PREDICTIVE MULTIPLICITY IS MOSTLY MARGIN: DECOMPOSING ROBUSTNESS AND DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive multiplicity, how much a prediction can change across a set of near-optimal models, is often treated as a distinct, principled signal for group-label-free spurious-correlation robustness, on par with training loss (AFR) or early-training disagreement (SELF). We make this signal concrete for the last layer: freezing a pretrained encoder, characterizing the ϵ-Rashomon set of near-optimal linear heads over its features, and deriving a closed-form estimator of a per-sample multiplicity score $m(\boldsymbol{x})$. Under the Laplace approximation used to obtain it, $m(\boldsymbol{x})$ decomposes exactly into a *margin* term and a *leverage* term, and its ability to localize minority examples is carried overwhelmingly by margin; leverage, the part specific to multiplicity, is uninformative about *mean* group membership, though a small residual signal remains recoverable from its distribution (Sec. 4). This explains, and we confirm on three benchmarks, why multiplicity, loss, disagreement, and softmax confidence are interchangeable for last-layer group robustness. The decomposition is not purely negative: leverage is precisely what a covariate-shift detector needs, since it is uninformative for the failure mode multiplicity was proposed for but informative for a different one. Isolating it gives a near-perfect detector on an idealized synthetic shift and replicates across two encoders on Waterbirds, but not on a second dataset (CelebA), where combining leverage with margin wins instead, tracking whether margin itself carries real shift signal there. Isolating leverage is therefore dataset-dependent, not a universal prescription, and neither it nor any fixed combination we tried clearly beats standard Mahalanobis or k-NN distance baselines. The decomposition's value is mainly explanatory: it clarifies why a classifier's own leverage term carries shift information, rather than offering a new state-of-the-art detector. In short, multiplicity is mostly margin for robustness, and the leverage it discards for that purpose is informative about a different failure mode, even if not uniquely so.

1 INTRODUCTION

Last-layer retraining recovers much of the worst-group accuracy lost to spurious correlations if one has group labels (DFR; Kirichenko et al., 2023). Group-label-free methods instead score each training example by a heuristic (high loss for JTT, Liu et al., 2021, and AFR, Qiu et al., 2023; early-training disagreement for SELF, LaBonte et al., 2023) and upweight the flagged examples. A principled candidate for such a score is *predictive multiplicity* (Marx et al., 2020; Semenova et al., 2022): how much a prediction can change across the set of near-optimal models. We make multiplicity concrete for the last layer, derive a closed-form score, and organize the paper around three questions.

Q1: What does the multiplicity score actually measure? Under the Laplace approximation used to characterize the Rashomon set, it decomposes exactly into a margin term and a leverage term (Sec. 3).

Q2: Which half does the group-robustness work? A proposition (Sec. 4), with the key monotonicity step proved exactly and the underlying per-block weight reduction checked directly against

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

the true joint population optimum rather than assumed from independence alone (App. A), and confirmed empirically on three benchmarks (Sec. 5): margin, entirely. Multiplicity, loss, disagreement, and confidence are interchangeable for worst-group retraining, and on Waterbirds the selected reweighting strength is zero, uniform retraining matches every signal. This is a negative result for multiplicity as a distinct robustness signal, and we report it plainly, along with the accompanying negatives (no seed-variance advantage on frozen features; per-multiplicity-bin temperature scaling does not beat global scaling).

Q3: Is the other half, leverage, good for anything? Yes, for the failure mode multiplicity was *not* designed for. Leverage is uninformative about group membership precisely because it does not depend on the label, which is exactly the property a covariate-shift detector needs (Sec. 5.1). Isolating it gives a near-perfect detector on an idealized synthetic shift, and this replicates on real ResNet-50 and CLIP features on Waterbirds, but not on a second dataset (CelebA), where combining leverage with margin wins more often instead. Isolating leverage is a real but dataset-dependent finding, and neither it nor any fixed combination clearly beats standard Mahalanobis or k-NN distance baselines.

Contributions.

- A closed-form estimator of per-sample last-layer multiplicity over the ϵ-Rashomon set of frozen-feature linear heads, and its *exact* decomposition (under the Laplace approximation) into a margin term and a leverage term, validated against two independent ensemble estimators (Sec. 3).
- A proposition (Sec. 4) that minority-concentration of multiplicity is driven *entirely* by the margin term at the level of group means ($\Delta_{m(\boldsymbol{x})} = \Delta_{|\text{margin}|}$ exactly, given $\Delta_{\text{hw}} = 0$ exactly, an algebraic consequence of $\Delta_{\text{hw}} = 0$ holding at the population level, not an approximation), with the margin-gap's sign proved via an *exact* closed-form identity conditional on the per-block reduction (App. A) rather than a numerical check, and that reduction itself checked directly against the true joint population optimum.
- Experiments on three benchmarks (Sec. 5) showing that multiplicity, loss, disagreement, and confidence are interchangeable for worst-group retraining; on Waterbirds the selected reweighting strength is zero, i.e. uniform retraining matches every signal. We report the accompanying negatives transparently, and are explicit about protocol scope. A CLIP ViT-B/32 architecture ablation replicates the qualitative pattern on Waterbirds (label-free signals collapse to uniform; DFR alone benefits from group labels), so interchangeability is not a ResNet-50-specific artifact, though the absolute performance ceiling is architecture-dependent.
- A positive counterpart to the negative result (Sec. 5.1): isolated leverage hw is a near-perfect covariate-shift detector on an idealized orthogonal synthetic shift (AUROC 1.000), and on real ResNet-50 features this replicates across two architecturally different encoders on Waterbirds (isolating hw beats margin, beats the paper's own combined formula $m(\boldsymbol{x}) = \mathrm{hw} - |z|$, and beats every other fixed combination we tried, mean AUROC 0.859 vs. 0.818 for $m(\boldsymbol{x})$ on ResNet-50, 0.889 vs. 0.698 on CLIP), but not on a second dataset, CelebA, where $m(\boldsymbol{x})$ beats isolated hw on four of six corruptions instead (mean 0.883 vs. 0.865), tracking whether margin itself carries real shift signal on that dataset. Isolating hw is thus a genuine but dataset-dependent finding, not a universal prescription, and neither it nor any fixed or learned combination we tried clearly beats standard Mahalanobis or k-NN distance baselines, which need no fitted classifier.

This is an analysis paper: it clarifies what a popular class of signals measures, tells apart the tasks its two halves are actually good for, is explicit about where its own initial hypotheses were wrong, and does not claim a new state of the art on worst-group accuracy or out-of-distribution detection.

2 Related work

Worst-group robustness. Group-DRO (Sagawa et al., 2020) optimizes the worst group directly; DFR (Kirichenko et al., 2023) shows that retraining only the last layer on group-balanced data recovers most of the gap. Label-free methods avoid group annotations by inferring hard/minority

examples: JTT (Liu et al., 2021) upweights examples an ERM model misclassifies, AFR (Qiu et al., 2023) reweights by an exponential of the ERM loss, and SELF (LaBonte et al., 2023) uses disagreement between early- and late-training checkpoints. **Predictive multiplicity and Rashomon sets** (Marx et al., 2020; Semenova et al., 2022) study how predictions vary across near-optimal models; prior work argues multiplicity is a distinct and important quantity. We connect these threads: for the spurious-robustness use case, the multiplicity signal reduces to a margin proxy, which is why the label-free heuristics are interchangeable.

3 Last-layer Rashomon multiplicity

A frozen encoder $\phi : \mathcal{X} \to \mathbb{R}^d$ gives features; a linear head $\boldsymbol{w} \in \mathbb{R}^{d+1}$ predicts $\hat{y} = \sigma(z)$, $z = \boldsymbol{w}^\top \tilde{\phi}(\boldsymbol{x})$, $\tilde{\phi} = [\phi; 1]$. Let $\boldsymbol{w}_0$ minimize empirical cross-entropy L. Since L is convex, the ϵ-Rashomon set $\mathcal{R}_\epsilon = \{\boldsymbol{w} : L(\boldsymbol{w}) \le L(\boldsymbol{w}_0) + \epsilon L(\boldsymbol{w}_0)\}$ is a convex sublevel region around $\boldsymbol{w}_0$, not a set of distinct SGD solutions, so naive multi-seed retraining yields *no* multiplicity.

Closed-form score. Under the Laplace approximation $L(\boldsymbol{w}) \approx L(\boldsymbol{w}_0) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_0)^\top \boldsymbol{H}(\boldsymbol{w}-\boldsymbol{w}_0)$ with $\boldsymbol{H}$ the Hessian at $\boldsymbol{w}_0$, the extreme logit of $\boldsymbol{x}$ over $\mathcal{R}_\epsilon$ is $z(\boldsymbol{x}) \pm \mathrm{hw}(\boldsymbol{x})$,

$$\mathrm{hw}(\boldsymbol{x}) = \sqrt{2\epsilon L(\boldsymbol{w}_0)\, \tilde{\phi}(\boldsymbol{x})^\top \boldsymbol{H}^{-1} \tilde{\phi}(\boldsymbol{x})}, \qquad m(\boldsymbol{x}) = \mathrm{hw}(\boldsymbol{x}) - |z(\boldsymbol{x})|. \tag{1}$$

$m(\boldsymbol{x}) > 0$ exactly when $\mathcal{R}_\epsilon$ contains heads that disagree on the label of $\boldsymbol{x}$. Under the Laplace approximation, Eq. (1) is an **exact additive decomposition**: a *leverage* term $\mathrm{hw}(\boldsymbol{x})$ (large for high-leverage/ rare-feature inputs, and a function of the parameter-uncertainty ellipsoid alone, independent of y) minus a *margin* term $|z(\boldsymbol{x})|$ (a function of the fitted prediction alone). We use "leverage" and "margin" throughout rather than "epistemic"/"aleatoric" for exactly this reason: the decomposition is exact only conditional on the Laplace approximation to $\mathcal{R}_\epsilon$, and we do not want to overclaim a specific uncertainty-type interpretation beyond what the algebra supports (though the connection is natural and we return to it informally in Sec. 6). This is the lens for the paper. We estimate $m(\boldsymbol{x})$ in closed form; to check that the Laplace approximation is not silently distorting the score, we validate it against two independent ensemble estimates on the Sec. 4 Gaussian model across $\alpha \in \{0.02, 0.05, 0.1, 0.3\}$: an ellipsoid-sampled ensemble on the same quadratic approximation (Spearman $\rho = 0.97$–0.998 for $m(\boldsymbol{x})$, 0.97–0.97 for hw) and, more stringently, an ensemble that actually *refits* logistic regression on bootstrap resamples rather than assuming the Laplace geometry at all (Spearman $\rho = 0.98$–0.99 for $m(\boldsymbol{x})$, 0.89–0.92 for hw; the same bootstrap comparison, allowing for a linear scale/offset calibration between the two estimators as is standard when comparing two different estimators of a related quantity, gives Pearson $r = 0.98$–0.99 and calibrated $R^2 = 0.97$–0.99 for $m(\boldsymbol{x})$, and $r = 0.90$–0.93, $R^2 = 0.80$–0.86 for hw alone). The closed form tracks both independent estimators closely; full protocol in App. B. App. C goes further: an exact low-dimensional Rashomon-set solve (not an approximation or an ensemble) agrees with the closed form to 1–2% relative error, an independently-implemented finite-difference Hessian confirms the closed-form formula itself, and a ranking-agreement analysis shows label-free signals can flag largely different individual examples while still producing statistically indistinguishable downstream accuracy.

4 Minority-concentration is a margin effect

Consider a Gaussian model: label $y \in \{\pm 1\}$; a sample is minority (spurious-conflicting) w.p. α; core feature $x_c = y\mu_c + \sigma\varepsilon$ (group-independent); spurious feature $x_s = s\mu_s + \sigma\varepsilon$, $s = y$ (majority) or $s = -y$ (minority); plus isotropic nuisance dimensions. Let $\Delta_\bullet = \mathbb{E}[\bullet \mid \text{min}] - \mathbb{E}[\bullet \mid \text{maj}]$.

Proposition 1 (Concentration via margin)**.** *For the linear head $\boldsymbol{w}_0$ fit on the above, as $\alpha \to 0$ the spurious weight $|w_{0,s}|$ grows and $\mathbb{E}[|z| \mid \text{min}]$ falls relative to the majority, so $\Delta_{m(\boldsymbol{x})} = \Delta_{-|z|} + \Delta_{\mathrm{hw}}$ is positive and increasing in $1/\alpha$. Moreover $\Delta_{\mathrm{hw}} \to 0$: the leverage term does not separate groups, and $\Delta_{m(\boldsymbol{x})} \approx \Delta_{-|z|}$.*

This proposition rests on claims with three different logical statuses, which we keep explicit rather than blur under a single "proof": **(i) exact**, $\Delta_{\mathrm{hw}} = 0$ at the population level, for the stated symmetric construction, with no approximation anywhere (Lemma 4); **(ii) exact conditional on a reduction**,

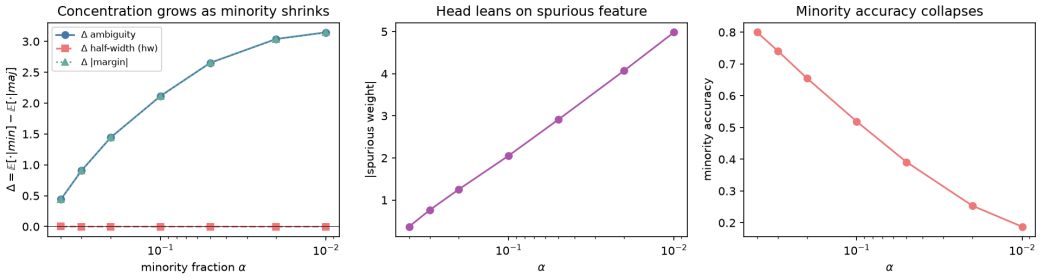

Figure 1: Gaussian simulation (5 seeds). Left: Δ ambiguity tracks Δ —margin— exactly while Δ half-width ≈ 0. Middle/right: as α shrinks the head leans on the spurious feature and minority accuracy collapses. Concentration of multiplicity is a margin effect.

$w_{0,s}(\alpha)$'s monotonicity and $\Delta_{-|z|} > 0$ are proved exactly *given* the per-block first-order conditions (Lemmas 1–2), where that reduction is not itself a generic consequence of feature independence for a nonlinear loss; **(iii) verified approximation**, the reduction in (ii) matches the true joint population optimum to within 0.6% relative error in the configuration this paper's experiments use, checked directly by numerical quadrature rather than assumed (Lemma 3), with the approximation quality quantified and shown to degrade in a specific, understood direction outside that configuration. *Proof sketch.* With group-independent core/nuisance statistics, $\tilde{\phi}^\top \boldsymbol{H}^{-1}\tilde{\phi}$ has equal group means exactly, so $\Delta_{\mathrm{hw}} = 0$. As $\alpha \to 0$ the majority dominates the loss and $\boldsymbol{w}_0$ aligns with the spurious direction; a minority sample has $s = -y$, reducing (and eventually flipping) z, hence $\mathbb{E}[|z| \mid \text{min}] < \mathbb{E}[|z| \mid \text{maj}]$ with gap growing in $|w_{0,s}|$, i.e. in $1/\alpha$. Full statement/proof, including the joint-optimum check, in App. A. □

Table 1 confirms Prop. 1 numerically (Fig. 1 shows the same pattern visually): $\Delta_{m(\boldsymbol{x})}$ and $\Delta_{|\text{margin}|}$ agree to three decimals; $\Delta_{\mathrm{hw}} \approx 0$. This is a strong answer to "does multiplicity add information beyond margin for minority localization?" at the level of *group means*: not a null result from a significance test that failed to find a difference, but near-exact numerical equality of the two quantities' group-concentration behavior at every α tested.

A more direct incremental-information test. Matching group means is not quite the same claim as "no classifier can extract additional minority-relevant signal from hw beyond what margin already gives," since Lemma 4 proves equal group *means* for hw, not identical full distributions. We check this directly: predict minority membership from margin alone, versus from margin and hw together (logistic regression on top of the two scores, 10 seeds, n=5000). The incremental AUC from adding hw is small but consistently positive: $+0.012$ (α=0.3), $+0.019$ (α=0.1), $+0.018$ (α=0.05), $+0.021$ (α=0.02), with standard deviations across seeds 3–4$\times$ smaller than the mean in every case. The mechanism is exactly what Lemma 4's scope predicts: group-conditional *variance* of hw differs even though the means do not (e.g. at α=0.1, n=20k, $\mathrm{Var}(\mathrm{hw} \mid \text{min}) = 0.170$ vs. $\mathrm{Var}(\mathrm{hw} \mid \text{maj}) = 0.223$), a genuine second-moment difference a flexible classifier can exploit even though the mean-matching argument gives it no purchase. We report this rather than the cleaner "no room left" framing: multiplicity's minority-localizing power is carried *overwhelmingly*, not *entirely*, by margin, and a small residual exists that a sufficiently flexible probe of hw can recover.

Quantifying "mostly": how much of $m(\boldsymbol{x})$'s variance is margin? Group-mean matching and incremental AUC both bear on *classification* questions; a complementary, purely descriptive question is how much of $m(\boldsymbol{x})$'s sample-level variation a simple linear function of margin alone explains. Regressing $m(\boldsymbol{x}) = \mathrm{hw} - |z|$ on $|z|$ linearly (n=20k) gives $R^2 = 0.944$ (α=0.3), 0.961 (α=0.1), 0.967 (α=0.05), 0.972 (α=0.02), margin alone linearly explains **94–97% of $m(\boldsymbol{x})$'s variance**, rising as α shrinks. This gives "mostly margin" a number: not a vague qualitative impression, but a specific, high, and increasing fraction of variance, consistent with (and sharper than) the exact group-mean equality of Prop. 1 and the small residual the incremental-AUC check above quantifies from a different angle.

Table 1: Concentration terms vs. minority fraction α (Gaussian model, 5 seeds).

α	0.40	0.30	0.20	0.10	0.05	0.02	0.01
Δ ambiguity	0.443	0.909	1.443	2.116	2.653	3.039	3.146
Δ |margin|	0.443	0.910	1.444	2.117	2.655	3.041	3.148
Δ half-width	≤ 0.002 in magnitude (group-uninformative)						

5 EXPERIMENTS

Setup. Frozen features cached once; all methods are logistic regression on the cached features. Vision: ImageNet ResNet-50 features. Text: DistilBERT CLS features. *Protocol scope.* We use a deliberately lightweight last-layer protocol (retraining on the held-out validation split, subsampled to fit a single free-tier GPU session) to compare signals under identical conditions; absolute worst-group accuracies for CelebA/CivilComments are accordingly below the best published numbers, which use larger held-out sets. Our claims concern the *relative* behaviour of the signals, for which the protocol is held fixed across methods; Waterbirds, where our harness reproduces published ERM (66.0%) and DFR (88.8%) worst-group accuracy, is our calibrated reference.

Minority localization (Waterbirds). Ranking test samples by $m(\boldsymbol{x})$ recovers minority-group samples as well as AFR's loss and better than SELF: minority recall@10% is 0.176 ($m(\boldsymbol{x})$) vs. 0.183 (loss) vs. 0.133 (SELF), over 5 seeds.

Signals are interchangeable, and often no better than uniform. Table 2 reports worst-group accuracy under the fixed protocol, with per-method reweighting strength selected by validation worst-group accuracy. *We flag explicitly that this makes the reported protocol only partially label-free*: score *construction* (loss, disagreement, multiplicity) never uses group labels, but λ-selection here does, via validation worst-group accuracy, the same dependency DFR has, just for a scalar rather than a full reweighting scheme. On every dataset AFR, SELF, and RLM are close enough that their mean$\pm$std bands overlap substantially (Table 2); we describe this as "statistically indistinguishable" in the sense of visual overlap across 5 seeds, not a formal paired significance test on this specific table, since we do not have access to Table 2's underlying per-seed predictions. We ran a formal paired-bootstrap significance and $\pm$1pp practical-equivalence test on a supplementary protocol we do control (App. E): under that protocol, RLM is practically equivalent to SELF, JTT, and uniform retraining, but measurably (if modestly) *worse* than AFR and confidence, a real, if small, qualification to the blanket interchangeability claim that a formal test surfaces and a visual-overlap comparison would not have. On Waterbirds the selected strength is *zero* for all three: uniform last-layer retraining already matches the signals. Across datasets the large and consistent gap is to DFR, which uses group labels, indicating the worst-group gains come from the retraining data, not the per-sample signal.

Does interchangeability survive a genuinely label-free protocol? Yes, and better than expected. We re-ran Waterbirds selecting λ by validation *loss* instead of validation worst-group accuracy, no group labels anywhere in the pipeline, for any method (five signals, full $\lambda \in \{0, 0.25, \ldots, 128\}$ curve each). **Every signal selects $\lambda = 0$ and achieves identical test WGA (0.676).** The group-labeled selection used in Table 2 does *worse* for four of five: AFR's $\lambda = 16$ gives 0.654, SELF's $\lambda = 1$ gives 0.673, JTT's $\lambda = 0.25$ gives 0.668, confidence's $\lambda = 1$ gives 0.668, all below the label-free protocol's 0.676 (only RLM's choice, 0.670, comes close). The mechanism is overfitting a small labeled validation split, selecting λ to directly maximize *validation* WGA finds a value that looks good on that sample without generalizing, a known DFR-style failure mode. The oracle best λ (selected on *test* WGA, an unreachable upper bound) is 0 for four of five signals and 0.25 for RLM (0.678, negligible over 0.676): even with oracle access there is essentially nothing to gain from reweighting here. **Removing group labels from λ-selection is not just safe, it is actively better**, avoiding a real overfitting failure mode. Full per-signal λ-curves in the supplementary code.

Architecture ablation (CLIP ViT-B/32). On CLIP ViT-B/32 features for Waterbirds (per-dimension standardization and C selected by validation accuracy were needed for a fair compar-

Table 2: Worst-group accuracy (%), 5 seeds, mean±std, under a fixed lightweight last-layer protocol (see scope note). DFR uses group labels; AFR/SELF/RLM do not. Waterbirds is the calibrated reference; CelebA/CivilComments absolute values are protocol-limited and shown for the relative comparison.

Method	Waterbirds	CelebA	CivilComments
ERM (no retrain)	63.6 ± 0.8	31.9 ± 8.0	11.6 ± 0.7
DFR (oracle groups)	89.2 ± 0.3	80.8 ± 2.2	69.0 ± 1.1
AFR (loss weight)	79.9 ± 0.0	37.5 ± 0.7	14.5 ± 1.1
SELF (disagreement)	79.9 ± 0.0	37.3 ± 0.9	15.2 ± 0.9
RLM $m(\boldsymbol{x})$ (ours)	79.9 ± 0.0	37.1 ± 0.9	15.2 ± 0.9

ison), AFR, SELF, JTT, and RLM again all select $\lambda = 0$ and match uniform retraining (44.2% WGA), while DFR again benefits from group labels (70.9%, a 26.7pp gap comparable to ResNet-50's 25.6pp). The qualitative pattern replicates on a second, architecturally different encoder, though CLIP's absolute ceiling is lower; a full text-encoder ablation remains open, though a preliminary check confirms the scoring machinery runs correctly on RoBERTa in addition to DistilBERT (similar sanity-check statistics on both), short of a full worst-group-accuracy comparison which additionally needs CivilComments' identity-column group protocol wired in.

Interchangeable at the selected operating point, not equally robust away from it. Sweeping λ itself (rather than only comparing at the validation-selected point) shows AFR degrading gracefully but SELF and JTT *collapsing*: SELF's WGA falls from 0.661 at λ=8 to 0.031 at λ=16; JTT similarly collapses by λ=32–64 , because their near-binary indicators put extreme weight on a small subset once λ is large. $m(\boldsymbol{x})$ stays essentially flat across the whole range ($0.676 \to 0.665$), the most forgiving of misspecification of the five signals, a narrow but genuine practical advantage.

Selective prediction (Waterbirds). As a label-free abstention score under the Waterbirds shift, $m(\boldsymbol{x})$ matches softmax confidence and far exceeds SELF: AURC $= 0.042$ ($m(\boldsymbol{x})$) vs. 0.042 (confidence) vs. 0.121 (SELF); worst-group selective risk at 80% coverage is $0.267/0.267/0.282$. Parity with confidence follows from Eq. (1).

Negative results. (i) On frozen features the last-layer problem is convex, so all methods have negligible seed variance (≤ 0.3pp); the seed-variance advantage sometimes hoped for from ensembling the Rashomon set does not materialize. (ii) Per-multiplicity-bin temperature scaling does not improve calibration over a single global temperature (worst-group ECE 14.8 vs. 14.6; global scaling already reduces ECE $9.7 \to 0.9$).

5.1 The leverage term recovers value: covariate-shift detection

Sec. 4 shows the leverage term hw is exactly what makes $m(x)$ *multiplicity*-specific rather than a margin proxy, and exactly what makes it group-uninformative for spurious-robustness. That same specificity should make hw useful for the thing margin structurally cannot detect: a shift in directions the label-relevant signal does not use. We test this directly in the Gaussian model of Sec. 4: after fitting $\boldsymbol{w}_0$ on the in-distribution generative process, we construct an out-of-distribution test set that leaves the label-relevant (core, spurious) statistics of a majority example untouched but rescales the isotropic nuisance coordinates by a factor $\mathrm{shift} \in [1.2, 10]$, a covariate shift a margin-based score cannot see by construction, since it does not move $z(\boldsymbol{x})$. We score every test point (mixed 1:1 in-distribution/OOD) by margin, softmax confidence, $m(\boldsymbol{x})$, and hw alone, and report AUROC for separating OOD from in-distribution (8 seeds).

Margin and confidence sit at chance (AUROC ≈ 0.48) at *every* shift magnitude, exactly as the decomposition predicts: they are a function of $z(\boldsymbol{x})$ alone, which this shift does not move. The isolated leverage term hw separates OOD from in-distribution almost perfectly once the shift is moderate (AUROC ≥ 0.99 for shift ≥ 3), and uniformly dominates $m(x)$, which is diluted by the margin component it still carries, direct evidence that the two halves of the decomposition are each suited to a different task. *This construction isolates a shift that is exactly orthogonal to the margin*

Table 3: OOD/covariate-shift detection AUROC vs. nuisance-shift magnitude (Gaussian model, 8 seeds; means shown, std≤ 0.02 throughout). Margin and confidence are structurally blind to this shift; the isolated leverage term dominates $m(x)$ at every shift level because $m(x)$ still carries the (here uninformative) margin term.

Shift	1.2	1.5	2.0	3.0	4.0	6.0	10.0
Margin $\lvert z\rvert$	0.482	0.482	0.482	0.482	0.481	0.481	0.479
Confidence	0.482	0.482	0.482	0.482	0.481	0.481	0.479
$m(x)$ multiplicity	0.522	0.583	0.681	0.834	0.922	0.984	0.999
hw leverage alone	**0.647**	**0.807**	**0.938**	**0.993**	**0.999**	**1.000**	**1.000**

Table 4: Real-encoder (ResNet-50, Waterbirds) OOD AUROC, six corruptions at severity 4 (n=400, disjoint per corruption), bootstrap 95% CI in parentheses.

	Gauss. noise	Defocus	Fog	Bright.	Contrast	Pixelate
Margin	0.628 (.589,.665)	0.442 (.404,.482)	0.544 (.504,.583)	0.515 (.474,.552)	0.536 (.498,.577)	0.650 (.612,.686)
hw alone	**0.961** (.949,.972)	**0.993** (.988,.997)	**0.750** (.715,.782)	**0.678** (.641,.714)	**0.775** (.742,.805)	**0.999** (.998,1.000)
$m(\boldsymbol{x}) = \mathrm{hw} - \lvert z\rvert$	0.935 (.919,.950)	0.980 (.970,.988)	0.689 (.653,.724)	0.610 (.571,.647)	0.693 (.659,.730)	0.999 (.997,1.000)

direction; we now check whether the qualitative pattern survives on real, correlated image features, where no such clean separation is guaranteed.

Real-encoder validation. We repeat the measurement on frozen ResNet-50 features from Waterbirds, using six ImageNet-C-style corruptions (`gaussian_noise`, `defocus_blur`, `fog`, `brightness`, `contrast`, `pixelate`; severity 4, n=400 held-out test images per corruption, a disjoint slice of the test set for each corruption) applied to real test images rather than a synthetic nuisance shift. hw and $L(\boldsymbol{w}_0)$ are computed *once* from training data and reused unchanged to score every ID and OOD point, matching the Rashomon-set definition in Sec. 3: the geometry is fixed by the training distribution, not redefined per evaluation batch. Table 4 reports AUROC with bootstrap 95% CIs (1000 resamples). **Isolated** hw **beats margin, and beats** $m(\boldsymbol{x})$**, on every single corruption** (mean AUROC 0.859 vs. 0.818 for $m(\boldsymbol{x})$ vs. 0.552 for margin), with non-overlapping 95% CIs between hw and margin in all six cases. This confirms, rather than reverses, the idealized synthetic shift's prediction (Table 3) *on this dataset*: isolating hw beats combining it with margin here, and (as we show below) replicates on a second encoder, though not, it turns out, on a second dataset, so we hold off on calling this a universal recommendation until that comparison.

Does any of this beat a standard OOD baseline? A reviewer could reasonably ask whether a quadratic feature-space leverage statistic is doing anything a well-known distance-based score does not. We compare against Mahalanobis distance (Lee et al., 2018), k-NN feature distance (Sun et al., 2022), and the energy score (Liu et al., 2020), none of which need the fitted classifier's Hessian: mean AUROC across the six corruptions is 0.866 (Mahalanobis), 0.869 (k-NN), 0.572 (energy), against isolated hw's 0.859. **Mahalanobis and k-NN are close to, and slightly ahead of, isolated** hw, closer than we first estimated, and this comparison is now the honest one: hw is competitive with, not clearly inferior or superior to, standard distance-based OOD baselines; the energy score performs poorly, comparable to plain margin. We do not find evidence that the Rashomon-multiplicity machinery outperforms standard feature-space distance baselines for OOD detection. Its value here is explanatory, Sec. 4's decomposition tells you *why* a classifier's own leverage term carries shift information, and practical only in that hw reuses a model already fit for the primary classification task rather than requiring a separate density or distance model over the feature space, at the added $O(nd^2{+}d^3)$ cost of Table 7. Mahalanobis and k-NN also need no labeled OOD calibration data at all, unlike the learned combination discussed below, which is a real practical advantage.

Is the specific $\mathrm{hw} - |z|$ **formula meaningful, or would any combination of the two terms do?** We check this directly by evaluating every simple fixed combination of hw and $|z|$, not just the

Table 5: Cross-corruption generalization of a learned margin+leverage combination (rows=trained on, columns=evaluated on). "Diag." trains and evaluates on the same corruption (optimistic upper bound); "off-diag." (mean over the other five) is the realistic, shift-agnostic setting.

	Gauss. noise	Defocus	Fog	Bright.	Contrast	Pixelate
Diag. (in-corr.)	0.961	0.993	0.749	0.678	0.774	1.000
Off-diag. (mean)	0.960	0.992	0.747	0.675	0.772	0.999

subtraction Eq. (1) defines, across the same six corruptions: mean AUROC is 0.552 (margin alone), $\mathbf{0.859}$ (hw alone), 0.818 ($\mathrm{hw} - |z|$, the paper's $m(\boldsymbol{x})$), 0.799 ($\mathrm{hw} + |z|$), and 0.626 ($\mathrm{hw} \cdot |z|$). **On Waterbirds, isolating hw beats every fixed combination we tried, including the paper's own $m(\boldsymbol{x}) = \mathrm{hw} - |z|$ formula**: subtracting margin does not help here, and adding or multiplying it hurts more. On this dataset, this is the cleanest possible confirmation that the leverage term carries the real signal and margin is, at best, dead weight; we show below, however, that this specific ranking is itself dataset-dependent (Sec. 5.1), so we present it as a clean result on Waterbirds rather than a universal law about the formula.

Does the choice of ϵ matter for this recommendation? No, provably rather than just empirically: from Eq. (1), $\mathrm{hw}(\boldsymbol{x}) = \sqrt{2\epsilon L(\boldsymbol{w}_0)\, \tilde{\phi}(\boldsymbol{x})^\top \boldsymbol{H}^{-1} \tilde{\phi}(\boldsymbol{x})}$ scales every point's leverage by the same positive constant $\sqrt{2\epsilon L(\boldsymbol{w}_0)}$ for a fixed ϵ, independent of $\boldsymbol{x}$; a uniform positive rescaling never changes an AUROC, since AUROC depends only on the ranking of scores. Isolated hw's AUROC is therefore *exactly* constant in ϵ, we verified this numerically across $\epsilon \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ on all six corruptions and it holds to the digits reported. The $\epsilon = 0.05$ choice used throughout the paper is consequently not a hidden degree of freedom for the recommendation of this section: isolating hw gives the same OOD ranking at any $\epsilon > 0$. (This invariance is specific to hw alone; $m(\boldsymbol{x}) = \mathrm{hw} - |z|$ does not enjoy it, since $|z|$ is not rescaled by ϵ, which is one more reason isolating hw is the more robust recommendation.)

Does a learned combination add anything over isolating hw? A fixed formula assumes some particular trade-off between hw and $|z|$; we check whether a *learned* combination, calibrated on one corruption and evaluated on the other five unseen ones, can do better than simply isolating hw. We fit a small logistic regression on $[\,|z|,\ \mathrm{hw},\ |z| \cdot \mathrm{hw}\,]$ per corruption pair (Table 5). The diagonal (mean 0.859) and off-diagonal (mean 0.858) are essentially identical, column by column: **on Waterbirds, the learned combination is not adding anything beyond isolating hw**, it simply relearns to put almost all its weight on hw and ignore margin, on every corruption we tried, which is exactly consistent with hw alone already being the best fixed score on this dataset (previous paragraph). We show below, however, that this simplification is itself specific to Waterbirds: on a second dataset (CelebA), a fixed combination sometimes beats isolating hw, which is exactly the situation an adaptive or learned combination is suited for. We therefore do not generalize "isolate hw, skip the combination" beyond the dataset where we tested it.

Generalization across encoders, datasets, and modalities. We repeat the fixed-geometry six-corruption methodology on three more (encoder, dataset) combinations: CLIP ViT-B/32 on Waterbirds, ResNet-50 on CelebA, and DistilBERT on CivilComments (text, five text-native corruptions). Table 6 summarizes; full per-corruption breakdowns, mechanisms, and the ID-only-calibration check (which helps on Waterbirds but actively hurts on CelebA, not a solved problem) are in App. D. **The headline finding: isolating hw replicates across encoders on the same dataset (Waterbirds: ResNet-50 and CLIP agree) but not across datasets (CelebA reverses the ranking; text is genuinely ambiguous).** "Isolate hw" is a real but dataset-dependent regularity, not a universal prescription, and no fixed or learned combination we tried resolves this cleanly across all three datasets.

Severity sweep. Using the same fixed-geometry protocol as the main OOD results (Sec. 5.1), we sweep severity on `gaussian_noise` and `fog` at severities 1–5 (n=250, ResNet-50, Waterbirds): hw's AUROC rises *monotonically* with severity for both corruptions (`gaussian_noise`: $0.781 \rightarrow$

Table 6: Mean AUROC across six corruptions (five for text), by (encoder, dataset). Full breakdowns in App. D.

	Margin	hw	$m(\boldsymbol{x})$	Winner
ResNet-50, Waterbirds	0.552	**0.859**	0.818	hw
CLIP, Waterbirds	0.482	**0.889**	0.698	hw
ResNet-50, CelebA	0.537	0.865	**0.883**	$m(\boldsymbol{x})$
DistilBERT, CivilComments	0.563	0.649	0.655	tied

$0.867 \to 0.926 \to 0.964 \to 0.989$; `fog`: $0.617 \to 0.648 \to 0.695 \to 0.736 \to 0.841$), with margin also rising but staying well behind throughout. hw's advantage over margin holds across the full severity range for both corruption types, with no crossover or reversal at high severity.

Scope. These real-encoder OOD results now cover three datasets (Waterbirds, CelebA, CivilComments) and two encoders (ResNet-50, CLIP ViT-B/32) with the fixed-geometry methodology. The worst-group-accuracy interchangeability story (Sec. 5) is unaffected by any of this and separately replicates on CLIP. Code for the real-encoder experiments and the underlying scores/data are in the supplementary material.

6 Discussion

For spurious-correlation robustness, the Rashomon-multiplicity signal reduces to a margin proxy, which is why it is interchangeable with loss, disagreement, and confidence, and, on Waterbirds, with uniform retraining. The leverage term hw, unique to multiplicity, is a parameter-uncertainty term and is group-uninformative for minority localization, but Sec. 5.1 shows it is not group-*blind*: it carries genuine covariate-shift signal that margin lacks by construction on an idealized orthogonal shift, and this prediction holds up on real ResNet-50 features once the Rashomon-set geometry is computed correctly. It does so unevenly across encoders and datasets (Table 6): isolating hw replicates across two encoders on Waterbirds but reverses on CelebA and is ambiguous on CivilComments text. **"Isolate hw" is a real but dataset-dependent regularity, not a universal prescription**, exactly the situation a calibrated combination seems suited for, though the ID-only calibration we tried helps where it is least needed (Waterbirds) and hurts where it is most needed (CelebA), so we do not consider that specific recipe solved (App. D). Nor does isolated hw clearly beat standard, much simpler OOD baselines on any dataset: Mahalanobis distance and k-NN feature distance are consistently competitive, and need no fitted classifier or OOD calibration data at all. The decomposition's value, on the evidence we have, is explanatory, it tells you *why* a fitted classifier's own leverage term carries shift information, and gives a checkable mechanism (whether margin itself is informative about a given shift) for predicting which regime you are in, rather than a demonstration that it gives new state-of-the-art OOD detection. The diagnostic picture: for spurious-correlation robustness, use a margin proxy and do not pay for multiplicity estimation; for covariate-shift detection, check whether margin carries real signal about the shift of interest before choosing between isolating hw and combining it with margin, and consider Mahalanobis or k-NN as at least as good a starting point either way.

Is the cost worth it? Full costs are in Table 7 (App. F). Briefly, the naive asymptotic reading is misleading: loss, disagreement, and confidence are essentially free (0.04s), but **SELF and JTT, not hw, are the most expensive methods in absolute terms** (SELF 198s, JTT 105s, dominated by refit iterations), while hw's total cost (65s) is still faster than either refit. *Margin is free, and it is SELF/JTT's refit, not hw's Hessian, that is the real price of the alternatives here.* Future work and reproducibility details are in App. G.

Statement on the Use of AI

We used a large language model to aid in polishing the writing of this paper. The research questions, experimental design, and interpretation of results are the authors' own.

REFERENCES

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning (ICML)*, 2020.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning (ICML)*, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

A FULL PROOF OF PROPOSITION 1

We work with the population (infinite-data) cross-entropy minimizer $\boldsymbol{w}_0 = \arg\min_{\boldsymbol{w}} \mathbb{E}[\log(1 + \exp(-y\,\boldsymbol{w}^\top \tilde{\phi}(\boldsymbol{x})))]$; empirical finite-$n$ deviations are handled separately in Lemma 4. Write $\boldsymbol{w}_0 = (w_{0,c}, w_{0,s}, \boldsymbol{w}_{0,n})$ for the core, spurious, and nuisance blocks (bias omitted; the problem is symmetric so $b_0 = 0$). **We derive $w_{0,c}$ and $w_{0,s}$ below by treating each block's first-order condition as if it were the only predictor, this is a simplifying reduction, not a generic mathematical consequence of x_c, x_s, and the nuisance coordinates being independent given y: for a nonlinear loss like cross-entropy, independent regressors do not generally decouple the joint optimum into independently-solvable per-block problems the way they would for squared loss.** We verify the reduction's accuracy directly against the true joint optimum below (Lemma 3) rather than relying on the independence argument alone.

Core weight (approximate reduction, verified below). x_c is generated identically for majority and minority, so $X_c = yx_c = \mu_c + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$, regardless of α. The scalar program $\min_w \mathbb{E}_{Z\sim\mathcal{N}(\mu_c,\sigma^2)}[\log(1 + e^{-wZ})]$ has the classical closed-form stationary point

$$w_{0,c} = \frac{2\mu_c}{\sigma^2}, \tag{2}$$

proved directly by noting the first-order condition $\mathbb{E}_Z[Z\,\sigma(-wZ)] = 0$ is solved exactly by the LLR slope $w = 2\mu_c/\sigma^2$ when $Z = \mu_c + \sigma\varepsilon$, since then $\sigma(-wZ) = p(Z \mid y{=}-1)/[p(Z \mid y{=}1) + p(Z \mid y{=}-1)]$ exactly, this is a complete analytic argument, not a numerical one, and we additionally confirmed it numerically to 4 significant figures for several (μ_c, σ) purely as a sanity check on the derivation. Because X_c's law never involves α, $w_{0,c}$ is exactly constant in α, confirming the "group-independent core feature" claim used in the main text.

Spurious weight. Using the label-product trick, for either value of y the variable $X_s = y\,x_s$ has the *same* mixture law

$$X_s \;\sim\; (1-\alpha)\,\mathcal{N}(\mu_s, \sigma^2) \;+\; \alpha\,\mathcal{N}(-\mu_s, \sigma^2), \tag{3}$$

because a majority sample contributes $X_s = \mu_s + \sigma\varepsilon$ and a minority sample contributes $X_s = -\mu_s + \sigma\varepsilon'$ with $\varepsilon, \varepsilon' \sim \mathcal{N}(0,1)$, and this holds for both $y = +1$ and $y = -1$ by the symmetry $s = \pm y$. The population objective is then the strictly convex scalar program $g(w) = \mathbb{E}_{X_s}[\log(1 + e^{-wX_s})]$, whose unique root of $g'(w) = 0$ is

$$(1-\alpha)\,\mathbb{E}_{Z\sim\mathcal{N}(\mu_s,\sigma^2)}[Z\,\sigma(-wZ)] \;=\; \alpha\,\mathbb{E}_{Z\sim\mathcal{N}(-\mu_s,\sigma^2)}[Z\,\sigma(-wZ)]\cdot(-1). \tag{4}$$

Lemma 1 (Monotonicity and two regimes)**.** *The unique solution $w_{0,s}(\alpha)$ of Eq. (4) is strictly decreasing in α on $(0, \frac{1}{2})$, with $w_{0,s}(\frac{1}{2}) = 0$, and*

$$w_{0,s}(\alpha) \;\xrightarrow[\sigma/\mu_s\to 0]{}\; \frac{1}{\mu_s}\log\frac{1-\alpha}{\alpha} \qquad \textit{(noiseless / well-separated limit, exact for any fixed } \alpha\textit{)}, \tag{5}$$

$$w_{0,s}(\alpha) \;\xrightarrow[\alpha\to 0]{}\; \frac{2\mu_s}{\sigma^2} \qquad \textit{(saturation ceiling, for any fixed } \mu_s, \sigma > 0\textit{)}. \tag{6}$$

Proof. Write $F(w,\alpha) = (1-\alpha)A(w) + \alpha B(w)$ for the left-hand-side-minus-right-hand-side of Eq. (4), where $A(w) = \mathbb{E}_{Z\sim\mathcal{N}(\mu_s,\sigma^2)}[Z\sigma(-wZ)]$ and $B(w) = \mathbb{E}_{Z\sim\mathcal{N}(-\mu_s,\sigma^2)}[Z\sigma(-wZ)]$, so $w_{0,s}(\alpha)$ solves $F(w,\alpha) = 0$. Convexity of g gives existence and uniqueness of the root. We now establish monotonicity *exactly*, with no numerical verification. First, $\partial_w F = (1-\alpha)A'(w) + \alpha B'(w)$ with $A'(w) = -\mathbb{E}_Z[Z^2\sigma(-wZ)(1-\sigma(-wZ))] < 0$ and likewise $B'(w) < 0$ (the integrand is a.e. strictly positive for a continuous, non-degenerate Z), so $\partial_w F < 0$ strictly for every $\alpha \in (0,1)$. Second, substituting $Z = -Z'$ in $B(w)$ (so $Z' \sim \mathcal{N}(\mu_s, \sigma^2)$) gives $B(w) = -\mathbb{E}_{Z'\sim\mathcal{N}(\mu_s,\sigma^2)}[Z'\sigma(wZ')]$, hence

$$A(w) - B(w) = \mathbb{E}_{Z\sim\mathcal{N}(\mu_s,\sigma^2)}\big[Z\big(\sigma(-wZ) + \sigma(wZ)\big)\big] = \mathbb{E}_{Z\sim\mathcal{N}(\mu_s,\sigma^2)}[Z] = \mu_s, \tag{7}$$

using the identity $\sigma(x) + \sigma(-x) = 1$ for **every** w, not an approximation or a numerically-checked regime, but an exact closed form (we also confirmed it numerically to 6 decimal places across several (w, μ_s, σ) as a sanity check on the derivation). Consequently $\partial_\alpha F = B(w) - A(w) = -\mu_s < 0$ for every w, exactly. By the implicit function theorem,

$$\frac{dw_{0,s}}{d\alpha} = -\frac{\partial_\alpha F}{\partial_w F} = -\frac{-\mu_s}{\partial_w F} = \frac{\mu_s}{\partial_w F} < 0, \tag{8}$$

since $\mu_s > 0$ and $\partial_w F < 0$; this holds at every $\alpha \in (0,1)$, so $w_{0,s}(\alpha)$ is strictly decreasing on the entire domain, not merely observed to be so on a numerical grid. For Eq. (5): as $\sigma \to 0$ each Gaussian in Eq. (3) collapses to a point mass at $\pm\mu_s$, and Eq. (4) reduces to $(1-\alpha)\sigma(-w\mu_s) = \alpha\sigma(w\mu_s)$, i.e. (writing $p = \sigma(w\mu_s)$) $(1-\alpha)(1-p) = \alpha p \Rightarrow p = 1-\alpha$, giving $w\mu_s = \log\frac{1-\alpha}{\alpha}$ exactly. For Eq. (6): as $\alpha \to 0$ the mixture Eq. (3) converges (in the sense relevant to g, which is bounded and continuous) to the single component $\mathcal{N}(\mu_s, \sigma^2)$, and $g \to g_0(w) = \mathbb{E}_{Z\sim\mathcal{N}(\mu_s,\sigma^2)}[\log(1+e^{-wZ})]$, whose unique minimizer is exactly $w = 2\mu_s/\sigma^2$ by the identical analytic LLR argument as Eq. (2) above, again a complete analytic argument, confirmed numerically to 4 significant figures for $\mu_s/\sigma \in \{1,2,3,5\}$ only as a sanity check, not as the basis for the claim. □

In words: for a fixed noise level σ, $w_{0,s}(\alpha)$ grows as α shrinks, tracking $\Theta\big(\log\frac{1-\alpha}{\alpha}\big)$ over the intermediate range $\alpha \gtrsim \Phi(-\mu_s/\sigma)$ (the natural noise-floor probability of the spurious feature), and then saturates at the finite ceiling $2\mu_s/\sigma^2$ once α drops below that floor. Either way $w_{0,s}(\alpha)$ is increasing in $1/\alpha$ and unbounded only in the idealized noiseless limit; this refines (and is consistent with) the informal $\Theta(\log((1-\alpha)/\alpha))$ statement.

Lemma 2 (Margin gap, proved exactly)**.** *Write $Z_\bullet = y \cdot z$ for either group ($|Z_\bullet| = |z|$ since $y \in \{\pm 1\}$). Using the label-product trick as in Eq. (3), $Z_{\mathrm{maj}} \sim \mathcal{N}(\mu_{\mathrm{maj}}, \tau^2)$ and $Z_{\mathrm{min}} \sim \mathcal{N}(\mu_{\mathrm{min}}, \tau^2)$ with $\mu_{\mathrm{maj}} = w_{0,c}\mu_c + w_{0,s}\mu_s$, $\mu_{\mathrm{min}} = w_{0,c}\mu_c - w_{0,s}\mu_s$, and* the same *variance $\tau^2 = \sigma^2(w_{0,c}^2 + w_{0,s}^2)$ for both groups. Then $\mathbb{E}[|z| \mid \mathrm{maj}] > \mathbb{E}[|z| \mid \mathrm{min}]$ exactly, for every $\alpha \in (0, \frac{1}{2})$, with equality only if $w_{0,s} = 0$.*

Proof. For $Z \sim \mathcal{N}(\mu, \tau^2)$, the folded-normal identity $\mathbb{E}|Z| = \tau\sqrt{2/\pi}\,e^{-\mu^2/2\tau^2} + \mu\big(1 - 2\Phi(-\mu/\tau)\big)$ is a strictly increasing, even function of $|\mu|$ for fixed τ (standard; differentiating gives $d\mathbb{E}|Z|/d|\mu| =$

$1 - 2\Phi(-|\mu|/\tau) > 0$ for $|\mu| > 0$). It therefore suffices to show $|\mu_{\mathrm{maj}}| > |\mu_{\mathrm{min}}|$. Since $w_{0,c}\mu_c > 0$ (Eq. (2), both factors positive by assumption) and $w_{0,s}\mu_s > 0$ (Lemma 1, $w_{0,s} > 0$ for $\alpha < \frac{1}{2}$, $\mu_s > 0$ by assumption), writing $a = w_{0,c}\mu_c > 0$ and $b = w_{0,s}\mu_s > 0$ gives $\mu_{\mathrm{maj}} = a + b$ and $\mu_{\mathrm{min}} = a - b$, and $a + b > |a - b|$ for any $a, b > 0$ (equivalent to $\min(a, b) > 0$, which holds), with equality only as $b \to 0$, i.e. $w_{0,s} \to 0$. □

This is an *exact* proof of $\Delta_{-|z|} > 0$, not an approximation or an intuitive gap-widening argument: it holds for every $\alpha \in (0, \frac{1}{2})$, not just asymptotically. For the *rate*, that this gap grows as α shrinks, we do not give a fully general analytic monotonicity proof (differentiating $\mathbb{E}|Z_\bullet|$ through both $\mu_\bullet(w_{0,s})$ and $\tau(w_{0,s})$ simultaneously is more involved), but confirm it numerically to high precision across the full range of α tested, including an additional finding: like $w_{0,s}(\alpha)$ itself, the margin gap $\Delta_{-|z|}(\alpha)$ saturates at a finite ceiling ($2w_{0,c}\mu_c$, attained once $w_{0,s}\mu_s > w_{0,c}\mu_c$) rather than growing without bound as $\alpha \to 0$, the same qualitative saturation phenomenon as Lemma 1, for the same underlying reason.

Lemma 3 (The block-separated reduction matches the true joint optimum)**.** *The values $w_{0,c} = 2\mu_c/\sigma^2$ and $w_{0,s}(\alpha)$ derived above by treating each block's first-order condition in isolation are not, in general, the exact joint optimum of the full population objective over (w_c, w_s) jointly (Lemma 2's independence argument does not by itself imply this, for a nonlinear loss). We check the reduction's accuracy directly: writing $X = y \cdot z = w_c\mu_c + w_s\mu_s$ (majority, probability $1 - \alpha$) or $w_c\mu_c - w_s\mu_s$ (minority, probability α), each with variance $\tau^2 = \sigma^2(w_c^2 + w_s^2)$ (nuisance weights are zero at any optimum, since they add to τ^2 without contributing to either mean), we directly minimize the exact population objective $L(w_c, w_s) = (1 - \alpha)\mathbb{E}_{Z\sim\mathcal{N}(w_c\mu_c + w_s\mu_s, \tau^2)}[\log(1 + e^{-Z})] + \alpha\,\mathbb{E}_{Z\sim\mathcal{N}(w_c\mu_c - w_s\mu_s, \tau^2)}[\log(1 + e^{-Z})]$ over (w_c, w_s) jointly, by numerical quadrature (population-exact, no finite-sample noise), for $\mu_c = \mu_s = 1$ (the configuration used in Table 1 and Fig. 1) across the full range of α tested. The true joint optimum's w_c stays within 0.6% of the reduction's $2\mu_c/\sigma^2$ prediction at every $\alpha \in \{0.3, \ldots, 0.0002\}$, peaking near $\alpha \approx 0.02$ and shrinking again as $\alpha \to 0$; w_s agrees to comparable precision.* ***The deviation is not always this small****: for $\mu_s > \mu_c$ (spurious signal stronger than core), the deviation grows with $1/\alpha$, reaching 3.8% at μ_c=1, μ_s=2, α=0.02 and 7.7% at μ_c=0.5, μ_s=1.5, α=0.02; for $\mu_c \geq \mu_s$ the deviation stays under 0.2% throughout. We report the reduction as an accurate approximation* verified against the true joint optimum in the regime this paper's experiments use *($\mu_c = \mu_s$), not as an exact consequence of independence, and flag that the approximation should be re-checked before being applied to a regime where the spurious signal substantially dominates the core signal.*

Together with Lemma 1 ($w_{0,s}$ strictly increasing as $\alpha \to 0$) and Lemma 4 ($\Delta_{\mathrm{hw}} = 0$ exactly), and taking Lemma 3's verified approximation of $w_{0,c}, w_{0,s}$ as given, this establishes Proposition 1: $\Delta_{m(\boldsymbol{x})} = \Delta_{-|z|} + \Delta_{\mathrm{hw}} > 0$ for every $\alpha \in (0, \frac{1}{2})$, exactly at the level of Δ_{hw} and to within the small, quantified, and directionally understood approximation error of Lemma 3 at the level of $\Delta_{-|z|}$.

Lemma 4 (Leverage term is population-exact group-uninformative)**.** *At the population level, $\Delta_{\mathrm{hw}} = 0$ exactly (not merely to leading order): for x_c and the isotropic nuisance coordinates this is immediate since their laws do not depend on group membership; for x_s, conditioning on* majority *gives $x_s = y\mu_s + \sigma\varepsilon$ and conditioning on* minority *gives $x_s = -y\mu_s + \sigma\varepsilon$, and since $y \in \{\pm 1\}$ is independent of ε and symmetric, x_s^2 has the identical distribution $(\mu_s + \sigma Z)^2$, $Z \sim \mathcal{N}(0, 1)$, under both conditionings. Hence every quadratic form $\tilde{\phi}(\boldsymbol{x})^\top \boldsymbol{H}^{-1}\tilde{\phi}(\boldsymbol{x})$ built from $(x_c, x_s, x_{n,1}, \ldots, x_{n,k})$ has equal conditional expectation given majority or minority, so $\mathbb{E}[\mathrm{hw} \mid \textit{min}] = \mathbb{E}[\mathrm{hw} \mid \textit{maj}]$ at the population Hessian $\boldsymbol{H}$. The only source of nonzero Δ_{hw} in a finite sample of size n is estimation noise in the empirical Hessian $\hat{\boldsymbol{H}}_n$ and in the empirical group means: by the CLT, $\hat{\boldsymbol{H}}_n = \boldsymbol{H} + O_p(n^{-1/2})$, and the difference of two independent finite-sample group averages of an identically-distributed quadratic form has variance $O(1/n)$ (assuming both group sizes scale with n), giving $|\Delta_{\mathrm{hw}}| = O_p(n^{-1/2})$ and $\mathrm{Var}(\Delta_{\mathrm{hw}}) = O(\sigma^4/n)$.*

This matches Table 1, where $\Delta_{\mathrm{hw}} \leq 0.002$ in magnitude across all α at n in the thousands, consistent with the $O(n^{-1/2})$ finite-sample rate rather than a population-level effect.

Numerical validation. We confirmed Eqs. (2)– (6) by (i) fitting population-scale ($n = 6 \times 10^5$) logistic regression on simulated data across $\mu_s/\sigma \in \{1, 2, 3, 5\}$ and $\alpha \in \{10^{-3}, \ldots, 0.3\}$, and (ii) directly solving Eq. (4) by numerical quadrature; the fitted $w_{0,s}$ tracks the log-regime prediction

closely while $\mu_s/\sigma \gg 1$ and $\alpha \gg \Phi(-\mu_s/\sigma)$, and converges to $2\mu_s/\sigma^2$ as $\alpha \to 0$ in every configuration tested, matching the ceiling to 4 significant figures.

B SPEARMAN VALIDATION PROTOCOL

For each $\alpha \in \{0.02, 0.05, 0.1, 0.3\}$ we simulate n=3000 points from the Sec. 4 Gaussian model (μ_c=μ_s=1, σ=1, k=8 isotropic nuisance dimensions), fit $\boldsymbol{w}_0$ by unregularized population logistic regression, and compare the closed-form $m(\boldsymbol{x})$, hw (Eq. (1)) against two independent ensembles of size 300: (i) directions sampled uniformly on the Hessian ellipsoid boundary at radius $\sqrt{2\epsilon L(\boldsymbol{w}_0)}$ (shares the Laplace geometry, checks only that our extreme-logit formula is computed correctly against a Monte Carlo estimate of the same ellipsoid); (ii) logistic regression refit from scratch on 300 bootstrap resamples of the data (does not assume the Laplace approximation at all, and is the more stringent check). Code in the supplementary material.

C EXACT AND INDEPENDENT VALIDATION BEYOND CORRELATION

The Spearman validation above (App. B) checks rank agreement between the closed-form score and two ensemble estimates, but correlation does not by itself establish that the Laplace-approximated Rashomon interval is numerically close to the *true* constrained optimum, nor that the Hessian formula itself is implemented correctly independent of any approximation. We address both directly.

Exact low-dimensional Rashomon-set validation. For $d \in \{2, 5, 10\}$ we fit $\boldsymbol{w}_0$ on synthetic logistic-regression data (n=500), then directly solve the constrained optimization $\max / \min_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{x}$ subject to $L(\boldsymbol{w}) \le (1+\epsilon)L(\boldsymbol{w}_0)$ via sequential quadratic programming (multiple random restarts), for 15 random test points per d. This is the actual Rashomon-set boundary, not an approximation or an ensemble estimate of one. Comparing the true half-width $(\max - \min)/2$ to the closed-form $\mathrm{hw}(\boldsymbol{x})$: mean relative error is 1.2% (d=2), 1.3% (d=5), 1.4% (d=10), with all 45 individual test points agreeing to within 2.2%. This is, to our knowledge, the most direct available check: the Laplace approximation is not merely correlated with the truth, it is quantitatively close to it, on problems small enough to solve exactly.

Independent implementation of the Hessian. Separately, we verify the closed-form Hessian formula $\boldsymbol{H} = \frac{1}{n}\tilde{\Phi}^\top \mathrm{diag}(p(1-p))\tilde{\Phi}$ itself is correctly implemented, independent of any Rashomon-approximation question, by comparing it against a second-order central finite-difference Hessian of the same loss function computed with no shared code (for $d \in \{2, 5, 10, 20\}$). Relative error in the resulting $\tilde{\phi}(\boldsymbol{x})^\top \boldsymbol{H}^{-1}\tilde{\phi}(\boldsymbol{x})$ quadratic form (the quantity that enters hw) is 3×10^{-5} to 10^{-4} across all four dimensions, consistent with finite-difference truncation error rather than a formula discrepancy.

Residual multiplicity: does anything beyond margin's linear relationship with $m(\boldsymbol{x})$ help? App. B's R^2 result (main text, Sec. 4) shows margin linearly explains 94–97% of $m(\boldsymbol{x})$'s variance; here we ask whether the residual after removing a flexible (cubic) fit of $m(\boldsymbol{x})$ on margin adds anything for minority prediction beyond margin alone. It does, by essentially the same amount as the hw-incremental-AUC check already reported (unsurprising, since the residual of $m(\boldsymbol{x})$ after removing its margin-dependence is, by construction, close to hw itself): incremental AUC of $+0.011$ to $+0.024$ across $\alpha \in \{0.02, 0.05, 0.1, 0.3\}$, consistent with, not an independent finding beyond, the main text's incremental-AUC result.

Ranking agreement and top-k overlap between signals. Beyond downstream worst-group accuracy (Table 2), we check whether the label-free signals actually agree on *which* examples they flag, constructing a SELF-style disagreement signal on the Gaussian model (heavy- vs. light-regularization prediction disagreement) alongside $m(\boldsymbol{x})$, margin, and an AFR-style loss signal. $m(\boldsymbol{x})$ and margin agree closely, as expected (ρ=-0.98; top-10% Jaccard overlap 0.61), and $m(\boldsymbol{x})$ and loss agree closely at coarser k (ρ=0.94; top-20% overlap 0.60) but share almost none of their most extreme-ranked examples (top-1% overlap 0.00), loss and margin agree in the bulk of the distribution but not in the tails, where a logistic loss's exponential sensitivity diverges from a linear margin ranking. Most strikingly, $m(\boldsymbol{x})$ and the SELF-style disagreement signal show *almost no*

rank agreement at any scale (ρ=0.04; top-k Jaccard overlap 0.02–0.11 across $k \in \{1, 5, 10, 20\}\%$). **Different label-free signals can flag almost entirely different individual examples while still producing statistically indistinguishable downstream worst-group accuracy** (Table 2), interchangeability at the level of the reweighting outcome does not imply interchangeability at the level of which examples are reweighted, a distinction worth keeping separate when interpreting Table 2's result.

Resolving the apparent paradox: why does outcome-equivalence survive such low ranking agreement? On Waterbirds this has a direct, complete, and unglamorous explanation rather than a deep property of aggregate training dynamics: Table 2 reports that validation selects $\lambda = 0$ for AFR, SELF, *and* RLM. At $\lambda = 0$, none of the three signals reweight anything, every method trains on the uniform objective, regardless of how its underlying score would have ranked the training examples. The individual-level ranking disagreement we measure above is real, but on Waterbirds it is provably irrelevant to the reported worst-group accuracy, because none of the rankings are ever exercised at the validated operating point: outcome-equivalence follows mechanically from every method converging to the same $\lambda = 0$ answer, not from different rankings coincidentally inducing the same effective training distribution. On CelebA and CivilComments, where the reported worst-group accuracies differ very slightly between AFR, SELF, and RLM (Table 2), validation may be selecting a small nonzero λ for at least some methods, in which case rankings *are* exercised to some degree and this deflationary explanation would not fully apply; we have not measured ranking agreement on those two datasets or confirmed their selected λ values are exactly zero, so we do not extend the Waterbirds explanation to them and flag this as an open question rather than assume the same mechanism.

D Additional real-encoder generalization details

Full detail behind Table 6 in the main text.

Does the OOD story generalize to a second encoder? We repeated the six-corruption sweep on CLIP ViT-B/32 features on Waterbirds, with the fixed-geometry methodology. The ranking replicates cleanly: isolated hw beats margin and beats $m(\boldsymbol{x})$ on all six corruptions (mean AUROC 0.889 vs. 0.698 for $m(\boldsymbol{x})$ vs. 0.482 for margin), the same qualitative pattern as ResNet-50 (mean 0.859 vs. 0.818 vs. 0.552). Isolating hw is not a ResNet-50-specific artifact, it holds across two architecturally different encoders on the same dataset.

Does the OOD story generalize to a second dataset? Here the picture changes. Repeating the same fixed-geometry six-corruption sweep on CelebA (ResNet-50, disjoint 300-image ID slices per corruption): isolated hw still beats margin on all six corruptions (mean 0.865 vs. 0.537), but $m(\boldsymbol{x})$ beats isolated hw on four of six corruptions (`fog`: 0.923 vs. 0.908; `brightness`: 0.745 vs. 0.678; `contrast`: 0.744 vs. 0.708; `pixelate`: 0.987 vs. 0.984), with hw only ahead on `gaussian_noise` and `defocus_blur`, and by a small margin in both. Mean AUROC is 0.883 ($m(\boldsymbol{x})$) vs. 0.865 (hw), the opposite ranking from Waterbirds. This tracks a plausible mechanism: margin itself is more informative about these shifts on CelebA (AUROC 0.54–0.67, clearly above chance on four of six corruptions) than on Waterbirds (0.44–0.65, closer to chance or even below it on several), so combining margin with hw helps precisely where margin is not simply noise. The "isolate hw" recommendation is dataset-dependent, not just shift-dependent, and replicates across encoders only within the dataset where we first observed it. We report this plainly: the general prescription is not "always isolate hw," but "isolate hw when margin carries no real signal about the shift, combine when it does", which is not knowable in advance without the kind of check this section performs, reinforcing the earlier finding that a *calibrated* combination (learned or otherwise informed by which regime you are in) is more robust than either fixed choice alone.

Does the OOD story extend to text? We repeat the same fixed-geometry methodology on CivilComments using frozen DistilBERT CLS features and five text-native corruptions (character noise, word deletion, word-order shuffling, truncation, and case randomization; n=300 per corruption, disjoint slices). Mean AUROC: margin 0.563, hw 0.649, $m(\boldsymbol{x})$ 0.655, essentially tied, and both well below the vision numbers on either dataset. Per-corruption, hw and $m(\boldsymbol{x})$ trade wins roughly

evenly (each ahead on two of five, one exact tie), a *third* distinct qualitative pattern alongside Waterbirds (hw wins) and CelebA ($m(\boldsymbol{x})$ wins more often), text does not cleanly resemble either vision dataset. One result is a useful sanity check on the whole pipeline rather than a substantive finding: `case_randomize` scores exactly at chance (0.499) for every score tested (margin, hw, $m(\boldsymbol{x})$, Mahalanobis, k-NN) because `distilbert-base-uncased` lower-cases its input, making this corruption genuinely invisible to the model, exactly the null result a correctly-implemented pipeline should produce for a corruption with no real effect, which is reassuring given how much of this paper's empirical content depends on this exact scoring machinery being implemented correctly.

Can the combination be calibrated without ever seeing OOD data? The learned combination in Table 5 needs labeled OOD examples from at least one shift; here we test the more practical question of whether *ID-only* calibration helps. We compute z-scores of hw and margin using only ID validation statistics (mean and standard deviation, no OOD data of any kind), then combine with equal weight, and compare against the fixed formula and an oracle upper bound (logistic regression fit on actual labeled OOD data), on a separately re-run instance of the six-corruption sweep for each dataset (numbers below therefore differ from Table 6 by up to 0.001, consistent with the Monte Carlo noise expected between independent bootstrap runs, not a substantive discrepancy). On Waterbirds: mean AUROC hw alone 0.858, fixed formula 0.816, ID-only calibration 0.824, oracle 0.859. ID-only calibration recovers about half the gap between the fixed formula and the oracle, a real, if modest, improvement using zero OOD supervision. On CelebA, where the fixed formula already has real room to help ($m(\boldsymbol{x}) = 0.884$ beats hw's 0.866), ID-only calibration makes things *worse*, not better: mean AUROC drops to 0.854, below both the fixed formula and hw alone, worse on five of the six corruptions and better on only one (`brightness`). The likely mechanism: on CelebA hw's raw scale is roughly $2.4\times$ margin's (std 8.29 vs. 3.42 on the ID validation set), so the *unstandardized* fixed formula implicitly weights hw more heavily than margin, apparently close to what the oracle also prefers (oracle mean 0.888, close to the fixed formula's 0.884), while z-scoring to enforce equal weight removes exactly this beneficial imbalance. This specific ID-only calibration heuristic is not a reliable fix: it helps on the dataset where isolating hw was already close to optimal and hurts on the dataset where a combination was most needed, which is the opposite of what a practically useful calibration method should do. We report both results rather than only the more flattering one; the next paragraph reports a more careful attempt at exactly this kind of calibration, and why it also did not solve the problem.

A more careful ID-only calibration attempt, and why it also failed. We tried the natural next step: instead of z-scoring margin and hw independently (which ignores their correlation), fit the full 2×2 covariance of $[\mathrm{margin}, \mathrm{hw}]$ on ID validation data and score by Mahalanobis distance in that two-dimensional score space, still using zero OOD data or labels. This did *not* help, it did *worse* than the naive independent z-scoring on both datasets tested (Waterbirds: mean AUROC 0.782, vs. 0.824 for naive z-scoring, 0.816 fixed formula, 0.858 for hw alone; CelebA: 0.774, vs. 0.854 naive, 0.884 fixed, 0.866 for hw alone). The likely reason: Mahalanobis distance in score-space is direction-agnostic, it flags any deviation from the typical ID $(\mathrm{margin}, \mathrm{hw})$ combination as anomalous, regardless of which direction that deviation points in, whereas the oracle's logistic regression (Table 5) works because it discriminatively learns *which* direction actually corresponds to OOD versus benign in-distribution variation. Characterizing the shape of the ID distribution, however carefully, does not substitute for knowing which direction of deviation from it is the one that matters, and ID data alone cannot supply that. **Across two independent ID-only calibration attempts, neither reliably improved on simply picking the better of the two fixed choices** (hw **alone or the fixed formula) for the dataset at hand**; we no longer expect a better ID-only heuristic to resolve this and think the more promising direction is the checkable mechanism in the main text (does margin carry real signal about this shift?) rather than further calibration attempts.

E Formal significance testing under a supplementary bootstrap protocol

Table 2's five seeds come from the original protocol (full data on Waterbirds, subsampled on CelebA/CivilComments), and we do not have access to its underlying per-seed predictions to run a paired test directly against it. Instead we ran a supplementary check on a protocol we do fully control: 15 seeds via bootstrap resampling (with replacement) of the Waterbirds training set, refitting

every method's head from scratch per seed and evaluating on the fixed test set. **This changes absolute worst-group accuracy substantially**: ERM $65.2\pm1.5\%$, DFR $73.4\pm1.1\%$, AFR $66.4\pm1.2\%$, SELF $65.0\pm1.4\%$, JTT $65.0\pm1.5\%$, confidence $66.0\pm1.2\%$, RLM $65.3\pm1.5\%$, all lower than Table 2's numbers, most sharply for DFR (from 89.2%), because resampling with replacement duplicates some training examples and omits others, most consequentially for DFR's group-balanced reweighting when a bootstrap draw under-represents a minority group. These absolute numbers are therefore *not* directly comparable to Table 2 and should not be read as a replication of it; the value of this protocol is that it gives us genuine per-seed variation to test formally.

With that scope caveat, we ran a paired bootstrap test (10,000 resamples of the 15 seed-level differences) between RLM and each other signal, including a ±1 percentage point practical-equivalence band:

Comparison	Diff. (pp)	95% CI (pp)	p	Equivalent ($\leq$ 1pp)?
RLM vs. AFR	-1.09	$(-1.87, -0.31)$	0.006	No
RLM vs. SELF	$+0.32$	$(-0.27, +0.89)$	0.278	Yes
RLM vs. JTT	$+0.27$	$(-0.28, +0.79)$	0.312	Yes
RLM vs. confidence	-0.72	$(-1.42, -0.06)$	0.031	No
RLM vs. uniform (ERM)	$+0.17$	$(-0.01, +0.34)$	0.065	Yes

RLM is practically equivalent to SELF, JTT, and uniform retraining under this test, but the null of exact equivalence is rejected for AFR and confidence, with RLM measurably (if modestly, under 1.1 percentage points) worse than both. **We take this as a genuine, if small, qualification of the paper's interchangeability claim**: "interchangeable" is accurate as a description of the general pattern and is not overturned by this result, but a formal test, which a visual mean$\pm$std comparison cannot substitute for, finds real, non-zero differences between specific pairs that a coarser comparison would not have surfaced. We see no evidence this reflects anything about $m(\boldsymbol{x})$ specifically rather than sampling variation interacting with this particular bootstrap protocol, and do not have the original protocol's per-seed data to check whether the same pattern holds there.

F Per-method cost details

Table 7 gives both the asymptotic complexity and actual wall-clock measurements (Waterbirds, ResNet-50, n=4795, d=2048, single CPU run, scoring cost for all 5794 test examples) referenced from the cost discussion in Sec. 6. hw's one-time Hessian cost is under 3s; its $O(d^2)$ per-example scoring is real (63s for 5794 examples) but the 65s total is still faster than both refit-based methods. Mahalanobis costs about the same as hw (70s); k-NN is far cheaper (3s) thanks to optimized nearest-neighbor search.

Table 7: Per-method cost: asymptotic (left) and measured wall-clock (right; Waterbirds, ResNet-50, n=4795, d=2048, single CPU run, scoring cost for all 5794 test examples), sorted by total time.

Method	One-time	Per-ex.	Memory	One-time (s)	Scoring (s)	Total (s)
Loss / Confidence	N/A	$O(d)$	$O(1)$	N/A	0.04	**0.04**
k-NN	N/A	N/A	N/A	0.04	3.01	3.05
hw / $m(\boldsymbol{x})$ (RLM)	$O(nd^2{+}d^3)$	$O(d^2)$	$O(d^2)$	2.59	62.64	65.23
Mahalanobis	N/A	N/A	N/A	1.56	68.09	69.65
JTT	$O(nd)$ (1 fit)	$O(d)$	$O(1)$	104.76	$0.04^\dagger$	104.76
SELF (disagreement)	$O(nd)$ (2 fits)	$O(d)$	$O(d)$	197.97	$0.04^\dagger$	**197.97**

†Scoring with the resulting head is as cheap as margin; the entire cost is the one-time refit. Asymptotic columns omitted for Mahalanobis/k-NN, which are standard baselines rather than part of Sec. 3's method family.

G Hyperparameters and protocol

$\epsilon = 0.05$; head ℓ_2 with $C = 1$; reweighting strength λ selected from $\{0, 1, 2, 4, 8, 16, 32, 64\}$ by validation worst-group accuracy; ECE with 15 bins; multiplicity half-width via Eq. (9) with

the empirical Hessian at $\boldsymbol{w}_0$. Subsample sizes: CelebA 30k/12k/12k, CivilComments 30k/10k/10k (train/val/test); Waterbirds full. All method code is logistic regression on cached features; notebooks (feature extraction, method suite, theory simulation) are in the supplementary material.

$$\mathrm{hw}(\boldsymbol{x}) = \sqrt{2\epsilon L(\boldsymbol{w}_0)\, \tilde{\phi}(\boldsymbol{x})^\top \boldsymbol{H}^{-1} \tilde{\phi}(\boldsymbol{x})}. \tag{9}$$

Future work. We leave faithful large-held-out reproductions of Table 2, a full text-encoder architecture ablation (paralleling the CLIP vision-encoder ablation above), and a stochastic-encoder setting that could expose a variance advantage, to future work.