GRaM Workshop at ICLR 2026

Intrinsic Dimension Dynamics in Active Learning: A Geometric Diagnostic of Acquisition Behavior

Poojith Thummala & Mohamed Abdelrazek

Applied Artificial Intelligence Initiative, Deakin University, Victoria, Australia

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Abstract

Active learning reduces annotation cost, but is typically evaluated only through predictive performance, offering limited insight into how different strategies shape the labelled set.

We analyze active learning through a geometric lens using intrinsic dimensionality (ID) as a simple, label-free diagnostic. Across strategies and datasets, we observe a consistent pattern: uncertainty-based methods select high-ID (complex) samples early, while coverage-based methods focus on lower-ID regions and expand more gradually.

Controlled experiments confirm that prioritizing low-ID samples early leads to more efficient learning. Overall, intrinsic dimensionality provides a simple and effective tool for understanding and diagnosing active learning behavior beyond accuracy.

Key finding: uncertainty-based strategies tend to select samples from higher estimated intrinsic-dimensional regions, while coverage-based strategies tend to yield labeled sets with lower and more stable estimated global intrinsic dimension.

1. Introduction

Active Learning (AL) reduces annotation cost by selecting informative samples, but is typically evaluated only through predictive performance.

Limitation: Accuracy alone does not explain how different strategies construct the labeled set or explore the data.

Geometric View: Active learning can be seen as progressive exploration of a data manifold, where each query shapes the structural complexity of the labeled set.

Key Gap: Most strategies are geometry-agnostic, uncertainty-based methods rely on model outputs and may concentrate sampling in localized, complex regions.

Contributions

1. **Intrinsic Dimension as a Diagnostic**: We introduce intrinsic dimensionality (ID) as a simple, label-free signal to analyze how active learning explores the data manifold.
2. **Geometric Bias of Acquisition Strategies:** We show that uncertainty-based methods prioritize high-complexity (high-ID) regions, while coverage-based methods focus on lower-dimensional regions and expand more gradually.
3. **Order Matters in Active Learning:** Through controlled experiments, we demonstrate that prioritizing low-ID samples early leads to significantly more efficient learning.

2. Intrinsic Dimensionality as Diagnostic

We focus on intrinsic dimensionality (ID) as a minimal, label-free, and locally computable geometric diagnostic. ID characterizes the effective number of degrees of freedom required to describe data in a representation space.

Local Intrinsic Dimensionality (Levina-Bickel MLE):

$$ID_l(x) = \left[(1/(k-1))\Sigma log\left(r_k(x)/r_j(x)\right)\right]^{-1}$$

r_j(x): distance to the j-th nearest neighbor
k: neighborhood size

Global Intrinsic Dimensionality:

$$\left[ID_G(S_t) = 1/|S_t|\Sigma ID_L^{(S_t)}(x)\right]$$

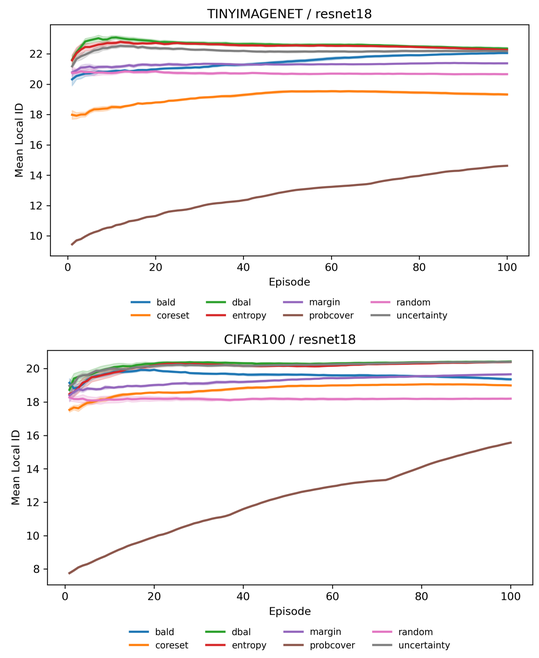
S_t: labeled set at round t
|S_t|: number of labeled samples

Higher LID = geometrically complex or ambiguous regions.

ID_G captures how structurally diverse the acquired set is across rounds. Tracks how the labeled set evolves during active learning

3. Results

Figure 1: ID Dynamics Induced by AL Strategies



Global Geometric Evolution:

Uncertainty-based strategies consistently induce higher ID_G, exhibiting rapid early growth. Coverage-based strategies maintain substantially lower and more stable global ID.

Local Geometric Selection:

Uncertainty-based strategies consistently query samples with high local ID. Coverage-based strategies initially select low-ID_L samples and only gradually expand.

Accuracy Results

Table 1: Test accuracy (%) at rounds 30, 60, and 100 using ResNet-18

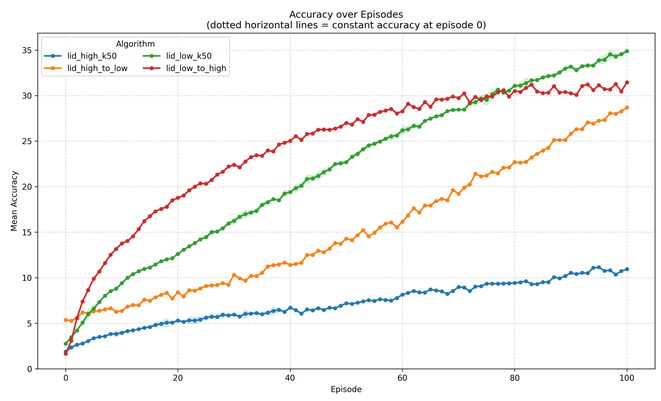
	CIFAR-100			TinyImageNet		
Method	**R30**	**R60**	**R100**	**R30**	**R60**	**R100**
ProbCover	**17.58**	**23.84**	**28.15**	**10.18**	**13.67**	**13.57**
CoreSet	13.94	18.25	23.26	5.66	7.49	8.77
Random	13.99	18.8	23.27	5.66	7.4	8.99
Margin	12.72	18.2	22.83	4.88	6.65	7.99
DBAL	12.14	17.5	21.92	4.48	5.44	6.87
Uncertainty	11.19	15.86	19.51	3.94	5.0	6.7

ProbCover tends to yield lower ID_G and achieves the highest test accuracies.

Uncertainty and DBAL produce higher-dimensional labeled sets and lower accuracies.

Controlled ID-based Schedules

Figure 2: Accuracy under ID-based acquisition schedules (k=50)



Controlled ID-based Schedules

We simulate acquisition by selecting samples based on intrinsic dimensionality.

- **Low → High ID:** Fast early learning, then saturation
- **High → Low ID:** Slow start, improves once low-ID samples are included
- **High-ID only:** Consistently poor performance

Key Result:
Prioritizing low-ID samples early is substantially more effective.

KEY INSIGHT

ID captures the temporal role of different geometric regions in learning. Low-ID regions correspond to structurally coherent areas valuable for early-stage learning; high-ID regions contribute primarily to later-stage refinement.

4. Discussion

The Geometric Dichotomy of Active Learning

Active learning strategies do not merely select samples with different predictive values; they induce measurably different geometric profiles in the labeled set.

Uncertainty-based

- Higher local ID (ID_L) (complex) regions
- Concentrates on geometrically complex regions
- Rapid ID_G growth
- May leave data support underexplored

Coverage-based

- Lower local ID (ID_L) (structured) regions first
- Anchors in denser, lower-dim regions first
- Stable ID_G trajectory
- More uniform exploration

Implications for Robustness:

Geometry-agnostic sampling may leave large portions of data support underexplored, potentially limiting generalization.

The more stable ID of coverage-based methods is consistent with more uniform exploration, which may contribute to robustness.

Practical Implications

Real-time diagnostic:
ID can monitor where labeling budget is being spent during the active learning loop.

Budget allocation:
Simple rules like deferring high-ID acquisitions until later rounds can improve efficiency.

Strategy design:
ID can inform new acquisition strategies as a regularizer that trades off uncertainty with geometric exploration.

Robustness axis:
Two strategies with similar accuracy may induce very different ID trajectories and geometric coverage.

Limitations & Future Work

- Analysis uses fixed self-supervised embeddings does not capture how geometric regimes interact with or distort representation space during end-to-end training.
- Quantifying downstream impact on model fragility from incomplete manifold coverage remains an open challenge.
- Future work: design generative evaluations capable of probing model behavior across the full data manifold to better assess geometric coverage and generalization robustness.

5. Conclusion

We studied active learning through a geometric lens, using ID as a simple diagnostic.

Across datasets and backbones, uncertainty-based methods query higher-ID regions, while coverage-based methods yield lower and more stable ID trajectories.

Controlled experiments show that prioritizing low-ID samples early leads to more efficient learning.

Active learning performance depends not only on *how many* samples are labeled, but *which geometric regions* are explored and *when*.

Contact

{s219110279, mohamed.abdelrazek}@deakin.edu.au

Key References

Levina & Bickel (2004) — MLE estimation of intrinsic dimension | Houle (2017) — Local ID foundations
Sener & Savarese (2018) — CoreSet | Yehuda et al. (2022) — ProbCover | Chen et al. (2020) — SimCLR
Ansuini et al. (2019) — ID in deep neural networks | Settles (2010) — Active learning survey