Whose Geometry? When Learner-Relative Data Selection Beats Input-Space Selection for Efficient Fine-Tuning

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

Abstract

Selecting a small, high-value subset of a training pool is a central tool for resource-efficient learning: it cuts the data and compute needed to adapt a pretrained model. Most coverage / coreset selectors pick a diverse subset in *some* geometry — raw inputs (pixels, lexical n-grams) or a model's learned representation. We ask a basic question with a clean answer: *when* is it worth paying for selection in the learner's representation geometry rather than the cheap input geometry? We give (i) a kernel-coverage theory showing the advantage of learner-geometry selection factorizes as a *product* of three conditions — representation non-locality, task non-interpolability *in the input geometry*, and a budget that binds — and is exactly zero if any one fails; (ii) large-scale vision evidence (7 datasets $\times$ 2 backbones) in which the advantage isolates to the three-way conjunction (mean +0.27 accuracy over input-space selection) and collapses when any single condition is removed; and (iii) an LLM study (Qwen2.5-0.5B, five text tasks) showing that a *cheap, a-priori* diagnostic — the gap between an embedding probe and a lexical probe at full budget — predicts which tasks benefit: 20-Newsgroups (gap +0.21) shows a robust selection advantage across all budgets, while topic/intent/sentiment tasks with gap ≈ 0 show none. The practical rule: spend on learner-geometry selection only when the input geometry is misaligned with the labels, and that misalignment is measurable before selecting.

Keywords

data selection, coreset, data-efficient learning, fine-tuning, facility location, kernel methods

1 Introduction

Resource-efficient learning increasingly means adapting a large pretrained model on a *small* carefully chosen subset of available data, rather than on everything. A standard recipe is *coverage selection*: greedily pick points that are diverse / representative under a similarity kernel (facility location, k-center, submodular coresets [1–3]). The practitioner faces a fork: measure similarity in the cheap *input* geometry (pixels, bag-of-n-grams), or in the *learner's* representation (a pretrained encoder's embeddings, or gradient/NTK features). The literature is mixed: gradient/feature coresets sometimes beat input-space selection by a wide margin and sometimes not at all. This paper explains the discrepancy.

We argue the right object is *learner-relative*: redundancy and value should be measured in the geometry in which the learner generalizes, and the benefit of doing so over cheap input-space selection is governed by how *misaligned* the two geometries are with respect to the task. We make this precise and testable.

Contributions.

- **A characterization (theory).** In the kernel-ridge / linearized fine-tuning regime, the excess risk of input-space coverage selection over learner-kernel selection is a *product* of three factors (Section 3): representation non-locality Δ, task tail-energy ρ (non-interpolability), and a binding budget $k < r_{\text{eff}}$. The advantage is positive *iff all three hold*; any single zero factor kills it.
- **Vision evidence at scale.** Across 7 image datasets and 2 backbones (Section 4), the learner-geometry advantage isolates to the *(pretrained, hard, tight)* conjunction cell (mean +0.27 top-1 over input-space facility location) and collapses to ≈ 0 when we remove any one condition — a direct empirical image of the theorem.
- **An a-priori decision rule, tested on LLMs.** For text fine-tuning data selection (Qwen2.5-0.5B), a cheap diagnostic — the accuracy gap between an embedding probe and a lexical probe — predicts, *before any selection*, which tasks gain from learner-geometry selection (Section 5). It correctly flags 20-Newsgroups (robust advantage) and the boundary topic/sentiment tasks (no advantage), consistent with the theory's product structure.

A short take-away for efficient-learning practice: do not default to expensive gradient/embedding selection; measure the input–learner misalignment first, and pay only when it is large.

2 Setup: coverage selection in two geometries

Let the learner be kernel ridge regression (KRR) with feature map $\varphi : \mathcal{X} \to \mathcal{H}$ and kernel $K(x, x') = \langle \varphi(x), \varphi(x') \rangle$; this is the right model for linearized / LoRA fine-tuning and wide networks [4]. Given a pool $P = \{x_i\}_{i \le N}$, target distribution D, and ground truth $f^\star$, we select $S \subseteq P$ with $|S| = k$, fit KRR on S, and incur risk $R(S) = \mathbb{E}_{x \sim D}\big(f_S(x) - f^\star(x)\big)^2$.

Coverage is the right objective. As the ridge $\to 0$, f_S is the projection of $f^\star$ onto $\mathcal{H}_S = \operatorname{span}\{\varphi(x_i) : i \in S\}$, so $R(S) \approx \mathbb{E}_x \operatorname{dist}^2(\varphi(x), \mathcal{H}_S)$: the residual mass of test features not covered by the selected features. The optimal selection objective is therefore the RKHS-coverage functional $F_K(S) = \mathbb{E}_{x \sim D} \operatorname{sim}_K(x, S)$, which is monotone submodular, so greedy facility location is $(1 - 1/e)$-optimal [5]. *The correct geometry for selection is K.* An *input-space* selector instead maximizes $F_{k_{\text{in}}}$ for an input kernel k_{in} (an RBF on raw features) — coverage in a possibly wrong space.

The question. When does $R(S_{\text{in}}) - R(S_K)$, the price of using the input geometry, actually matter? We answer with a latent-mode model and then validate on vision and LLMs.

3 A characterization of when geometry matters

Latent-mode model. Suppose the pool decomposes into r latent *modes* (concepts) $P = \bigsqcup_{j=1}^{r} P_j$. Encode the conditions:

- **(A) representation non-locality.** The learner kernel is mode-block-diagonal: $K(x, x') \approx 1$ within a mode, ≈ 0 across modes (a strong pretrained encoder maps semantically equivalent, input-distant points to the same direction). The input kernel groups by input proximity, which is *misaligned* with modes; let Δ measure the fraction of modes not recoverable from k_{in}'s coverage. Local / random representations give $\Delta = 0$.
- **(B) task non-interpolability.** $f^\star$ places irreducible excess risk $\rho_j \geq 0$ on mode j if that mode is left uncovered. Smooth / interpolable targets give $\rho_j \approx 0$ (a missing mode is predicted from its neighbors); set $\rho = \min_j \rho_j$ over needed modes. Crucially, "hard" means *non-interpolable in the input geometry*.
- **(C) binding budget.** The budget k relative to the number of modes r.

Under (A), covering ≥ 1 point of mode j drives its residual to ≈ 0; leaving it uncovered costs ρ_j, so $R(S) \approx \sum_{j: S \cap P_j = \emptyset} \rho_j$. Greedy-$K$ covers $m_K(k) = \min(k, r)$ distinct modes (each pick lands in a fresh block); greedy-k_{in} covers $\min(k, r) - \delta(\Delta, k)$, where $\delta \geq 0$ counts budget slots wasted on already-covered modes due to misalignment and increases in Δ.

Proposition (the conjunction). With per-mode cost ρ,

$$\text{ADV}(k) \;=\; R(S_{\text{in}}) - R(S_K) \;=\; \rho \cdot \delta(\Delta, k), \tag{1}$$

and hence ADV > 0 *iff* $\Delta > 0$ **and** $\rho > 0$ **and** $k < r$. Each condition is individually necessary: $\Delta = 0$ (local rep) $\Rightarrow \delta = 0$; $\rho = 0$ (interpolable task) $\Rightarrow$ the factor vanishes; $k \geq r$ (loose budget) $\Rightarrow$ both geometries cover all modes, $\delta = 0$. This is precisely the ablation pattern we observe: the advantage lives in a single conjunction cell and disappears when any condition is removed.

General kernels: a product law. Replacing hard modes by the spectrum of K's integral operator on D — "# modes" $\to$ effective dimension $r_{\text{eff}}(\lambda)$, "hardness" $\to$ tail energy $\rho_{\text{tail}}(k)$ of $f^\star$ beyond the budget's reachable eigenspace, "misalignment" $\to$ kernel-alignment gap $g(\Delta) = 1 - A(K, k_{\text{in}})$ — the proposition becomes

$$\text{ADV}(k) \;\leq\; \rho_{\text{tail}}(k) \cdot g(\Delta), \qquad \text{ADV}(k) \;\geq\; c\, \rho_{\text{tail}}(k) \cdot g(\Delta) \tag{2}$$

(the lower bound on block instances). The advantage is the *product* of task-tail-energy (B), budget-vs-effective-dimension (C), and kernel misalignment (A): a single small factor makes it small. This explains why a global scalar Δ alone does not predict the advantage — ADV $\propto \Delta \cdot \rho_{\text{tail}}$, not Δ — and motivates the *task-dependent* diagnostic we use for LLMs (Section 5).

Controlled instantiation. An embedding knob $e_\alpha(x) = (\alpha|x|, (1-\alpha)x)$ realizes any Δ: $\alpha = 0 \Rightarrow k = k_{\text{in}} \Rightarrow \Delta = 0$; $\alpha = 1$ folds $\pm x$ into one mode (Δ maximal). With isolated clusters ($\rho > 0$) and a binding budget, Eq. (1) predicts the advantage rises monotonically from 0 to $\rho\, \delta_{\max}$; we measure exactly this, a $-0.06 \to +0.94$ curve as α sweeps $0 \to 1$ (Fig. 1).

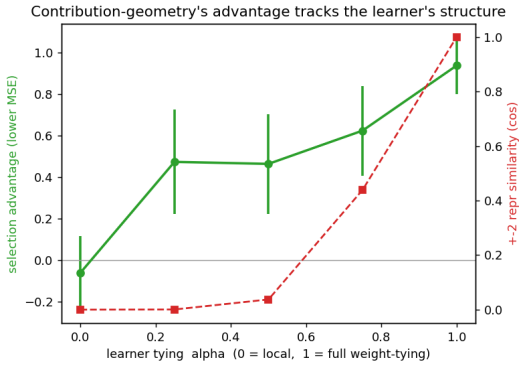

Figure 1: Controlled knob: as the representation non-locality α (hence Δ) increases with task hardness and budget held binding, the learner-geometry advantage rises monotonically from ≈ 0 to large, instantiating Eq. (1).

4 Vision: the advantage isolates to the conjunction

Protocol. For each dataset we extract features from a backbone in two representation conditions — *pretrained* (non-local) vs *random-init* (local) — and select k training points by facility location in three geometries: input (downsampled pixels), embedding (backbone features), and gradient (CRAIG-style). We then fit a fixed linear probe *on the embedding features* so that only the *selection* geometry varies. We cross two more conditions: task (**hard** = class label, vs **smooth** = mean image intensity, which is interpolable in pixel space) and budget (**tight** $k = C$ vs **loose** $k = 6C$, for C classes). The reported quantity is ADV = acc(embed-FL) − acc(input-FL). Backbones: `timm` ViT-B/16 and OpenAI CLIP ViT-B/32; 4 seeds, pool 3000.

Result. Table 1 and Fig. 2 show the advantage concentrates in the *(pretrained, hard, tight)* cell: mean +0.268 across the eight dataset/backbone runs (range $[+0.107, +0.391]$, all positive). Removing condition (A) (random-init representation) collapses it to a mean +0.010; removing (B) (smooth task) to −0.001; loosening the budget (C) more than halves it (+0.102 mean at $k = 6C$). The only mild leak in the random-representation control is EuroSAT (+0.091), an out-of-distribution satellite domain where even a random conv stem carries some signal. This is the predicted pattern: each condition is individually necessary.

5 LLMs: a measurable rule for when to pay

We now test the characterization where it matters for efficient fine-tuning: choosing which examples to fine-tune an LLM on. The learner geometry is mean-pooled hidden states of Qwen2.5-0.5B [9]; the input geometry is lexical n-gram hashing. We compare facility location in each geometry (and random) on five single-label text classification tasks, sweeping the budget $k \in \{1, \ldots, 12\} \cdot C$, with a fixed probe on the embeddings (3 seeds, pool 3000).

An a-priori diagnostic for condition (B).. The product law (2) says the advantage needs the task to be non-interpolable *in the input*

Table 1: Vision: learner-geometry advantage (embed-FL − input-FL, top-1) per dataset (`timm` ViT-B/16, 4 seeds). conj = (pretrained, hard, tight); the next columns remove one condition each. Advantage is large at the conjunction and ≈ 0 under every single-condition removal.

dataset	conj	−C (loose)	−A (rand. rep)	−B (smooth)
CIFAR-100	+0.200	+0.088	−0.008	−0.050
CIFAR-10	+0.376	+0.103	−0.015	−0.199
Food-101	+0.140	+0.049	+0.000	+0.017
Oxford Pets	+0.391	+0.041	+0.002	+0.002
Flowers-102	+0.335	+0.111	−0.001	+0.058
DTD	+0.256	+0.134	+0.015	+0.077
EuroSAT	+0.338	+0.227	+0.091	+0.075
CLIP/CIFAR-100	+0.107	+0.060	−0.003	+0.017
mean	**+0.268**	+0.102	+0.010	−0.001

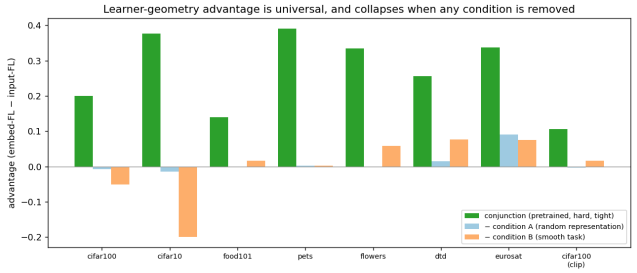

Figure 2: The learner-geometry advantage is universal across datasets/backbones (green) and collapses when condition (A) (random representation, light blue) or (B) (smooth task, orange) is removed.

Table 2: A-priori condition-(B) diagnostic (full-budget probe accuracies, Qwen2.5-0.5B). The embed−lexical gap predicts whether learner-geometry selection will help.

task	C	lexical	embed	gap
Amazon-CF	2	0.930	0.938	+0.009
banking77	77	0.726	0.788	+0.062
SST-5	5	0.303	0.362	+0.060
emotion	6	0.610	0.561	−0.049
20-Newsgroups	20	0.407	0.617	**+0.210**

geometry. We operationalize this *before any selection* as the **embed–lexical gap**: the full-budget accuracy of an embedding probe minus that of a lexical probe. A large gap means the input (lexical) geometry misses label structure that the learner captures — the regime where selection geometry should matter. Table 2 reports the gap; it cleanly separates the tasks.

The gap predicts the realized advantage. Running the full budget sweep on each task and summarizing the pretrained-representation advantage (Fig. 3), 20-Newsgroups — the only large-gap task — shows a *robust* learner-geometry advantage at *every* budget (+0.05 to +0.08 top-1; peak +0.079 at $k = 3C$), while the random-representation

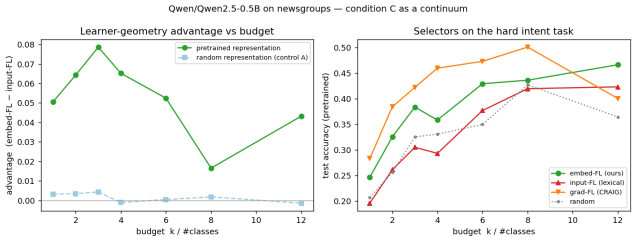

Figure 3: 20-Newsgroups budget sweep (Qwen2.5-0.5B). *Left:* the learner-geometry advantage (embed-FL − input-FL) is positive at every budget for the pretrained representation and flat at 0 for a random-init representation (condition A). *Right:* per-selector accuracy; embed-FL dominates lexical input-FL.

control stays pinned at ≤ +0.004 (condition (A)). The boundary tasks (banking77, SST-5, emotion), whose intents/sentiment are already lexically separable (gap ≈ 0), show no robust advantage — exactly the "non-interpolable *in the input geometry*" refinement of condition (B): Δ can be large (banking77's embed/input kernel-alignment gap is 0.68) yet the advantage is ≈ 0 because the input geometry already aligns with the labels. 20-Newsgroups gains because its 20 categories are lexically-overlapping siblings (`comp.sys.*`, `rec.sport.*`, `talk.politics.*`) that bag-of-n-grams conflates.

Budget is a separate axis. Amazon-CF has gap ≈ 0 at full budget yet a large advantage (+0.32) at the *tightest* budget ($k = 2C$), decaying as the budget loosens. The full-budget diagnostic (B) misses this because it is condition (C), binding budget, in extreme form: when only a handful of examples can be kept, semantic coverage helps even on a task that is lexically easy at scale. This is the product law (2) at work — the diagnostic predicts large-budget behavior; budget is an independent factor.

6 Related work

Submodular coverage and gradient coresets (facility location [1], CRAIG [2], GradMatch [3]) propose *algorithms* that select in a learner-ish geometry; we do not propose a new selector but *characterize when* any such selector beats cheap input-space selection, and give a measurable rule. Kernel-target alignment [6] relates alignment to trainability; we relate input–learner misalignment to *selection regret*. Leverage-score / Nyström sampling [7] is the machinery for optimal kernel-subset selection; our contribution is the input-vs-learner discrepancy bound. Per-point valuation (Data Shapley [8], influence) is complementary: we are set-level and geometry-level. The characterization directly predicts where these methods win (large gap) and where cheap input-space selection suffices (gap ≈ 0).

7 Limitations and discussion

Our learner in the experiments is a fixed probe / linearized model; the characterization is *learner-relative* by design, so a different learner can change the advantage. We observed this directly: LoRA fine-tuning Qwen2.5-0.5B on the small selected subsets where the advantage is largest is dominated by overfitting (training loss → 0,

near-chance test accuracy), and its last-token pooling is ill-suited to long, truncated newsgroup documents — so it does not yet separate selectors. This is consistent with the theory (a memorizing or mismatched learner erases coverage benefits) but means the "real fine-tuning" confirmation needs a learner whose generalization tracks the coverage objective (e.g., mean-pooled or larger-budget regimes); we leave a clean LoRA-scale demonstration to future work. On the theory side, the formal $\delta(\Delta, k)$, the soft-block spectral version of Eq. (1), and matching constants in Eq. (2) are stated as a model and sketch here and deserve full proofs.

8 Conclusion

The geometry in which to measure data redundancy is the learner's, but paying for it is only worthwhile when the input geometry is *misaligned with the task*. We gave a product characterization of that advantage — non-locality × non-interpolability × binding budget — validated it across 7 vision datasets and 2 backbones, and turned it into a cheap a-priori diagnostic that correctly predicts, on five LLM text tasks, which ones benefit. For resource-efficient learning the message is actionable: measure the input–learner gap first, and spend on expensive selection only when it is large.

References

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

[2] B. Mirzasoleiman, J. Bilmes, J. Leskovec. Coresets for data-efficient training of machine learning models (CRAIG). *ICML*, 2020.

[3] K. Killamsetty et al. GRAD-MATCH: gradient matching based data subset selection for efficient deep model training. *ICML*, 2021.

[4] A. Jacot, F. Gabriel, C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *NeurIPS*, 2018.

[5] G. Nemhauser, L. Wolsey, M. Fisher. An analysis of approximations for maximizing submodular set functions. *Math. Prog.*, 1978.

[6] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola. On kernel-target alignment. *NeurIPS*, 2002.

[7] P. Drineas, M. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *JMLR*, 2005.

[8] A. Ghorbani, J. Zou. Data Shapley: equitable valuation of data for machine learning. *ICML*, 2019.

[9] Qwen Team. Qwen2.5 technical report. 2024.

[10] E. Hu et al. LoRA: low-rank adaptation of large language models. *ICLR*, 2022.