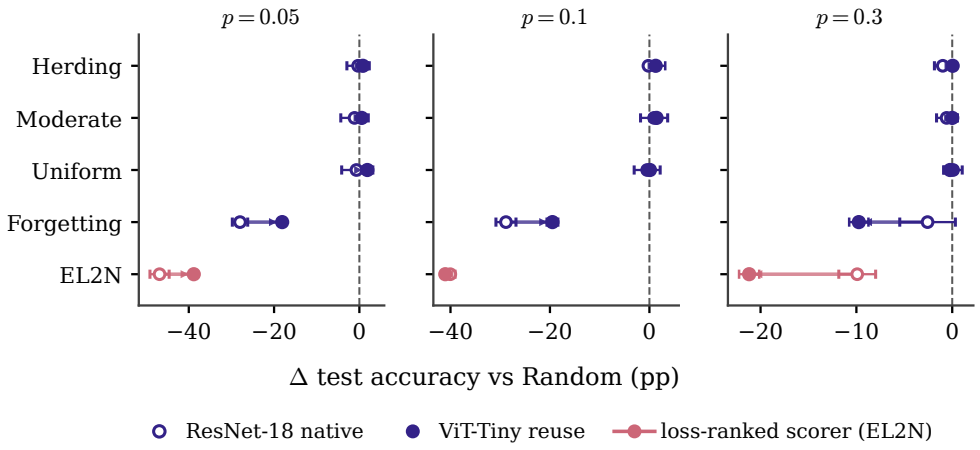

Figure 13: Accuracy edge over Random, native ResNet-18 training against reuse in ViT-Tiny training, per fraction on CIFAR-10 with 3-seed error bars.