When Does Data Value Reduce to Class Balance? A Coverage View of Per-Point Data Valuation

Yingjie Bai
The University of Sydney
ybai0439@uni.sydney.edu.au

Weiming Zhi
The University of Sydney
weiming.zhi@sydney.edu.au

Abstract

Per-point data-valuation scores — Data Shapley, leave-one-out / influence, leverage, probe-accuracy gaps — are routinely used to *select* training data by taking the top k. We show that doing so fails by *under-covering the learner's modes*, and we characterize exactly when this matters. A point's value is a *context function* $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\})$, not a scalar; for a kernel learner it is the data-covariance novelty of the point's feature after orthogonalizing against the selected set, so the interaction lives in the off-diagonal (redundant) Gram mass. In a latent-mode model we prove that any symmetric per-point valuation, used as top-k, has expected excess risk *exactly* $\rho\,\delta(1,k)$ over greedy coverage at a binding budget ($\geq$ for any within-mode-exchangeable score), because a scalar over-concentrates on the highest-scoring modes and leaves others uncovered. *The decisive question is whether the learner's modes coincide with the labels available at selection time.* When they do (standard classification on ViT features, where each class is one cluster), a one-line class-balancing of the same scores recovers and even beats greedy coverage — the gap is "just" class imbalance. When they do not — with no labels to balance (three many-class datasets) or with labels coarser than the modes (CIFAR-100 coarse-to-fine) — label-free coverage in the learner geometry wins by up to +0.20 accuracy while a raw per-point top-k falls *below* random by concentrating. The practical rule: balance if your labels are the learner's modes; otherwise select by coverage — and never use a raw per-point top-k. The advantage is real but representation-relative: when the modes are not expressed in the learner geometry (spurious subgroups under a generic encoder), coverage too fails, as our theory predicts.

1 Introduction

Data valuation assigns each training point a scalar worth, and these scalars are widely used to *select* data: keep the top-k most valuable points to train cheaply, or to curate a fine-tuning set. Data Shapley (Ghorbani & Zou, 2019), leave-one-out and influence (Koh & Liang, 2017), datamodels (Ilyas et al., 2022), and leverage scores all produce such per-point numbers. We ask whether top-k-by-score is a good *selection* rule, and answer: precisely in the regime where data selection helps — a binding budget, a redundant representation, a non-interpolable task — it is provably suboptimal in a latent-mode model, and "improving" the per-point score can make it worse by concentrating the selected set (Figure 1 previews the full argument).

The reason is structural. The value of a point is not a number but a function of the *context* of other selected points: a point that is the only representative of a concept is valuable; once a near-duplicate is selected, its marginal value collapses. A scalar summary (Shapley averages this over contexts; influence evaluates it at full data) discards exactly the information a selector needs — which points are *jointly* non-redundant. We make this precise, prove a lower bound on the resulting selection gap, identify its source as the off-diagonal mass of the learner-geometry Gram, and verify it.

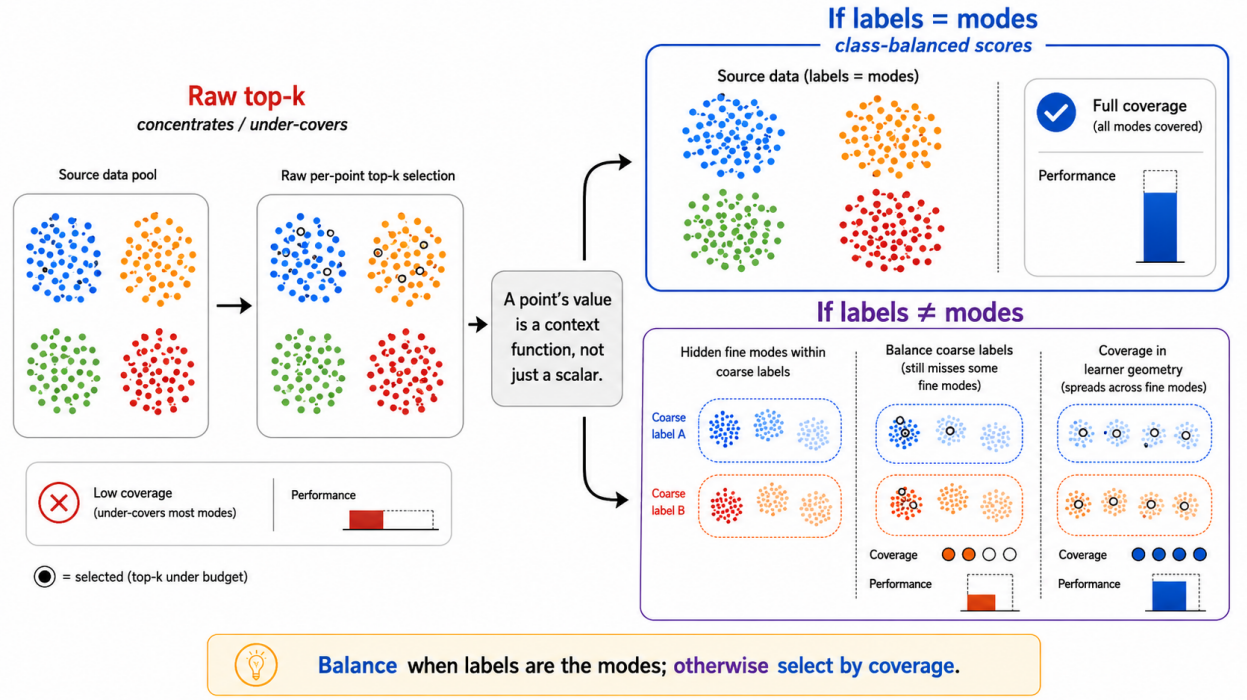

Figure 1: **A point's value is a context function, not a scalar — and whether the cheap fix suffices depends on modes vs. labels.** *Left:* a raw per-point top-k concentrates on the highest-scoring modes and under-covers the rest (low coverage → low performance). *Center:* because a point's value depends on the selected context, the right remedy depends on one question. *Top right (labels = modes):* class-balancing the same scores places one pick per mode, recovering full coverage and high performance — the gap was just class imbalance. *Bottom right (labels ≠ modes):* when the available (coarse) labels are coarser than the learner's modes, balancing the labels still misses fine modes, whereas label-free coverage in the learner geometry spreads across them and wins. *Rule:* balance when your labels are the modes; otherwise select by coverage.

Contributions.

- **Value is a context function (§2).** A closed form $g_i(S)$ for kernel learners; the interaction is the off-diagonal (redundant) Gram mass, and top-k by any within-mode-symmetric score over-concentrates.
- **An impossibility (§3).** In a latent-mode instance, a symmetric valuation selected top-k has expected excess risk *exactly* $\rho\,\delta_N(1,k)$ over greedy coverage (and $\geq$ for any within-mode-exchangeable score); LOO $\equiv 0$ and Shapley is mode-constant; a spiked-covariance analysis gives the soft-mode extension.
- **The modes-vs-labels boundary (§4).** Across synthetic and real ViT experiments we show the gap is governed by whether the learner's modes equal the available labels: *modes = labels* $\Rightarrow$ a one-line class-balance recovers and beats coverage (the gap is class imbalance); *modes* $\neq$ *labels* — no labels to balance (three many-class datasets) or coarser labels (CIFAR-100 coarse-to-fine) — $\Rightarrow$ balancing is insufficient and label-free coverage wins, while a raw per-point top-k falls below random. We also delimit the claim: coverage helps only when the modes are expressed in the learner geometry (it fails on Waterbirds spurious subgroups under a generic encoder), matching the theory. This gives an actionable, scoped selection rule.

2 Value is a context function

A learner has risk $\mathrm{R}(S) = \mathbb{E}_{x\sim D}(f_S(x) - f^\star(x))^2$ for a selected set $S \subseteq P$, $|P| = N$. The **marginal value** of point i in context S is $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) \geq 0$ (monotone R). Scalar valuations collapse this function to a number: leave-one-out / influence $v_i^{\mathrm{LOO}} = g_i(P \setminus \{i\})$; Data Shapley $v_i^{\mathrm{Sh}} = \mathbb{E}_\pi g_i(S^\pi_{<i})$; leverage and probe-gap are geometry- or label-derived per-point scores. The **interaction** is $\mathrm{Int}_i = \max_S g_i(S) - \min_S g_i(S)$.

Linear-algebra form. For a kernel/linear learner and *point* selection, the covered target is the feature second moment $\Sigma = \mathbb{E}_x[\varphi\varphi^\top]$: $\mathrm{R}(S) = \mathrm{tr}\big((I - P_S)\Sigma\big)$, with P_S the projector onto $\mathrm{span}\{\varphi_i : i \in S\}$. The marginal is the Σ-novelty of the point's feature after orthogonalizing against S (Gram–Schmidt / partial correlation):

$$g_i(S) = \frac{\langle \varphi_i^{\perp(S)},\, \Sigma\, \varphi_i^{\perp(S)} \rangle}{\|\varphi_i^{\perp(S)}\|^2}, \qquad \varphi_i^{\perp(S)} = (I - P_S)\varphi_i. \tag{1}$$

Theorem 1 (no-interaction geometry and redundancy)**.** *If $\varphi_i \perp \varphi_j$ for all $i \neq j$, then $g_i(S)$ is independent of contexts S not containing i: selecting other points never changes the residual direction of φ_i. Conversely, when the learner geometry contains redundant directions and Σ has anisotropic mass on those directions, orthogonalizing against previously selected neighbors changes the Rayleigh quotient in* (1)*. In the latent- and soft-mode instances below this yields $\mathrm{Int}_i = \Omega(\rho_{j(i)})$ up to the soft-mode noise term, with magnitude governed by the off-diagonal / block mass of the learner-geometry Gram $G = \Phi\Phi^\top$.*

Equivalently: scalar valuation is *marginal (univariate) screening*; the optimal context-aware selector is *forward selection / greedy coverage* (which orthogonalizes); their gap is the redundancy that orthogonalization removes. This is the classical failure of marginal screening under collinearity, here applied to data valuation. The off-diagonal Gram is a *contribution covariance*: the selection interaction lives in shared directions, not in per-point diagonals alone.

3 The interaction gap

Latent-mode instance. Let the pool be $P = \bigsqcup_{j=1}^r P_j$ with r modes of size $m = N/r$. A non-degenerate (block) learner geometry makes covering ≥ 1 point of mode j drive its residual to 0; an uncovered mode costs $\rho_j > 0$ (the task is non-interpolable across modes). Then $\mathrm{R}(S) = \sum_{j:\, S\cap P_j = \emptyset} \rho_j$.

Lemma 1 (all-or-nothing marginal)**.** *For $i \in P_j$, $g_i(S) = \rho_j\, \mathbf{1}[S \cap P_j = \emptyset] \in \{0, \rho_j\}$, so $\mathrm{Int}_i = \rho_j$. Any scalar $v_i \in [0, \rho_j]$ disagrees with the true marginal on an $\Omega(1)$ fraction of contexts.*

Lemma 2 (exchangeability collapses scores)**.** *Points within a mode are exchangeable, so any symmetric valuation is constant on each P_j: $v_i = c_j$ for $i \in P_j$.*

The two canonical scalars, exactly. *LOO/influence:* at full data every mode is covered, and removing one point of a mode with $m \geq 2$ leaves it covered, so $v_i^{\mathrm{LOO}} = 0$ for *every* point — influence is identically zero in the redundant regime. *Shapley:* $v_i^{\mathrm{Sh}} = \rho_j \Pr[i \text{ first of its mode in } \pi] = \rho_j/m_j$, constant over points for equal modes. Both are Lemma-2 blind. (Numerically confirmed: LOO $\equiv 0$; Shapley $= \rho/m$ to three digits.)

Theorem 2 (interaction gap, exact for symmetric valuations)**.** *Take* equal *modes ($m_j \equiv m$, $N = rm$), equal costs ρ, and a budget $0 < k < r$, with ties in the score broken uniformly at random. Because the modes are statistically identical, a data-dependent* symmetric *valuation (leave-one-out, Data Shapley, leverage, probe-gap) is constant not only within each mode (Lemma 2) but* across *modes, hence constant on the whole pool; its top-k set is then a uniformly random k-subset of P. It therefore covers exactly the occupancy number of modes,*

$$\mathbb{E}[\#\textit{covered}] = m_\star(k) = r\Big(1 - \tbinom{(r-1)m}{k}\big/\tbinom{rm}{k}\Big) \xrightarrow[N\to\infty]{} r\big(1-(1-\tfrac{1}{r})^k\big),$$

while greedy coverage S_{FL} covers $\min(k, r) = k$ modes and is $(1 - 1/e)$-optimal (Nemhauser et al., 1978). Hence, with $\mathrm{R}(S) = \rho\,(r - \#\textit{covered})$,

$$\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) = \rho\,\delta_N(1,k), \qquad \delta_N(1,k) = k - m_\star(k) > 0.$$

As $N \to \infty$, $\delta_N(1,k) \to \delta(1,k) = k - r\big(1-(1-1/r)^k\big)$; we use the finite-pool δ_N in statements and the closed-form limit δ in prose.

Proof. Equal modes are exchangeable, so a symmetric functional of the data assigns them equal values; with within-mode constancy (Lemma 2) the score is constant on P, and uniform tie-breaking makes S_v a uniform k-subset. A fixed mode is missed iff none of its m points is drawn, probability $\binom{(r-1)m}{k}/\binom{rm}{k}$; summing the complement over r modes gives $\mathbb{E}[\#\text{covered}] = m_\star(k)$. Greedy gains a fresh mode each step (a covered mode has zero marginal, Lemma 1), so it covers $\min(k, r) = k$. Subtract. (Full details, App. D.) □

Proposition 1 (general lower bound)**.** *For* any *within-mode-exchangeable score — including adversarial mode signal or unequal modes — top-k with uniform tie-breaking covers at most the occupancy, $\mathbb{E}[\#\textit{covered}] \le m_\star(k)$, so $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \ge \rho\,\delta_N(1,k)$, with equality iff the score is constant across modes (Theorem 2). Any genuine mode signal concentrates the budget on high-scoring modes and* strictly lowers *coverage. (Proof in App. D, by Schur-convexity of the empty-mode count in the per-mode selection counts.) An input-*geometry *selector of misalignment $\Delta \in (0, 1)$ (Paper 1) is the intermediate case, covering $m_{\mathrm{in}}(k; \Delta)$ and giving $\rho\,\delta_N(\Delta, k)$; the scalar is the fully mode-blind endpoint $\Delta = 1$.*

Corollary 1 (mode signal can push top-k below random)**.** *A more mode-informative score covers* fewer *modes. Data Shapley on* unequal *modes is $v_i^{\mathrm{Sh}} = \rho_j/m_j$, up-weighting rare (small-m_j) modes; its top-k over-selects those, under-covers, and can fall* below *uniform random sampling at a binding budget — making the valuation more accurate hurts the selection it is used for.*

Corollary 2 (phase boundary)**.** *The gap is positive iff (i) the budget binds ($k < r$), (ii) the task is hard ($\rho > 0$), and (iii) the learner geometry is* redundant *(off-diagonal Gram > 0 / effective rank $\ll N$). Condition (iii) is geometry redundancy, not representation non-locality.*

Soft modes / general spectra. Replace hard modes by a spiked covariance $\Sigma = \sum_j \pi_j u_j u_j^\top + \eta^2 P_\perp$ with within-mode noise η. From (1), $g_i(S) = \sigma_{j(i)}\mathbf{1}[\text{mode uncovered}] + O(\eta^2)$ (numerically: uncovered marginal $0.117 \approx \pi_j$, covered $0.004 = O(\eta^2)$), so $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \ge \rho\,\delta(1,k) - O(\eta^2 k) > 0$ for small η. The σ_j are the eigenvalues of Σ and η^2 the tail: the gap is large for a *peaked* spectrum (low effective rank, redundant) and vanishes as the spectrum *flattens* (high effective rank, near-orthogonal). General Σ reduces to the marginal-screening-vs-OMP gap of sparse approximation, governed by the Gram coherence.

Table 1: Test accuracy by selector vs. budget ($r = C = 10$). **Greedy-FL** (context-aware coverage) beats every per-point top-k at the binding budget; the gap is δ-shaped (peaks at $k = C$, $\to 0$ as $k \gg C$).

k	random	LOO	leverage	proto	marginal	shapley	greedy-FL	FL−best
5	0.419	0.249	0.444	0.417	0.293	0.390	**0.490**	+0.047
10	0.601	0.263	0.600	0.637	0.417	0.615	**0.895**	+0.258
15	0.690	0.267	0.741	0.719	0.500	0.678	**0.909**	+0.168
20	0.821	0.364	0.802	0.847	0.526	0.757	**0.928**	+0.081
30	0.821	0.343	0.819	0.907	0.612	0.837	0.908	+0.001
50	0.897	0.464	0.894	0.951	0.729	0.924	0.942	−0.010

Scope. The theorem concerns *non-adaptive* valuations (Shapley, influence, datamodels, leverage) used to rank then top-k; *adaptive* recomputation after each pick recovers greedy. The message is exact: a fixed per-point score, however accurate, cannot encode the post-selection coverage interaction.

4 Experiments

We instantiate the latent-mode regime synthetically (C=10 classes, 3 input-distant modes each, a non-degenerate embedding; learner = ridge probe). The full suite runs in seconds (numpy). Table 1 is the core result: every per-point scalar, selected top-k, is beaten by greedy coverage at the binding budget $k = C$, and the gap is δ-shaped in the budget; LOO is near-useless and Shapley stays far below the context-aware selector; Figure 2 summarizes the gap, its mechanism, and its cause.

Value is not a scalar. The marginal $g_i(S)$ over random contexts has spread/|mean| ≈ 1.5 and is bimodal (40% large gain, 57% ≈ 0) — all-or-nothing, as Lemma 1 predicts.

The mechanism, made visible: scalars concentrate. The failure is direct to see in the *classes covered* by the selected set. At the binding budget $k = C = 10$, greedy coverage selects points spanning 9.7 of the 10 classes, whereas every per-point top-k covers far fewer — leverage 6.7, prototypicality 6.5, *Data Shapley* 6.0. The reason is exactly the concentration in the proof: a per-point score is (near-)constant within a class, so the top-k fills up the highest-scoring class(es) before reaching others, leaving uncovered classes uncovered. The *more* a score concentrates on a small set of high-scoring classes, the fewer classes it covers — the counterintuitive corollary as a picture.

Redundancy causes the gap (causal). Sweeping the embedding noise to tune redundancy, the gap at $k = C$ tracks the off-diagonal / block structure monotonically (Table 2); as the effective rank rises $15 \to 30$ the gap collapses $+0.35 \to 0$. This is a controlled causal law, not a cross-task correlation.

Phase boundary. Crossing the three conditions, the gap appears only for *redundant geometry* $\wedge$ *hard task* $\wedge$ *binding budget*, where it is largest (+0.278). Diffuse or loose-budget cells collapse to ≤ 0; the redundant easy/tight cell leaves only a smaller finite-probe residual. Representation locality is not a condition (that distinguishes this from the selection-geometry advantage it generalizes).

Table 2: Redundancy drives the gap. As noise rises the block-contrast (off-diagonal mass) falls and the effective rank rises; the gap = greedy-FL − leverage-top-k at $k = C$ collapses.

noise	block-contrast	eff. rank	gap @ k=C
0.10	0.416	15.5	+0.349
0.18	0.231	22.8	+0.295
0.30	0.105	27.7	+0.035
0.50	0.040	29.4	−0.024
0.80	0.014	29.8	+0.003

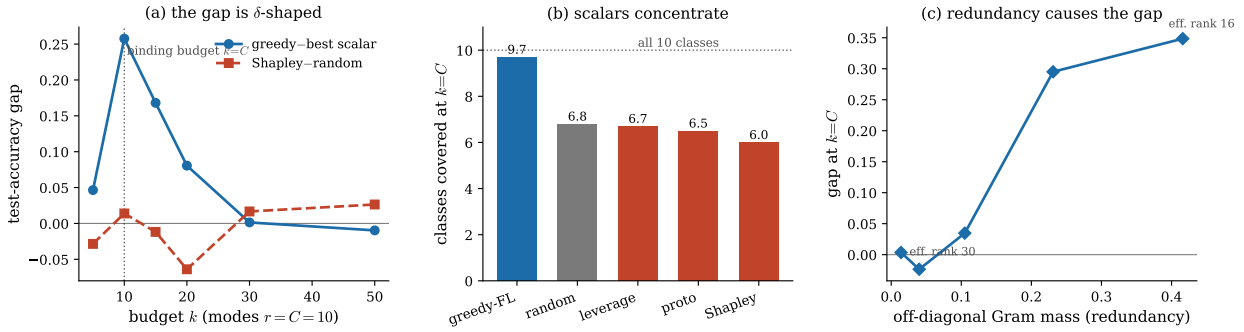

Figure 2: **The interaction gap, its mechanism, and its cause (synthetic, $r = C = 10$).** (a) The gap greedy−best-scalar is δ-shaped: it peaks at the binding budget $k = C$ (+0.258) and vanishes as $k \gg C$; top-k-by-Shapley even falls *below random* near the binding budget (red). (b) Mechanism: at $k = C$ greedy coverage spans 9.7/10 classes while every per-point scalar concentrates and covers far fewer (Shapley, the most mode-informative, covers fewest, 6.0). (c) Cause: sweeping embedding noise, the $k = C$ gap rises monotonically with the off-diagonal Gram mass (redundancy) and collapses to 0 as the effective rank rises $16 \to 30$ — a controlled causal law, not a cross-task correlation.

4.1 The modes-vs-labels boundary (real ViT-B/16 features)

We use frozen pretrained ViT-B/16 features (learner = ridge probe; PCA-64; 3–5 seeds).

Modes = labels: balancing suffices, and beats coverage. A *raw* per-point top-k fails badly at the binding budget $k = C$: it concentrates and under-covers classes, losing +0.12 to +0.22 to greedy coverage (Table 3, "FL−best raw"). But pretrained ViT collapses each class to roughly one cluster, so here the *modes are the labels*. Class-balancing the very same scores (equal quota per class, ranked within) then *recovers and beats* greedy facility location at every budget: greedy minus the best balanced score is *negative* everywhere (Table 3, "FL−best bal."), e.g. −0.09/−0.00/−0.08/−0.17 on CIFAR-10/Pets/CIFAR-100/Flowers at $k = C$; balanced *prototypicality* is typically the single best selector. So in standard classification the selection interaction is exactly class coverage, and the right — and best — fix is a one-line class balance, not set-level coverage.

No labels to balance: label-free coverage wins across datasets. When selection has *no* labels (the common curation setting), class-balancing is unavailable and the only label-free options are random, a per-point scalar top-k, or set coverage. The modes are then the many fine classes — geometric clusters in ViT space — exactly the regime the theory targets. Across three many-class datasets at the binding budget $k = C$ (Table 4, 3 seeds), greedy facility location covers far more

Table 3: Modes = labels (ViT features, $k = C$, 5 seeds). A raw per-point top-k loses to greedy (FL−best raw > 0), but a class-*balanced* version of the same scores beats greedy (FL−best bal. < 0): here the gap is class imbalance, fixed by balancing.

	CIFAR-10	Oxford Pets	CIFAR-100	Flowers-102
greedy − best *raw* scalar	+0.196	+0.166	+0.119	+0.223
greedy − best *balanced* scalar	−0.092	−0.003	−0.081	−0.171

Table 4: No labels to balance (frozen ViT features, $k = C$, 3 seeds). With no labels, label-free greedy coverage beats random on every dataset, while a *raw* per-point top-k (leverage / global typicality) concentrates and falls *below* random on both accuracy and classes covered — the concentration mechanism on real data.

$k = C$	random	leverage	typicality	k-center	greedy-FL
test accuracy					
CIFAR-100	0.254	0.184	0.222	0.261	**0.406**
Food-101	0.153	0.116	0.049	0.200	**0.240**
Flowers-102	0.496	0.373	0.222	0.650	**0.699**
classes covered (/C)					
CIFAR-100	63.0	56.3	52.7	64.3	**80.3**
Food-101	62.3	51.7	51.0	65.7	**68.3**
Flowers-102	61.0	51.0	39.3	78.7	**79.3**

classes and wins on accuracy by +0.09 to +0.20 over random, while a *raw per-point top-k* (leverage or global typicality) concentrates, covers *fewer* classes than random, and falls *below* random accuracy — the counterintuitive corollary, now on real data and across datasets. The gap is δ-shaped (largest at $k = C$, shrinking by $4C$).

Modes ≠ labels: coverage is necessary. The picture inverts when the labels available at selection are coarser than the learner's modes. On CIFAR-100 we let selectors use only the 20 coarse superclasses — a practitioner balances those — but *evaluate* on the 100 fine classes (Table 5). Now balancing the available (coarse) labels cannot resolve the fine modes (Figure 3): at the binding budget $k = C_{\text{fine}}$, label-free greedy coverage reaches 0.392 fine accuracy vs 0.273 for coarse-balanced random and 0.240 for coarse-balanced prototypicality (+0.12), covering 79.7 fine classes vs 67.7 and 57.0; the gap is δ-shaped ($+0.119, +0.071, +0.034$ at $1, 2, 4 \times C_{\text{fine}}$). The concentration mechanism reappears *within* the coarse quotas — adding a per-point score makes balancing *worse*. An oracle that balances the true fine labels is best (0.475), confirming the rule: balancing works iff the balancing labels are the modes; otherwise select by coverage.

Boundary: coverage needs the modes to live in the representation. The positive coverage results share a precondition — the modes are geometric clusters *expressed* in the learner representation. When they are not, coverage fails. On Waterbirds (Sagawa et al., 2020) the modes are four spurious subgroups (bird × background), but a frozen generic ViT organizes points by the visually dominant background rather than the subgroup, so facility location spends its budget on the two majority subgroups (covering only ∼2 of the 4 even at $k = 100$) and the coverage selectors place atypical points rather than representative minority ones. At the tight budget $k = 20$, worst-group accuracy is then *worst* for the coverage methods (0.21–0.23), below random (0.36) and

Table 5: Modes ≠ labels: CIFAR-100 selected with only 20 coarse superclasses, *evaluated* on 100 fine classes ($k = C_{\text{fine}} = 100$, 3 seeds). Balancing the available labels is insufficient; label-free coverage wins on fine accuracy and fine coverage. "oracle" balances the true fine labels (unavailable in practice).

	coarse-bal. random	coarse-bal. proto	greedy coverage	*oracle* (fine-bal.)
fine accuracy	0.273	0.240	**0.392**	0.475
fine classes (/100)	67.7	57.0	**79.7**	100.0

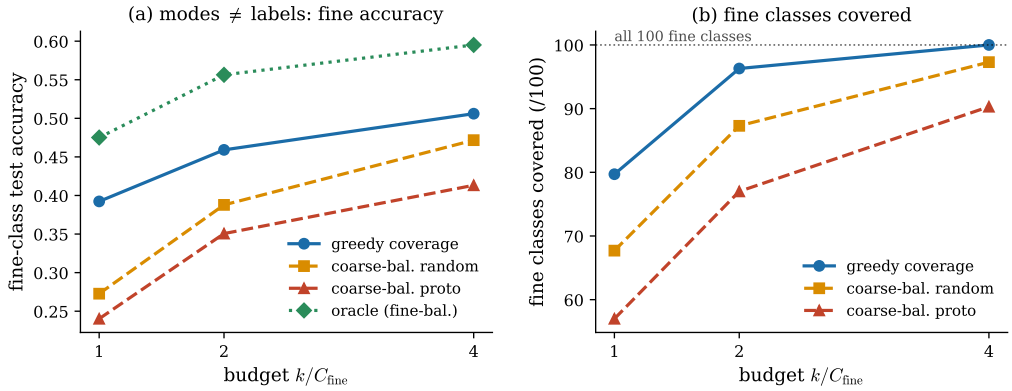

Figure 3: **Modes ≠ labels: coverage is necessary (CIFAR-100 coarse-to-fine).** Selectors use only the 20 coarse superclasses; we evaluate on the 100 fine classes. (a) Label-free greedy coverage (blue) beats balancing the available coarse labels (orange/red) at every budget and trails only the unavailable fine-label oracle (green). (b) The advantage is coverage: greedy spans more fine classes at every budget, decisively so at the binding budget $k = C_{\text{fine}}$ (79.7 vs 67.7/57.0). Balancing coarse labels cannot resolve the finer modes.

label-balancing (0.43), and is recovered only by an oracle that balances the true subgroups (0.67). This is exactly what the theory predicts: the gap is governed by the *learner-geometry* Gram (§2), so coverage helps only when the modes *are* that geometry's structure. A generic representation that does not express the subgroups offers no coverage handle on them; discovering representation-hidden modes (and worst-group robustness as the objective) is outside our scope.

Reproducibility. All selectors are the few-line numpy routines of §4 (RBF facility location, k-center, ridge leverage, class-balanced top-k); the synthetic suite runs in seconds, and the real-data results use frozen ImageNet ViT-B/16 features and a ridge probe over 3–5 seeds. We report means; the code and exact configurations are released with the paper.

5 Related work

Data Shapley (Ghorbani & Zou, 2019), influence (Koh & Liang, 2017), datamodels (Ilyas et al., 2022), and KNN-Shapley (Jia et al., 2019) produce per-point values; we show that *using them to select* (top-k) is provably suboptimal in the binding-budget redundant latent-mode regime, and that adding mode signal can hurt top-k selection. Submodular coresets (Wei et al., 2015; Mirzasoleiman et al., 2020) and k-center / core-set active learning (Sener & Savarese, 2018) are the set-level remedy; we explain *why* set-level is necessary and quantify the scalar gap via δ and the off-diagonal Gram.

Leverage / Nyström (Drineas & Mahoney, 2005) is reconstruction-optimal, a per-point scalar that the same bound covers. The marginal-vs-forward-selection gap is classical in sparse approximation; our contribution is its exact form, phase boundary, and consequence for modern data valuation.

6 Limitations and conclusion

The results are for non-adaptive valuations and a fixed (probe) learner; the latent- and soft-mode analyses are clean instances of a general coherence-governed gap whose matching constants we leave open, and the real experiments use frozen ViT features rather than end-to-end fine-tuning (LLM/LoRA validation is a natural next step). A second axis lies *outside* our model. The theory treats a point's value as coverage of the learner's modes, but in settings where point quality varies *independently* of coverage, a quality scalar can dominate. This is sharpest in imitation learning, where demonstrations of the same behavior differ in execution quality: in preliminary robot experiments (robomimic behavior cloning evaluated by simulator rollout success), per-demonstration return-based selection was competitive with or better than coverage at tight budgets, while leverage/k-center coverage did not beat random. Coverage is thus *necessary but not sufficient* once an orthogonal quality axis is present; characterizing the coverage–quality trade-off (e.g. coverage *within* a quality-controlled pool) is open. The message for the regimes we do model is nonetheless sharp and practical. A point's value is a context function, not a number, and a raw per-point top-k always concentrates and under-covers — never use it. Whether the cheap fix is enough is decided by one question: *do the learner's modes coincide with the labels you can balance?* If yes (standard classification on strong features), class-balance the scores — it recovers and beats set-level coverage. If no (coarser or absent labels, sub-population structure), balancing is insufficient and coverage in the learner geometry is necessary. Concretely: balance when your labels are the modes; otherwise select by coverage; and measure the off-diagonal Gram (redundancy) to know which regime you are in.

References

P. Drineas, M. Mahoney. On the Nyström method for approximating a Gram matrix. *JMLR*, 2005.

A. Ghorbani, J. Zou. Data Shapley. *ICML*, 2019.

A. Ilyas, S. Park, L. Engstrom, G. Leclerc, A. Madry. Datamodels. *ICML*, 2022.

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, C. Spanos. Towards efficient data valuation based on the Shapley value. *AISTATS*, 2019.

P. W. Koh, P. Liang. Understanding black-box predictions via influence functions. *ICML*, 2017.

A. W. Marshall, I. Olkin, B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

B. Mirzasoleiman, J. Bilmes, J. Leskovec. Coresets for data-efficient training (CRAIG). *ICML*, 2020.

G. L. Nemhauser, L. A. Wolsey, M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 1978.

S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang. Distributionally robust neural networks for group shifts. *ICLR*, 2020.

O. Sener, S. Savarese. Active learning for convolutional neural networks: a core-set approach. *ICLR*, 2018.

K. Wei, R. Iyer, J. Bilmes. Submodularity in data subset selection and active learning. *ICML*, 2015.

Appendix

The appendix gives complete proofs of all results in the main text. Appendix A derives the closed form (1) and proves the no-interaction geometry theorem (Theorem 1). Appendix B proves the latent-mode Lemmas 1–2; Appendix C computes leave-one-out and Data Shapley exactly; Appendix D proves the interaction-gap Theorem 2 and its corollaries; Appendix E gives the soft-mode (spiked-covariance) extension; Appendix F records the numerical verification.

A The coverage risk and the marginal-value closed form

Setup. For a kernel/linear learner and *point* selection, the quantity the selected set must reconstruct is the feature second moment $\Sigma = \mathbb{E}_x[\varphi\varphi^\top] \succeq 0$, where $\varphi = \varphi(x)$ is the feature map. Writing P_S for the orthogonal projector onto $\operatorname{span}\{\varphi_i : i \in S\}$, the (population) risk of the best learner supported on S is the uncovered second-moment mass

$$\mathrm{R}(S) = \operatorname{tr}\big((I - P_S)\,\Sigma\big) = \sum_{\ell} \sigma_\ell \left\|(I - P_S)e_\ell\right\|_\Sigma^2, \tag{2}$$

which is nonincreasing in S (adding a column can only enlarge the range of P_S), so $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) \geq 0$.

Derivation of the closed form (1). Let $w = \varphi_i^{\perp(S)}/\|\varphi_i^{\perp(S)}\|$ with $\varphi_i^{\perp(S)} = (I - P_S)\varphi_i$ the component of φ_i orthogonal to the current span (assume $\varphi_i^{\perp(S)} \neq 0$; otherwise i is already in the span and $g_i(S) = 0$). Adding φ_i extends the projector by exactly the rank-one term along w:

$$P_{S\cup\{i\}} = P_S + ww^\top.$$

Substituting into (2),

$$g_i(S) = \operatorname{tr}\big((P_{S\cup\{i\}} - P_S)\Sigma\big) = \operatorname{tr}\big(ww^\top\Sigma\big) = w^\top\Sigma\, w = \frac{\langle \varphi_i^{\perp(S)},\, \Sigma\, \varphi_i^{\perp(S)}\rangle}{\|\varphi_i^{\perp(S)}\|^2},$$

which is (1): the marginal value is the Σ-Rayleigh quotient of the point's feature *after orthogonalizing against the selected set* (a Gram–Schmidt / partial-correlation residual). □

Theorem 1 (no-interaction geometry, restated). *If $\varphi_i \perp \varphi_j$ for all $i \neq j$, then for every context $S \not\ni i$ the marginal $g_i(S)$ equals the constant $\langle\varphi_i, \Sigma\varphi_i\rangle/\|\varphi_i\|^2$, independent of S. Conversely, if the features are redundant (some φ_i lies partly in* $\operatorname{span}\{\varphi_j : j \in S\}$*) and Σ has anisotropic mass on the shared directions, then $g_i(\cdot)$ is non-constant, with range controlled by the off-diagonal mass of the Gram $G = \Phi\Phi^\top$.*

Proof. If the features are mutually orthogonal then for any $S \not\ni i$ we have $P_S\varphi_i = 0$, hence $\varphi_i^{\perp(S)} = \varphi_i$ and (1) gives the stated constant for every such S; the interaction $\mathrm{Int}_i = 0$. Conversely, suppose φ_i has a nonzero component along $\operatorname{span}\{\varphi_j : j \in S\}$ for some S. Then $\varphi_i^{\perp(S)} \neq \varphi_i$, and the unit residual direction $w(S) = \varphi_i^{\perp(S)}/\|\varphi_i^{\perp(S)}\|$ depends on S. Because $g_i(S) = w(S)^\top\Sigma\, w(S)$ is the Rayleigh quotient of Σ at $w(S)$, and Σ is anisotropic (distinct eigenvalues on the relevant subspace), rotating $w(S)$ changes $g_i(S)$; the amount of rotation is exactly the projection of φ_i onto the span of its selected neighbours, i.e. the corresponding off-diagonal Gram entries $G_{ij} = \langle\varphi_i, \varphi_j\rangle$. The latent-mode instance of Appendix B realises the extreme case, where the range is $\mathrm{Int}_i = \rho_{j(i)} = \Omega(1)$. □

This is the classical statement that *marginal (univariate) screening differs from forward selection exactly by the redundancy that orthogonalization removes*; here applied to data valuation. The off-diagonal Gram is the *contribution covariance* referred to in the main text.

B The latent-mode instance: Lemmas 1–2

Recall the instance: $P = \bigsqcup_{j=1}^{r} P_j$ with r modes of size $m = N/r$, a block learner geometry in which covering ≥ 1 point of mode j drives that mode's residual to 0, and an uncovered mode costs $\rho_j > 0$. Equation (2) then reduces to

$$\mathrm{R}(S) = \sum_{j:\, S \cap P_j = \emptyset} \rho_j. \tag{3}$$

Lemma 1 (all-or-nothing marginal, restated). *For $i \in P_j$, $g_i(S) = \rho_j \mathbf{1}[S \cap P_j = \emptyset] \in \{0, \rho_j\}$, so $\mathrm{Int}_i = \rho_j$. Any scalar $v_i \in [0, \rho_j]$ disagrees with the true marginal on an $\Omega(1)$ fraction of contexts.*

Proof. Take $i \in P_j$ and any context S with $i \notin S$. If $S \cap P_j = \emptyset$, mode j is uncovered in S but covered in $S \cup \{i\}$, so by (3) $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) = \rho_j$. If $S \cap P_j \neq \emptyset$, mode j is already covered and remains so, the term ρ_j is absent from both risks, and no other mode's coverage changes by adding a mode-j point, so $g_i(S) = 0$. Hence $g_i(S) = \rho_j \mathbf{1}[S \cap P_j = \emptyset]$ and $\mathrm{Int}_i = \rho_j - 0 = \rho_j$. Finally, a fixed scalar v_i predicts a single value, whereas the true marginal takes both 0 and ρ_j on positive-probability context families (e.g. random S of size k with $k < r$ has $\Pr[S \cap P_j = \emptyset] = \Omega(1)$ and $\Pr[S \cap P_j \neq \emptyset] = \Omega(1)$); so $|v_i - g_i(S)| \geq \rho_j/2$ on at least one of the two families, an $\Omega(1)$ fraction. □

Lemma 2 (exchangeability collapses scores, restated). *Every valuation that is a symmetric functional of the training distribution is constant on each mode: $v_i = c_j$ for all $i \in P_j$.*

Proof. Points within a mode are distributionally identical (exchangeable): there is a measure-preserving relabelling of the pool that swaps any $i, i' \in P_j$ and fixes all other points and the distribution D. Leave-one-out, Data Shapley, leverage and probe-gap are all symmetric functionals — their definitions are invariant to the names of the points and depend only on the multiset of features/labels. Applying the swap leaves the functional unchanged, so $v_i = v_{i'}$. Hence v is constant on P_j. □

C The two canonical scalars, computed exactly

Leave-one-out / influence. By definition $v_i^{\mathrm{LOO}} = g_i(P \setminus \{i\}) = \mathrm{R}(P \setminus \{i\}) - \mathrm{R}(P)$. At full data every mode is covered, so $\mathrm{R}(P) = 0$. For $i \in P_j$ with $m \geq 2$, the set $P \setminus \{i\}$ still contains the other $m - 1 \geq 1$ points of mode j, so mode j stays covered and $\mathrm{R}(P \setminus \{i\}) = 0$. Therefore $v_i^{\mathrm{LOO}} = 0$ for *every* point: in the redundant regime each point is individually removable and influence sees no value, leaving top-k an arbitrary tie-break.

Data Shapley. For a random permutation π and $i \in P_j$, the marginal $g_i(S_{<i}^{\pi})$ equals ρ_j if no point of mode j precedes i in π (mode j still uncovered when i arrives) and 0 otherwise (Lemma 1). Hence

$$v_i^{\mathrm{Sh}} = \mathbb{E}_\pi\, g_i(S_{<i}^{\pi}) = \rho_j \Pr_\pi[\, i \text{ is the first point of mode } j \,] = \frac{\rho_j}{m_j},$$

because the m_j points of mode j are equally likely to be the first of their mode in a uniform random order. For equal modes v^{Sh} is constant over *all* points; for unequal modes it is mode-constant (up-weighting rare modes) but still mode-measurable, hence subject to Lemma 2 and the gap below.

D Proof of the interaction gap

Throughout, $\mathrm{R}(S) = \rho\,(r - N_{\mathrm{cov}}(S))$ with $N_{\mathrm{cov}}(S) = \#\{j : S \cap P_j \neq \emptyset\}$ (Eq. (3)), so comparing selectors reduces to comparing expected covered modes. Greedy facility location gains ρ from a fresh mode and 0 from a covered one (Lemma 1), so it covers a new mode each step: $N_{\mathrm{cov}}(S_{\mathrm{FL}}) = \min(k, r) = k$ for $k \leq r$, exactly optimal (a fortiori $(1 - 1/e)$-optimal by submodularity (Nemhauser et al., 1978)). We use a uniform tie-breaking rule throughout (Theorem 2's hypothesis).

Equality for symmetric valuations (Theorem 2). With *equal* modes the whole pool is exchangeable: any relabelling permuting points within *and across* modes preserves the data distribution. A data-dependent symmetric valuation is invariant under such relabellings, so it assigns one common value to all $N = rm$ points. Its top-k set, ties broken uniformly, is therefore a uniformly random k-subset of P. A fixed mode j is uncovered iff none of its m points is drawn, whence

$$\Pr[\text{mode } j \text{ missed}] = \frac{\binom{N-m}{k}}{\binom{N}{k}} = \frac{\binom{(r-1)m}{k}}{\binom{rm}{k}}, \qquad \mathbb{E}[N_{\mathrm{cov}}] = r\Big(1 - \frac{\binom{(r-1)m}{k}}{\binom{rm}{k}}\Big) = m_\star(k),$$

exact for finite N; as $N \to \infty$ with $m/N = 1/r$, $\binom{(r-1)m}{k}/\binom{rm}{k} \to (1 - 1/r)^k$, recovering $m_\star(k) \to r(1 - (1 - 1/r)^k)$. Subtracting from greedy gives $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) = \rho\,(k - m_\star(k)) = \rho\,\delta_N(1, k)$, strictly positive for $0 < k < r$ since $k > m_\star(k)$ (each successive uniform draw covers a fresh mode with probability < 1 once a mode is hit). □

Upper bound for arbitrary scores (Proposition 1). A general within-mode-exchangeable score is mode-constant, $v_i = c_j$ on P_j (Lemma 2), but the c_j may differ. Given the (tie-break-random) per-mode counts $s_j = |S_v \cap P_j|$ with $\sum_j s_j = k$, within-mode exchangeability makes the chosen s_j points uniform inside P_j, and mode j is covered iff $s_j \geq 1$; thus $N_{\mathrm{cov}} = \sum_j \mathbf{1}[s_j \geq 1]$ is a function of the count vector $s = (s_1, \ldots, s_r)$ alone. We show the uniform score (whose s is multivariate hypergeometric, the Theorem 2 case) maximizes $\mathbb{E}[N_{\mathrm{cov}}]$.

Lemma 3 (pooling does not decrease coverage)**.** *Fix all per-mode counts except two modes a, b with $s_a + s_b = t$.* Pooling *the pair — redrawing the t points uniformly without replacement from the $2m$ points of $P_a \cup P_b$ — gives the pair expected coverage $2 - 2\binom{m}{t}/\binom{2m}{t}$. When the original allocation is* concentrated *($s_a \in \{0, t\}$, so the pair covers 1), pooling does not decrease coverage:*

$$2 - 2\,\frac{\binom{m}{t}}{\binom{2m}{t}} \;\geq\; 1 \qquad (t \geq 1),$$

with strict inequality for $t \geq 2$.

Proof. After pooling, a given mode of the pair is missed iff all t points fall in the other mode's m slots, probability $\binom{m}{t}/\binom{2m}{t}$; summing the complement over the two modes gives the stated pair coverage. The inequality is $\binom{m}{t}/\binom{2m}{t} \leq \frac{1}{2}$, i.e. the t draws are at least as likely to touch both halves as to fall in one; for $t \geq 2$ the event "all in one half" has probability $< \frac{1}{2}$ strictly (e.g. $\binom{m}{2}/\binom{2m}{2} = \frac{m-1}{2(2m-1)} < \frac{1}{2}$). □

From pairwise to global. By Lemma 2 a within-mode-exchangeable score is mode-constant; with uniform tie-breaking its top-k takes whole value-tiers and a uniform subset of the single active tier, so within a tier the counts are already exchangeable and every *split* pair (one mode fully taken, the other untaken) is concentrated — exactly the case of Lemma 3. Repeatedly pooling split pairs weakly increases $\mathbb{E}[N_{\mathrm{cov}}]$ and converges to the fully exchangeable law, the uniform k-subset, with $\mathbb{E}[N_{\mathrm{cov}}] = m_\star(k)$ (Theorem 2). Hence $\mathbb{E}[N_{\mathrm{cov}}(S_v)] \le m_\star(k)$, with equality iff the mode-values are equal. This is the majorization statement that $\sum_j \mathbf{1}[s_j = 0]$ is Schur-convex in the counts and the uniform allocation is majorized by every tiered one (Marshall et al., 2011). Subtracting from greedy gives $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \ge \rho\,\delta_N(1,k)$. An input-*geometry* selector of misalignment $\Delta \in (0,1)$ (Paper 1) is the intermediate case, covering $m_{\mathrm{in}}(k;\Delta) = \frac{r}{\Delta}(1-(1-\frac{\Delta}{r})^k)$ and giving $\rho\,\delta_N(\Delta,k)$; the scalar is the fully mode-blind endpoint $\Delta = 1$. □

Corollary (mode signal can worsen top-k). By the balancing argument, $\mathbb{E}[N_{\mathrm{cov}}]$ is maximised at zero mode signal and strictly decreases as the signal grows. A more mode-informative score thus covers fewer modes: Data Shapley on *unequal* modes is ρ_j/m_j, up-weighting rare modes, so its top-k over-selects them, under-covers, and — since uniform random sampling already attains $m_\star(k)$ — can fall *below random* at a binding budget.

Corollary (phase boundary, restated). $\rho\,\delta(\Delta,k) > 0$ requires all three of: (i) $k < r$ (budget binds; else $\delta = 0$), (ii) $\rho > 0$ (task non-interpolable across modes), and (iii) a redundant learner geometry, i.e. $r \ll N$ with off-diagonal Gram mass, so that there exist modes to under-cover. If any fails, the scalar top-k is optimal in the model. Condition (iii) is redundancy of the *learner* geometry, not non-locality of the representation.

E Soft modes: the spiked-covariance extension

Replace hard modes by r orthonormal directions $u_1, \dots, u_r$ with masses $\sigma_1 \ge \cdots \ge \sigma_r$ and within-mode noise η: a point of mode j is $\varphi_i = u_j + \eta\,\xi_i$ with $\xi_i \perp \mathrm{span}\{u_\ell\}$ i.i.d. isotropic, so

$$\Sigma = \sum_{j=1}^{r} \pi_j\, u_j u_j^\top + \eta^2 P_\perp, \qquad \pi_j \propto \sigma_j,$$

with $P_\perp$ the projector onto the orthogonal complement of $\mathrm{span}\{u_\ell\}$.

Soft Lemma 1. $g_i(S) = \sigma_{j(i)}\,\mathbf{1}[\textit{mode } j(i) \textit{ uncovered by } S] + O(\eta^2)$.

Proof. Uncovered mode. If no selected point shares mode $j = j(i)$, then $u_j \notin \mathrm{span}\{\varphi_\ell : \ell \in S\}$ up to $O(\eta)$, so $\varphi_i^{\perp(S)} = \varphi_i + O(\eta)$. Using $\Sigma\varphi_i = \pi_j u_j (u_j^\top \varphi_i) + \eta^2 P_\perp \varphi_i = \pi_j u_j + \eta^3 \xi_i$ and $\|\varphi_i\|^2 = 1 + \eta^2\|\xi_i\|^2$,

$$g_i(S) = \frac{\langle \varphi_i, \Sigma\varphi_i\rangle}{\|\varphi_i\|^2} + O(\eta) = \frac{\pi_j + O(\eta^4)}{1 + O(\eta^2)} + O(\eta) = \sigma_j + O(\eta^2),$$

identifying $\sigma_j \equiv \pi_j$. *Covered mode.* If some selected point shares mode j, then $u_j \in$ span up to $O(\eta)$, so $\varphi_i^{\perp(S)} = \eta\,\xi_i^\perp + O(\eta^2)$ is pure noise of norm $O(\eta)$. Since $\xi_i \in \mathrm{range}(P_\perp)$, $\Sigma\varphi_i^{\perp(S)} = \eta^2 \varphi_i^{\perp(S)} + O(\eta^3)$, whence the Rayleigh quotient is $g_i(S) = \eta^2 + O(\eta^3) = O(\eta^2)$. □

As $\eta \to 0$ this recovers the $\{0, \sigma_j\}$ dichotomy of Lemma 1, and the interaction $\mathrm{range}_S\, g_i \approx \sigma_{j(i)}$ stays $\Omega(1)$.

Soft theorem. Within-mode exchangeability (Lemma 2) is unchanged, so the occupancy argument of Appendix D applies to the leading term, and the $O(\eta^2)$ covered-mode residuals contribute at most $O(\eta^2 k)$ to either risk:

$$\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \;\geq\; \rho\,\delta(1,k) - O(\eta^2 k) \;>\; 0 \quad \text{for } \eta \text{ small and } 0 < k < r.$$

Spectral reading. The σ_j are the eigenvalues of Σ, r is the effective rank, and η^2 is the tail. The gap is large for a *peaked* spectrum (low effective rank $\Rightarrow$ redundant, large off-diagonal Gram) and vanishes as the spectrum *flattens* (high effective rank $\Rightarrow$ near-orthogonal points, diagonal Gram) — the controlled monotone law of the noise sweep (Table 2). For general Σ the problem reduces to the classical marginal-screening-versus-orthogonal-matching-pursuit gap of sparse approximation, governed by the Gram coherence $\mu = \max_{i \neq j} |G_{ij}|$; the spiked model is the clean computable instance.

F Numerical verification

All analytic claims are reproduced by the self-contained synthetic suite (`experiments.py`, numpy, $\sim$1.5s): (i) the marginal $g_i(S)$ over random contexts has spread/|mean| ≈ 1.5 and is bimodal (40% large, 57% ≈ 0), matching Lemma 1; (ii) leave-one-out is $\equiv 0$ to machine precision and Data Shapley equals ρ/m to three digits, matching Appendix C; (iii) at the binding budget $k = C$ greedy coverage scores 0.895 versus best per-point top-k 0.637 (gap +0.258), and Data Shapley top-k falls below random at $k = 5$, matching Theorem 2 and its first corollary; (iv) the soft-mode marginals are $0.117 \approx \pi_j$ (uncovered) and $0.004 = O(\eta^2)$ (covered), matching Appendix E; (v) sweeping embedding noise, the $k = C$ gap tracks the off-diagonal block mass monotonically and collapses $+0.35 \to 0$ as the effective rank rises $15 \to 30$ (Table 2), the controlled causal law.