RETHINKING DATA SCALING THROUGH DATA GEOMETRY: A UNIFORMITY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling training data does not necessarily scale the amount of useful learning signal available to a model. We study this discrepancy through the geometry of training data and ask why carefully structured subsets can learn more efficiently than substantially larger datasets. Across controlled regression and LLM fine-tuning, subsets with more uniform geometry optimize faster than equal-size random subsets, and in several settings a half-size subset even outperforms the full dataset, revealing that scaling data can add redundancy faster than useful geometric coverage. We provide two complementary theoretical explanations for this phenomenon. From an *optimization* perspective, closely spaced examples can induce nearly dependent rows in the model Jacobian, leading to poorly conditioned data geometry and slower convergence under gradient descent. From an *approximation* perspective, data spacing and local simplex shape determine the conditioning of the interpolation geometry, with more uniform data simplices yielding tighter approximation bounds. Both views connect learning efficiency to data uniformity through the minimum pairwise separation $h_{\min}$ and related geometric quantities, with consistent empirical benefits across datasets, model sizes, and optimization settings. Together, these findings provide a uniformity-based explanation for why training on smaller, geometrically structured subsets can be more efficient than training on larger datasets.

1 INTRODUCTION

Modern machine learning is built on a powerful scaling intuition: more training data should provide more learning signal (Kaplan et al., 2020; Touvron et al., 2023). This intuition has driven the construction of increasingly large pre-training and fine-tuning corpora, together with growing interest in data selection and pruning methods that aim to reduce the computational cost of training (Xia et al., 2024; Albalak et al., 2024). Yet dataset size is only a count. It does not reveal how examples are arranged, how well they cover the relevant data space, or how much geometric redundancy they contain. As a result, adding more examples may not yield a proportional increase in the useful learning signal available to a model.

Our investigations reveal this mismatch between *data volume* and *learning efficiency*. Across controlled regression and large-language-model fine-tuning, subsets with more uniform geometry consistently optimize faster than equal-size random subsets. More strikingly, across several settings, including WizardLM (Xu et al., 2024), Tulu (Lambert et al., 2024), and data already selected by LESS (Xia et al., 2024), subsets containing roughly half of the available examples can match or outperform the full dataset while reaching target losses substantially earlier. These results suggest a simple but important possibility: *scaling data can increase redundancy faster than useful geometric coverage.*

This observation raises a fundamental question: **Why can a smaller but better-distributed dataset learn more efficiently than a larger one?** We study this question through the geometry of the training set, with particular focus on *data uniformity* as a concrete geometric property of how samples are arranged in the data space. Our main premise is that the effect of a training set depends not only on the number of samples it contains, but also on their spacing and local configuration. These geometric properties further influence both the optimization dynamics and the approximation error. We

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

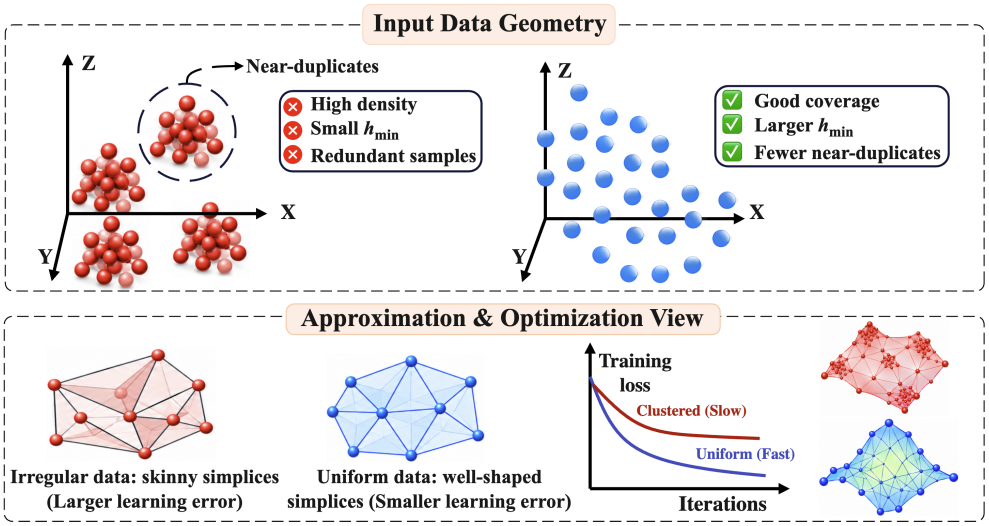


Figure 1: **Why data geometry matters beyond nominal data volume. Top:** Uniformly distributed data reduce near-collisions, increase the minimum pairwise separation $h_{\min}$, and provide broader coverage of the input space compared with clustered data. **Bottom:** These geometric differences affect learning through two complementary mechanisms. From the *approximation* view, irregularly distributed data can induce poorly shaped local simplices, whereas better-distributed data yield better-conditioned simplices and tighter approximation bounds through smaller $H_{\mathcal{T}}$ and larger $s_{\mathcal{T}}$. From the *optimization* view, closely spaced inputs can induce poorly conditioned model-induced geometry and a small $\mu_{\text{low}} = \lambda_{\min}(JJ^{\top})$, weakening guaranteed optimization progress; reducing such near-collisions can therefore facilitate faster convergence. Together, these two views explain how data geometry can improve learning efficiency even when nominal dataset size is unchanged.

formalize this connection through two complementary geometric perspectives, showing how data uniformity enters the geometry of optimization and approximation.

The first is an *optimization view from model-induced geometry*. For a dataset $X = (x_1, \ldots, x_N)$, consider the stacked Jacobian $J_\theta(X)$ of the model outputs with respect to the parameters. Its Gram matrix $J_\theta(X)J_\theta(X)^{\top}$ captures how training examples are arranged in the model-induced geometry. A key quantity is

$$\mu_{\text{low},\theta,X} = \lambda_{\min}\big(J_\theta(X)J_\theta(X)^{\top}\big), \tag{1}$$

which directly enters our convergence analysis of Gradient Descent (GD). When examples become very close, their corresponding Jacobian rows can become nearly dependent, driving $\mu_{\text{low},\theta,X}$ toward zero and slowing down the optimization progress, such $\mu_{\text{low},\theta,X}$ is indeed upper bounded by the level of data uniformity through the minimum pairwise distance $h_{\min}$.

The second is an *approximation view from local simplex geometry*. We triangulate the region covered by the training data into local simplices and study the singular values of their edge matrices. The resulting geometry is governed jointly by data spacing and simplex shape. In particular, we introduce the largest and smallest local stretching, $H_{\mathcal{T}}$ and $s_{\mathcal{T}}$, where better conditioned local simplices, roughly smaller $H_{\mathcal{T}}$ and larger $s_{\mathcal{T}}$, yield tighter approximation bounds. The $H_{\mathcal{T}}$ and $s_{\mathcal{T}}$ are also controlled by the minimum pairwise separation $h_{\min}$ and the local diameter $h_{\max_{d+1}}$ through $H_{\mathcal{T}} \le \sqrt{d}\, h_{\max_{d+1}}$, and $s_{\mathcal{T}} \ge \hat{s}_{\mathcal{T}} h_{\min}$, which similarly control the approximation error. This view highlights an important distinction: large separation alone is not sufficient, effective data geometry requires both good coverage and well-conditioned local structure.

Together, these two views demonstrate a common geometric principle. Increasing the amount of training data can improve coverage, but may simultaneously introduce greater local redundancy as pairwise sample distances become smaller. A more uniformly structured subset can preserve broad coverage while reducing such local crowding, improving both model-induced optimization geometry and local approximation geometry, and can accelerate training and improve approximation. This naturally leads to a *separation–coverage trade-off*: too few examples may leave the space insuf-

ficiently covered, whereas too many can reduce separation without providing commensurate new information. Figure 1 illustrates these two mechanisms.

We evaluate this perspective across diverse and comprehensive learning settings. Our experiments span California Housing, WizardLM (Xu et al., 2024), Tulu (Lambert et al., 2024), and LESS (Xia et al., 2024); squared-loss Stochastic Gradient Descent (SGD) and cross-entropy training with AdamW (Kingma & Ba, 2014; Zhou et al., 2024); and models ranging from LLaMA (Touvron et al., 2023) to Qwen3 (Yang et al., 2025). Across these settings, geometry-aware subsets consistently improve optimization relative to equal-size random subsets, and in several cases, a half-size subset exceeds the downstream performance of the full dataset while requiring substantially less training time. These results indicate that the observed effect is not specific to a single dataset, model, optimizer, or selection baseline.

Our focus is on understanding the role of data geometry in learning rather than introducing a new data-selection algorithm. To this end, we leverage a simple farthest-first traversal to construct subsets with more uniform geometry in a fixed representation space. We compare the resulting subsets with equal-size random subsets, varying the geometric configuration of the selected data while holding nominal dataset size fixed. This design enables us to examine differences in learning dynamics that cannot be explained by sample count alone.

Our contributions are threefold:

- **Data geometry beyond nominal data volume.** We show across controlled learning and LLM fine-tuning that equal-size datasets can exhibit substantially different optimization dynamics, and that smaller subsets with more uniform geometry can match or outperform substantially larger datasets. These results demonstrate that scaling data can increase redundancy faster than useful coverage.
- **A model-induced geometric view of optimization.** We connect local data spacing to the smallest eigenvalue $\mu_{\text{low}} = \lambda_{\min}(JJ^\top)$ that governs the convergence analysis. Near-collisions in input space, i.e., small minimum pairwise distance $h_{\min}$, can induce nearly dependent Jacobian rows and force this quantity to become small, providing a concrete mechanism through which clustered data can lead to slower optimization.
- **A local geometric view of approximation.** We characterize local interpolation geometry through simplex conditioning and derive approximation bounds controlled by local maximal and minimal streching $H_{\mathcal{T}}$ and $s_{\mathcal{T}}$. Relating these quantities to data separation by $h_{\min}$, local coverage through $h_{\max_{d+1}}$, and simplex shape explains why a smaller but better-structured dataset can retain strong approximation quality.

Taken together, our results suggest a geometric view of data scaling:

> *Effective data scaling is not only about increasing data volume, but about expanding useful coverage while avoiding excessive redundancy and poorly conditioned data geometry.*

Due to space limits, we provide more related work in Appendix A.

2 PRELIMINARIES

In this section, we consider the data $\{(x_i, y_i)\}_{i=1}^N$ where input $x_i \in \mathbb{R}^d$, output $y_i \in \mathbb{R}^d$, and define

$$h_{ij} = \|x_i - x_j\|, \text{ and } h_{\min} = \min_{1 \le i < j \le N} h_{ij}.$$

We assume that the input data are not necessarily independently sampled as in Wang (2026).

Assumption 1 (Data)**.** *Let $x_1, \cdots, x_N \in \mathbb{R}^d$ be input data, and $y_1, \cdots, y_N \in \mathbb{R}^d$ be the corresponding outputs. Assume*

$$(x_1, \cdots, x_N) \sim \pi(z_1, \cdots, z_N) dz_1 \cdots dz_N,$$

where $\pi(z_1, \cdots, z_N) \in L^\infty$ is supported on an open set $\Omega \times \cdots \times \Omega \subseteq \mathbb{R}^{Nd}$ with $|\Omega| < \infty$, and $\pi \ll \text{Leb}_{Nd}$.

This assumption allows dependencies among data points, which is compatible with the data selection and preprocessing used in our experiments.

We consider a family of neural networks as follows:

$$\begin{aligned} u_{0,i} &= x_i; \\ u_{\ell+1,i} &= u_{\ell,i} + \varphi_\ell(\theta_\ell; u_{\ell,i}), \forall \ell = 0, \cdots, L-1; \\ f(\theta; x_i) &= \varphi_L(\theta_L; u_{L,i}) \end{aligned} \tag{2}$$

where $\theta = (\theta_0, \cdots, \theta_L)$ is the collection of weights with θ_i to be the weight of the ith layer, $\varphi_\ell : \mathbb{R}^{\dim\theta_\ell} \times \mathbb{R}^d \to \mathbb{R}^d$ is the block of the ℓth layer for $0 \le \ell \le L$, $u_{\ell,i} \in \mathbb{R}^d$ is the output of the ℓth layer corresponding to the input x_i.

Next, we introduce some quantities that characterize the geometry of the data and relates to data uniformity.

Model-induced data geometry. We define for a fixed parameter θ and dataset $X = (x_1, \cdots, x_N)$, the minimum eigenvalue of the matrix

$$\mu_{\text{low},\theta,X} := \lambda_{\min}\left[\begin{pmatrix} \nabla_\theta f(\theta; x_1) \\ \vdots \\ \nabla_\theta f(\theta; x_N) \end{pmatrix}\begin{pmatrix} \nabla_\theta f(\theta; x_1) \\ \vdots \\ \nabla_\theta f(\theta; x_N) \end{pmatrix}^\top\right].$$

This quantity $\mu_{\text{low},\theta,X}$ characterizes the geometric arrangement of the parameter gradients of the neural network's output values at the data points. A small value means that some combinations of these gradients nearly cancel, which can produce weak training gradients and slow optimization.

Local simplex geometry. We first introduce the simplex notations used to describe the local geometry of the interpolation region. A d-simplex is the convex hull of $d+1$ affinely independent points in $\mathbb{R}^d$. Let $\mathcal{T}$ be a finite triangulation of $\mathcal{I} = \text{conv}\{x_1, \ldots, x_N\}$ into nondegenerate d-simplices. Each simplex $\mathcal{I}_i$ represents a local interpolation region whose geometry is encoded by its edge vectors

$$\mathcal{I}_i = \text{conv}\{x_{i,0}, x_{i,1}, \ldots, x_{i,d}\}, \quad \text{with edges } \delta x_{i,j} = x_{i,j} - x_{i,0}, \quad \text{for } 1 \le j \le d,$$

where $\{x_{i,0}, x_{i,1}, \ldots, x_{i,d}\} \subseteq \{x_1, \ldots, x_N\}$. The edge matrix based at $x_{i,0}$ is

$$\Delta x_i = (\delta x_{i,1}, \delta x_{i,2}, \ldots, \delta x_{i,d}).$$

Its largest singular value $\sigma_{\max}(\Delta x_i)$ measures the maximal stretching of each simplex, whereas $\sigma_{\min}(\Delta x_i)$ measures its minimal stretching. We aggregate these local quantities over the triangulation by defining

$$H_{\mathcal{T}} = \max_{\mathcal{I}_i \in \mathcal{T}} \sigma_{\max}(\Delta x_i), \quad \text{and} \quad s_{\mathcal{T}} = \min_{\mathcal{I}_i \in \mathcal{T}} \sigma_{\min}(\Delta x_i).$$

Here $H_{\mathcal{T}}$ and $s_{\mathcal{T}}$ describes the largest and the smallest local scales respectively. We will use these two terms to characterize the approximation error.

Notations. We use the Sobolev space $W^{m,p}(\Omega)$ (Adams & Fournier, 2003; Evans, 2022), which is defined as $W^{m,p}(\Omega) = \{\phi \in L^p(\Omega) : D^\alpha \phi \in L^p(\Omega), \forall |\alpha| \le k\}$, with Sobolev norm $\|\phi\|_{m,p} = \left(\sum_{|\alpha|\le m} \|D^\alpha \phi\|_p^p\right)^{1/p}$ for $1 \le p < \infty$, and $\max_{|\alpha|\le m} \|D^\alpha \phi\|_\infty$ for $p = \infty$ where $D^\alpha \phi = \frac{\partial^{|\alpha|}\phi}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d}$. We use $\|\cdot\|$ to denote ℓ^2 norm for vectors and matrices, use $\|\cdot\|_p$ to denote ℓ^p norm for vectors and $L^p(\Omega)$ norm for functions, use $\|\cdot\|_{r,p,\Omega}$ to denote the Sobolev norm of $W^{r,p}(\Omega)$. For the function $f : \mathbb{R}^m \to \mathbb{R}^n$, we use $\nabla f \in \mathbb{R}^{n\times m}$ to denote the Jacobian of the map f w.r.t. all variables, and use $\nabla_x f$ to represent the Jacobian w.r.t. x. For a vector α, we denote $|\alpha| = |(\alpha_1, \cdots, \alpha_d)| = \sum_{i=1}^d \alpha_i$; for a set Ω, we use $|\Omega|$ to denote the volume of the set. The closure and interior of a set Ω are denoted as $\bar{\Omega}$ and Ω^o. For two measures ν and μ, we use $\nu \ll \mu$ to denote the absolute continuity of ν w.r.t. μ. We denote $\text{Leb}_d(\cdot)$ to be the Lebesgue measure of a set in $\mathbb{R}^d$. We use $\text{conv}\{v_1, \cdots, v_n\}$ to denote the convex hull over $v_1, \cdots, v_n$.

3 Theory: two geometric views of data through data uniformity

In this section, we introduce the model-induced data geometry, measured by $\mu_{\text{low},\theta,X}$, and the local simplex geometry, measured by $H_{\mathcal{T}}$ and $s_{\mathcal{T}}$, to data uniformity through the minimum pairwise distance $h_{\min}$. Futhermore, these relationships explain how a larger $h_{\min}$ can lead to faster training and smaller approximation error. The theoretical results also motivate a geometry-guided criterion for uniform data selection.

3.1 Model-induced data geometry: optimization

We first study the optimization perspective: the model transforms input data into training directions whose geometric arrangement affects the optimization dynamics; hence, closely spaced data can slow convergence while greater separation, particularly a larger $h_{\min}$, can facilitate training.

We define the ℓ^2 loss $\mathcal{L}(\theta; \{x_i, y_i\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^N l(f(x_i), y_i)$, where $l(\theta; x_i, y_i) = \frac{1}{2}\|y_i - f(\theta; x_i)\|^2$. We consider minimizing $\mathcal{L}(\theta)$ by gradient descent $\theta^{k+1} = \theta^k - \eta\nabla_\theta\mathcal{L}(\theta^k)$, where $\eta > 0$, and θ^k is the vectorization of all parameters θ at the kth iteration.

Along the GD trajectory, write $\mu_{\text{low},s,X} := \mu_{\text{low},\theta^s,X}$. Larger values of $\mu_{\text{low},s,X}$ give a stronger guaranteed decrease in the loss at each iteration, as stated in the following convergence result.

Theorem 1 (Wang (2026))**.** *Suppose Assumption 1 and the model regularity, nondegeneracy, and local dissipativity conditions in Wang (2026) hold with some $\mathsf{L} > 0$. Then, for almost every initialization and with probability 1 over the joint distribution of the input data $x_1, \cdots, x_N$, for almost all learning rate $\eta < \frac{2}{\mathsf{L}}$, GD converges to a region $\|\nabla\mathcal{L}(\theta)\| \le \epsilon_{\mathcal{L}}$ and, for all $k \ge 1$ before entering this region,*

$$\mathcal{L}(\theta^k) \le \prod_{s=0}^{k-1}\left(1 - \eta\left(1 - \frac{\eta\mathsf{L}}{2}\right)\frac{2\mu_{\text{low},s,X}}{N}\right)\mathcal{L}(\theta^0).$$

We then derive the following theorem, which highlights the relationship between data pairwise distance, particularly $h_{\min}$, and the $\mu_{\text{low},s,X}$ that controls the convergence rate.

Theorem 2. *Under the same assumptions as Theorem 1, fix some x_i. Let $X = (x_1, \cdots, x_N)$, and $\mathcal{D}_{i,H} = \{j \ne i \,|\, h_{ij} = \|x_i - x_j\| \le H\}$. Assume $\mathcal{D}_{i,H} \ne \emptyset$. Then for any $k \ge 1$, there exists $r_{k,i,H} > 0$ and $L_{k,i,H} > 0$, s.t., when $\sqrt{\sum_{j\in\mathcal{D}_{i,H}} h_{ij}^2} \le r_{k,i,H}$, we have*

$$\mu_{\text{low,s,X}} \le L_{k,i,H}\sqrt{\sum_{j\in\mathcal{D}_{i,H}} h_{ij}^2},\ \forall s \le k.$$

Specifically, if $H = h_{\min}$ and $\mathcal{D}_{i,H} \ne \emptyset$, there exists $L_{k,i}, r_{k,i} > 0$, s.t., when $h_{\min} \le r_{k,i}$, we have

$$\mu_{\text{low,s,X}} \le L_{k,i}h_{\min},\ \forall s \le k.$$

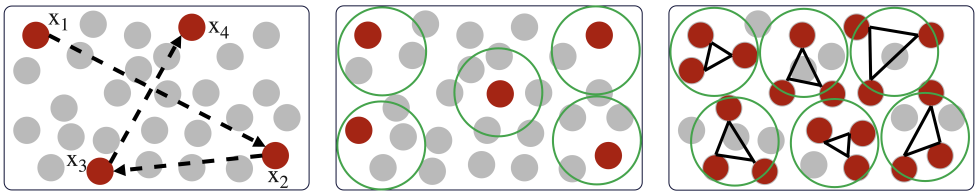
The above theorem shows that, for a fixed data point x_i, when nearby points become even closer to x_i, the training process potentially slows down, which is verified experimentally in Section 4. More precisely, there exists a subsequence of $\{\mu_{\text{low,s,X}_t}\}_t$ depending on the data X_t, s.t., as the nearby distance $\sqrt{\sum_{j\in\mathcal{D}_{i,H}} h_{ij}^2}$ decreases for some X_t, the value of $\mu_{\text{low,s,X}_t}$ also monotonically decreases. Consequently, the value of $1 - \eta\left(1 - \frac{\eta\mathsf{L}}{2}\right)\frac{2\mu_{\text{low},s,X_t}}{N}$ increases, and thus leads to larger rate bound. The dependence on $h_{\min}$ is a special case of this phenomenon and follows the same reasoning. The proof mainly relies on the local Lipschitz continuity of eigenvalues for real-analytic matrix functions (see Theorem 4.1 in Kurdyka & Paunescu (2008)); the convergence ingredients are from Wang (2026). See Appendix B for the proof.

We remark that the theorem provides an upper bound on $\mu_{\text{low},s,X}$. Thus, increasing $h_{\min}$ is motivated as a way to mitigate local crowding, but does not by itself guarantee faster convergence. Even datasets with the same $h_{\min}$ can induce different geometries and hence different convergence behavior. Moreover, increasing separation by selecting fewer samples may weaken local coverage. These considerations motivate the separation–coverage tradeoff studied in the following sections.

3.2 Local simplex geometry: approximation

Next, we establish the approximation perspective by triangulating the interpolation region formed by the input data. Each simplex provides a local coordinate map whose singular values describe its scale and shape. Relating these singular values to the minimum data separation $h_{\min}$ and the largest local diameter $h_{\max_{d+1}}$ allows us to identify how data spacing enters the approximation error.

We first present how data uniformity characterized by $h_{\min}$, together with the largest local diameter $h_{\max_{d+1}}$, controls the local simplex geometry in the following lemma.



(a) Farthest-first selection (b) Insufficient local coverage (c) Sufficient local coverage

Figure 2: **Geometry-aware stopping criterion for uniform data selection. (a) Farthest-first selection.** Starting from an initial point, the algorithm iteratively adds the point farthest from the current selected set, promoting a larger minimum separation $h_{\min}$ and reducing near-collisions. **(b) Insufficient local coverage.** When the selected set is still too sparse, some $x \in S$ do not have enough neighboring selected points within the local ball $B_x(r)$ to form a well-covered local neighborhood. **(c) Sufficient local coverage.** Selection continues until every selected point has at least $\hat{d}$ neighboring selected points within radius r, equivalently $r_{\hat{d}}(S) = \max_{x \in S} d_{\hat{d}}(x; S) \le r$. This stopping rule balances pairwise separation with the local coverage scale associated with $h_{\max_{\hat{d}+1}}$, yielding a geometry-motivated stopping principle rather than a fixed data-retention ratio.

Lemma 1. *Suppose Assumption 1 holds. With probability 1 over the joint input data distribution, the data admit a finite nondegenerate triangulation $\mathcal{T}$. For every such triangulation,*

$$H_{\mathcal{T}} \le \sqrt{d}\, h_{\max_{d+1}}, \qquad s_{\mathcal{T}} \ge \hat{s}_{\mathcal{T}} h_{\min}, \tag{3}$$

where $h_{\max_{d+1}} = \max_{\mathcal{I}_i \in \mathcal{T}} \max_{0 \le q < j \le d} \|x_{i,q} - x_{i,j}\|$, $\hat{s}_{\mathcal{T}} = \min_{\mathcal{I}_i \in \mathcal{T}} \sigma_{\min}(\widehat{\Delta x}_i) > 0$, *and* $\widehat{\Delta x}_i = \left(\frac{\delta x_{i,1}}{\|\delta x_{i,1}\|}, \frac{\delta x_{i,2}}{\|\delta x_{i,2}\|}, \ldots, \frac{\delta x_{i,d}}{\|\delta x_{i,d}\|} \right)$

Lemma 1 relates the minimum pairwise distance $h_{\min}$ to the smallest local singular value $s_{\mathcal{T}}$ through the shape factor $\hat{s}_{\mathcal{T}}$. However, a lower bound on $h_{\min}$ alone does not prevent nearly flat simplices. The ratio $h_{\max_{d+1}}/h_{\min}$ bounds the variation in edge lengths, while $\hat{s}_{\mathcal{T}}$ measures the angles between the normalized edge directions. Together, they bound the condition number of every simplex edge matrix: $\kappa_2(\Delta x_i) := \sigma_{\max}(\Delta x_i)/\sigma_{\min}(\Delta x_i) \le \sqrt{d}\, h_{\max_{d+1}}/(\hat{s}_{\mathcal{T}} h_{\min})$ for all $\mathcal{I}_i \in \mathcal{T}$.

We next show that a larger $h_{\min}$ and smaller local maximal distances $h_{\max_{d+1}}$ can lead to smaller approximation error in the interpolation regime $\mathcal{I} = \operatorname{conv}\{x_1, \cdots, x_N\}$.

Theorem 3. *Suppose Assumption 1 holds with $N \ge d+1$. Under some regularity conditions of ground truth g and network f, there exist a function $\phi \in W^{m,p}(\mathcal{I}^\circ; \mathbb{R}^d)$ and a finite-width analytic network $f : \mathbb{R}^d \to \mathbb{R}^d$ with probability 1 such that*

$$\|f - g\|_{p,\mathcal{I}^\circ} \le 3C \max\{H_{\mathcal{T}}, 1\}^m \min\{s_{\mathcal{T}}, 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ}. \tag{4}$$

Moreover, we have

$$\|f - g\|_{p,\mathcal{I}^\circ} \le 3C_{\mathcal{T}} \max\{h_{\max_{d+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ}, \tag{5}$$

where $C_{\mathcal{T}} > 0$ depends on m, p, d and the finite triangulation $\mathcal{T}$.

The above theorem examines how local geometry and distances between data points affect approximation error. More precisely, the error bound is determined directly by the largest local stretching $H_{\mathcal{T}}$ and the smallest local stretching $s_{\mathcal{T}}$: a smaller $H_{\mathcal{T}}$ and a larger $s_{\mathcal{T}}$ tighten the bound, showing that better-conditioned simplices can yield tighter approximation error. The bound can further be interpreted in terms of pairwise distances by Lemma 1: using fewer but more uniformly distributed data points, i.e., increasing $h_{\min}$[1], can achieve similar approximation error as increasing the number of data points, i.e., decreasing $h_{\max_{d+1}}$. This is also validated in our experiments (see Section 4). We remark that simplex shape, characterized by $\hat{s}_{\mathcal{T}}$, also affects the approximation bound. Thus, data selection methods yielding similar values of $h_{\min}$ and $h_{\max_{d+1}}$ may still produce different bounds. Moreover, the selected datasets may define different interpolation regions. The proof mainly involves establishing a data-dependent Bramble-Hilbert lemma (Bramble & Hilbert, 1970). See Theorem 4 for the formal version and Appendix C for the proofs.

[1]Larger $h_{\min}$ also contributes to smoother approximation as r specifies the order of smoothness in $W^{r,p}(\Omega)$.

3.3 Practical Implications for Geometry-Guided Uniform Data Selection

Our theoretical analysis motivates selecting subsets that balance sample separation and local coverage. We translate this insight into a simple geometric selection procedure. Let $\mathcal{D}$ denote the candidate dataset, $\phi(x)$ a fixed feature representation of sample x, and $d(\cdot,\cdot)$ the distance measure in the representation space. Starting from an initial sample $S = \{x_1\}$, we construct a *uniform* subset through farthest-first selection:

$$x^\star = \arg\max_{x\in\mathcal{D}\setminus S} \min_{x'\in S} d(\phi(x),\phi(x')) \,. \tag{6}$$

At each iteration, we update $S \leftarrow S \cup \{x^\star\}$. This procedure prioritizes samples farthest from their nearest selected neighbors, promoting separation and reducing local redundancy.

The remaining question is when to stop adding samples. Theorem 3 motivates a *separation–coverage trade-off*: a smaller subset can improve separation, but excessive pruning may weaken local coverage and increase $h_{\max_{d+1}}$. To operationalize this trade-off, let $\hat{d}$ be a positive integer estimated from the local intrinsic dimension, and let $d_{\hat{d}}(x;S)$ denote the distance from $x \in S$ to its $\hat{d}$-th nearest neighbor in $S\setminus\{x\}$, measured in the same representation space. We define

$$r_{\hat{d}}(S) = \max_{x\in S} d_{\hat{d}}(x;S), \tag{7}$$

with $r_{\hat{d}}(S) = +\infty$ when $|S| \le \hat{d}$, and stop when $r_{\hat{d}}(S) \le r$, where $r > 0$ is a prescribed neighborhood radius. Equivalently, every selected point must have at least $\hat{d}$ other selected points within distance r, as illustrated in Figure 2. This criterion checks local neighborhood sufficiency, complementing the separation promoted by farthest-first selection.

Motivated by Theorem 3 and the optimization analysis in Theorem 2, this procedure determines subset size through a geometric criterion rather than a predefined retention ratio. It remains a geometry-motivated stopping principle, not a guarantee of the downstream-performance-optimal subset size.

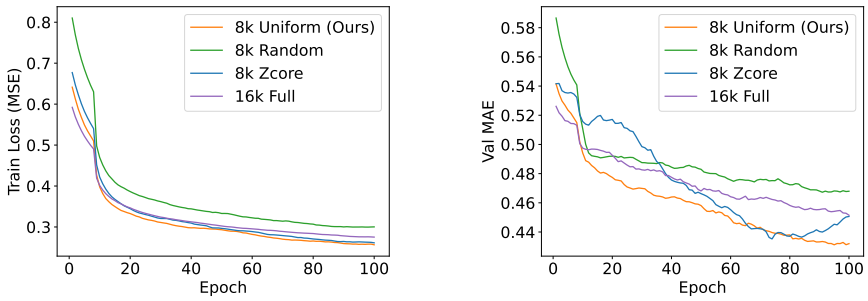
4 Experiments

4.1 Experimental Setup

We conduct comprehensive experiments across diverse settings, including California Housing (Nugent, 2017; Pace & Barry, 1997), WizardLM (Xu et al., 2024), Tulu (Lambert et al., 2024), and LESS-selected data (Xia et al., 2024). Our evaluations span LLaMA-1-7B/13B (Touvron et al., 2023) and Qwen3-1.7B (Yang et al., 2025). Motivated by the geometric analysis in Section 3, we construct uniform subsets using fixed representations, with Word2Vec embeddings for instruction-tuning data. We compare equal-size uniform and random subsets against their corresponding full datasets: 8k versus 16k on California Housing, 10k versus 20k on WizardLM, and 7k versus 14k on LESS. We also compare against strong data-selection baselines, including Zcore (Griffin et al., 2024), an embedding-based coreset-selection method for constructing compact and representative subsets, and LESS (Xia et al., 2024), a model-aware approach that selects instruction-tuning examples using gradient-based importance scores. These controlled comparisons isolate the effect of data geometry at a fixed subset size and assess whether geometric structure provides complementary benefits beyond random sampling, full-data training, and existing data-selection criteria, indicating that geometric structure captures an important aspect of data quality not fully reflected by conventional selection objectives. The selection procedure is computationally lightweight: all subsets used in our experiments can be constructed in less than one minute on a single H100 GPU. Moreover, Additional training configurations is available in Appendix E.

4.2 ℓ^2-SGD Training Experiments

To validate our theoretical insights, we follow the ℓ^2-loss SGD setting on the California Housing dataset (Nugent, 2017; Pace & Barry, 1997) and compare our uniform subset with an equal-size random subset, the full dataset, and Zcore (Griffin et al., 2024), a strong coreset-selection baseline that uses foundation-model embeddings to identify compact and representative subsets. As shown in Figure 3, our uniform subset converges faster and achieves the lowest training MSE and validation MAE among all compared methods, outperforming all baselines despite using only half as much data as the full dataset. These results support our theoretical points that improved data geometry can



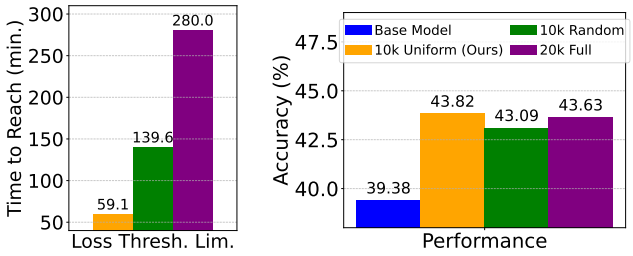
(a) Training loss (lower is better) (b) Validation loss (lower is better)

Figure 3: **Training and validation performance on California Housing with squared-loss SGD.** We compare an 8k uniform subset with an equal-size 8k random subset, an 8k Zcore (Griffin et al., 2024) subset, and the full 16k dataset. **(a)** Training MSE, where the uniform subset converges faster and reaches the lowest final training loss. **(b)** Validation MAE, where the uniform subset achieves the best overall validation performance while using only half of the full dataset. These results show that data geometry can substantially affect both optimization efficiency and generalization.

enhance both optimization efficiency and generalization. Consistent results across other datasets, model sizes, and baselines such as LESS and WizardLM confirm the efficiency of uniformity-based selection, supporting theory (Section 3) and demonstrating improved efficiency and performance.

4.3 Cross Entropy Loss-Adam Training Experiments

Beyond ℓ^2-SGD, we also conduct training experiments using cross-entropy loss with Adam (Kingma & Ba, 2014; Zhou et al., 2024), the standard optimization setup for LLMs (Gunel et al., 2021; Mao et al., 2023). Across WizardLM, Tulu, and LESS-selected data, using both LLaMA (Touvron et al., 2023) and Qwen (Yang et al., 2025) models, more geometrically uniform subsets consistently train faster than equal-size random subsets and often match or exceed the performance of the full dataset while using only half of the training data. The same pattern holds when uniform selection is applied on top of LESS (Xia et al., 2024), a strong model-aware data-selection method based on gradient-derived importance scores, suggesting that geometric structure is a key factor in training efficiency that is not fully captured by model-based importance scores.

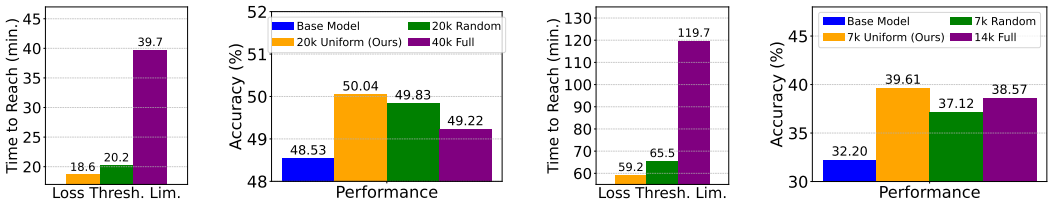


(a) Efficiency comparison (lower is better). (b) Performance comparison (higher is better).

Figure 4: **Training efficiency and downstream performance on WizardLM with LLaMA-1-13B.** Compared with the 10k random subset and full 20k dataset, the 10k uniform subset **(a)** reaches the target loss of 0.6 fastest and **(b)** achieves better downstream performance at matched-loss checkpoints while using only half of the training data.

Evaluation on WizardLM datasets (Xu et al., 2024). We evaluate our method on WizardLM using LLaMA-1-13B under standard fine-tuning settings, and observe consistent improvements in both training efficiency and downstream performance. As shown in Figure 4(a), the 10k uniform subset reaches the target loss substantially faster than both the equal-size 10k random subset and the full 20k dataset. At matched-loss checkpoints, Figure 4(b) shows that the uniform subset also achieves stronger performance on ARC (Clark et al., 2018) and TruthfulQA (Lin et al., 2022), despite using only half of the full training data.

Evaluation on Tulu-3 Datasets (Lambert et al., 2024). We further evaluate our method on Tulu-3 using Qwen3-1.7B-Base. As shown in Figures 5 (a) and (b), the 20k uniform subset reaches the target loss in 18.6 minutes, compared with 20.2 minutes for the equal-size random subset and 39.7 minutes for the full 40k dataset, corresponding to a 2.13× speedup over full-data training. At matched-loss checkpoints, the uniform subset also achieves the highest overall downstream accuracy



(a) Efficiency comparison (lower is better).

(b) Performance comparison (higher is better).

(c) Efficiency comparison (lower is better).

(d) Performance comparison (higher is better).

Figure 5: **(a) and (b)Training efficiency and downstream performance on Tulu-3 (Lambert et al., 2024) with Qwen3-1.7B-Base (Yang et al., 2025).** Compared with an equal-size random subset and the full dataset, the 20k uniform subset **(a)** reaches the target loss fastest, with a 2.13× speedup over full-data training, and **(b)** achieves the highest average downstream accuracy at matched-loss checkpoints while using only half of the training data. **(c) and (d) Training efficiency and downstream performance on LESS-selected data with LLaMA-1-7B (Touvron et al., 2023; Xia et al., 2024).** Compared with an equal-size 7k random subset and the full 14k dataset, the 7k uniform subset **(c)** reaches the target loss of 0.3 fastest and **(d)** achieves the highest downstream accuracy at matched-loss checkpoints while using only half of the training data.

across ARC Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), TruthfulQA (Lin et al., 2022), and IFEval (Zhou et al., 2023). Thus, using only half of the full training data, uniform selection improves both optimization efficiency and downstream performance, further supporting the benefit of data geometry over simply increasing nominal data volume.

Data Geometry Further Improves Model-Aware Selection. We examine whether data uniformity remains beneficial after strong model-aware selection. LESS (Xia et al., 2024) is a strong baseline that selects instruction-tuning examples using gradient-based importance scores. Starting from its 14k selected set, we construct a 7k uniform subset and compare it with an equal-size 7k random subset and the full 14k set. As shown in Figures 5 (c) and (d), the 7k uniform subset reaches the target loss substantially faster than both baselines and achieves the highest MMLU (Hendrycks et al., 2020; Wang et al., 2024b) accuracy, despite using only half of the full dataset. These results show that data geometry provides a complementary benefit beyond importance-based selection: even high-value examples can retain geometric redundancy and uneven coverage, which uniform refinement can further reduce.

Taken together, these controlled comparisons show that nominal data volume alone is insufficient to explain learning efficiency. Across WizardLM, Tulu-3, and LESS-selected data, more geometrically uniform subsets consistently train faster than equal-size random subsets and can match or outperform substantially larger datasets. The consistency of these results across different data sources, model families, and selection paradigms further indicates that data geometry captures a fundamental aspect of training efficiency beyond sample count and model-based importance. Additional results are provided in Appendix D.

5 Conclusion

In this work, we study data scaling through the lens of data geometry, focusing on data uniformity as a concrete characterization of the geometric configuration of training samples. We provide two complementary explanations for why geometry matters: from the optimization perspective, closely spaced samples can induce nearly dependent rows in the model Jacobian, deteriorating the conditioning of the underlying data geometry and slowing convergence; from the approximation perspective, data spacing and local simplex shape jointly determine the conditioning of the interpolation geometry, with better-conditioned simplices yielding tighter approximation bounds. Empirically, across our comprehensive experiments, model families, and data-selection settings, subsets with more uniform geometry consistently train faster than both equal-size random subsets and the full dataset. They also achieve better downstream performance than equal-size random subsets, full-data training, and strong baselines, despite using only half of the training data. Taken together, our findings suggest that *the effectiveness of data scaling is governed not only by sample count alone, but also by how additional data reshape the geometry of learning.*

AI USE STATEMENT

We used generative AI tools to assist with language editing and parts of code. The authors reviewed and revised the AI-assisted material, checked technical statements, and verified that the illustrations accurately represented the intended concepts. The authors take full responsibility for the final content of this work, including all text, theoretical claims, experimental results, citations, and figures.

ETHICS STATEMENT

This work studies how data geometry and uniform selection affect learning efficiency. This paper presents work whose goal is to advance machine learning, and to provide empirical insights. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

Section 3 presents the theoretical setting and assumptions, and Section 3.3 specifies the farthest-first selection procedure and geometry-aware stopping criterion. Section 4.1 describes the datasets, models, controlled comparisons, and evaluation protocol, with additional training configurations provided in the appendix. Appendix D.2 describes the PCA visualization procedure, while Appendix D.3 details the geometric measurements and empirical calibration of the stopping radius, etc. These descriptions support reproduction of the work. For more details, please see the attached code.

REFERENCES

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=XfHWcNTSHp. Survey Certification, Featured Certification.

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International conference on machine learning*, pp. 1908–1916. PMLR, 2014.

James H Bramble and SR Hilbert. Estimation of linear functionals on sobolev spaces with application to fourier transforms and spline interpolation. *SIAM Journal on Numerical Analysis*, 7(1): 112–124, 1970.

Miaomiao Cai, Min Hou, Lei Chen, Le Wu, Haoyue Bai, Yong Li, and Meng Wang. Mitigating recommendation biases via group-alignment and global-uniformity in representation learning. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–27, 2024.

Yongqiang Cai. Achieve the minimum width of neural networks for universal approximation. *arXiv preprint arXiv:2209.11395*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mark De Berg. *Computational geometry: algorithms and applications*. Springer Science & Business Media, 2000.

Jesús A. De Loera, Jörg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*, volume 25 of *Algorithms and Computation in Mathematics*. Springer, 2010. doi: 10.1007/978-3-642-12971-1.

Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:202120592.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *Ieee Access*, 7: 64323–64350, 2019.

Brent A Griffin, Jacob Marks, and Jason J Corso. Zero-shot coreset selection: Efficient pruning for unlabeled data. *arXiv preprint arXiv:2411.15349*, 2024.

Shangding Gu, Alois Knoll, and Ming Jin. Teams-rl: Teaching llms to generate better instruction datasets via reinforcement learning. *Transactions on Machine Learning Research*, 2024.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep relu neural networks in w s, p norms. *Analysis and Applications*, 18(05):803–859, 2020.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=cu7IUiOhujH.

Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental*, 4:1–13, 2020.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *arXiv preprint arXiv:2307.14023*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on learning theory*, pp. 2306–2327. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. *arXiv preprint arXiv:2110.03303*, 2021.

Krzysztof Kurdyka and Laurentiu Paunescu. Hyperbolic polynomials and multiparameter real-analytic perturbation theory. 2008.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Shengjie Luo, Shanda Li, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. Your transformer may not be as powerful as you expect. *Advances in Neural Information Processing Systems*, 35: 4301–4315, 2022.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pp. 23803–23828. PMLR, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. *Advances in neural information processing systems*, 36: 61115–61157, 2023.

Cam Nugent. California housing prices. Kaggle, 2017. URL `https://www.kaggle.com/datasets/camnugent/california-housing-prices`. Accessed: 2025-04-01.

Zhongyu Ouyang, Chunhui Zhang, Shifu Hou, Chuxu Zhang, and Yanfang Ye. How to improve representation alignment and uniformity in graph-based collaborative filtering? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 1148–1159, 2024.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

Josiah Park and Stephan Wojtowytsch. Qualitative neural network approximation over r and c: Elementary proofs for analytic and polynomial activation. *arXiv preprint arXiv:2203.13410*, 2022.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. *arXiv preprint arXiv:2006.08859*, 2020.

Lei Sang, Yu Zhang, Yi Zhang, Honghao Li, and Yiwen Zhang. Towards similar alignment and unique uniformity in collaborative filtering. *Expert Systems with Applications*, 259:125346, 2025.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Approximation with cnns in sobolev space: with applications to classification. *Advances in neural information processing systems*, 35: 2876–2888, 2022.

Ha Min Son, Moon-Hyun Kim, Tai-Myoung Chung, Chao Huang, and Xin Liu. Feduv: uniformity and variance for heterogeneous federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5863–5872, 2024.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2022.

Qitao Tan, Xiaoying Song, Guanghui Ye, and Chuan Wu. An effective negative sampling approach for contrastive learning of sentence embedding. *Machine Learning*, 112(12):4837–4861, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1816–1825, 2022.

Liang Wang, Xiang Tao, Qiang Liu, and Shu Wu. Rethinking graph masked autoencoders through alignment and uniformity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15528–15536, 2024a.

Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20050–20060, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024b.

Yuqing Wang. Convergence of gradient descent for general neural network architectures beyond the ntk regime. *arXiv preprint arXiv:2606.23364*, 2026.

David S Watkins. A generalization of the bramble-hilbert lemma and applications to multivariate interpolation. *Journal of approximation theory*, 26(3):219–231, 1979.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yu Yu, Shahram Khadivi, and Jia Xu. Can data diversity enhance learning generalization? In *Proceedings of the 29th international conference on computational linguistics*, pp. 4933–4945, 2022.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL `https://arxiv.org/abs/2311.07911`.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6486–6493, 2024.

Appendix

A Related Work

Data Geometry and Selection. Data selection seeks compact training subsets that preserve useful information, with existing approaches drawing on geometric coverage, redundancy, and model-dependent influence. Sener & Savarese (2018) relate core-set selection to data geometry, while ZCore (Griffin et al., 2024) evaluates coverage and redundancy in foundation-model embedding spaces. LESS (Xia et al., 2024) instead estimates data influence through optimizer-aware gradient similarity for targeted instruction tuning. Related work has investigated data composition through the broader notion of diversity (Gong et al., 2019). In language modeling, Muennighoff et al. (2025) incorporate question-domain diversity into reasoning-data selection, while WizardLM (Xu et al., 2024) and TeaMs-RL (Gu et al., 2024) explore instruction-data generation to improve generalization. The benefits of diverse training data have also been studied in supervised learning (Yu et al., 2022), medical image segmentation (Hofmanninger et al., 2020), and offline RL (Nguyen-Tang & Arora, 2023). These notions of diversity are related to, but do not necessarily imply, *geometric uniformity*: broad semantic or task coverage can coexist with local clustering and near-duplicate samples. Our work focuses on the spatial arrangement of samples in a fixed representation space, examining separation, local coverage, and simplex conditioning through optimization and approximation analyses. This perspective complements existing selection methods by investigating how geometric uniformity affects the relationship between data volume and learning efficiency, including when applied to subsets already selected by a model-aware method.

Uniformity on the hypersphere. Different from our Euclidean distance $h_{\min}$, there is a line of empirical work studying the uniformity on the hypersphere (Ethayarajh, 2019). Wang & Isola (2020) illustrated the importance of uniformity on the sphere in representation learning, which captures the distances between angles. Wang et al. (2022) later on showed that uniformity on hypersphere contributes to higher recommendation performance. Wang et al. (2023) tested the effect of uniformity on inference tasks. Additional work related to uniformity on the hypersphere includes Tan et al. (2023); Wang et al. (2024a); Ouyang et al. (2024); Sang et al. (2025); Cai et al. (2024); Son et al. (2024), etc.

Approximation of Neural Networks. A lot of literature has investigated the universal approximation properties of neural networks through the lens of Sobolev spaces. Early work by Andoni et al. (2014) considered the learnability of a two-layer neural network with analytic activation functions. Building on classical approximation theory, Lu et al. (2021) applied the Bramble–Hilbert theorem to characterize the expressivity of deep ReLU networks for smooth functions, deriving error bounds in the L^∞ norm. Extending the analysis to Sobolev norms, Gühring et al. (2020) established approximation rates for ReLU networks in $W^{s,p}$. Similarly, De Ryck et al. (2021) derived approximation error bounds in $W^{k,\infty}$ norm for tanh networks, while Shen et al. (2022) focused on convolutional architectures and their ability to approximate functions in Sobolev spaces. More recently, Jiang & Li (2024) analyzed the approximation capabilities of transformer architectures in the L^∞ norm. Regarding the general L^p norm for $1 \le p \le \infty$, there are works studying, for example, transfromers Yun et al. (2019); Kajitsuka & Sato (2023); Kratsios et al. (2021); Edelman et al. (2022); Luo et al. (2022), residual networks (ResNets) Lin & Jegelka (2018); Tabuada & Gharesifard (2022), feedforward networks (Hanin & Sellke, 2017; Kidger & Lyons, 2020; Park et al., 2020; Cai, 2022, etc.).

B Optimization: proof of Theorem 2

In this section, we only prove the data-distance consequence used in Theorem 2. The convergence guarantee and the assumptions ensuring the lower frame bound in Theorem 1 are taken from Wang (2026).

Proof of Theorem 2. Fix an iteration $s \le k$ and condition on the parameter value θ^s along the GD trajectory supplied by Theorem 1. For any data array $Z = (z_1, \cdots, z_N)$, let

$$J_s(Z) = \begin{pmatrix} \nabla_\theta f(\theta^s; z_1) \\ \vdots \\ \nabla_\theta f(\theta^s; z_N) \end{pmatrix}, \qquad A_s(Z) = J_s(Z) J_s(Z)^\top.$$

By definition, the lower frame bound in Theorem 1 is

$$\mu_{\mathrm{low,s,X}} = \lambda_{\min}(A_s(X)).$$

The model regularity assumptions in Wang (2026) imply that the maps $Z \mapsto A_s(Z)$ are locally Lipschitz for all fixed $s \le k$. Since each $A_s(Z)$ is symmetric, the map $Z \mapsto \lambda_{\min}(A_s(Z))$ is locally Lipschitz as well.

Fix x_i and let $\mathcal{D}_{i,H} \neq \varnothing$. Define the collapsed data set $X^{(i,H)}$ by replacing each x_j with x_i for $j \in \mathcal{D}_{i,H}$ and leaving all other points unchanged. Then

$$\|X - X^{(i,H)}\| = \sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2}.$$

At $X^{(i,H)}$, the stacked Jacobian has repeated block rows corresponding to x_i and the collapsed points. Hence $A_s(X^{(i,H)})$ is singular and

$$\lambda_{\min}(A_s(X^{(i,H)})) = 0.$$

By local Lipschitz continuity, for each $s \le k$ there exist $r_{s,i,H} > 0$ and $L_{s,i,H} > 0$ such that, whenever $\|X - X^{(i,H)}\| \le r_{s,i,H}$,

$$\begin{aligned}\mu_{\mathrm{low,s,X}} &= \lambda_{\min}(A_s(X)) \\ &\le L_{s,i,H}\|X - X^{(i,H)}\| \\ &= L_{s,i,H}\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2}.\end{aligned}$$

Taking $r_{k,i,H} = \min_{s \le k} r_{s,i,H}$ and $L_{k,i,H} = \max_{s \le k} L_{s,i,H}$ gives

$$\mu_{\mathrm{low,s,X}} \le L_{k,i,H}\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2}, \qquad \forall s \le k.$$

Finally, if $H = h_{\min}$ and $\mathcal{D}_{i,H} \neq \varnothing$, then $h_{ij} = h_{\min}$ for all $j \in \mathcal{D}_{i,H}$. Therefore

$$\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2} = \sqrt{|\mathcal{D}_{i,H}|}\, h_{\min}.$$

Taking $L_{k,i} = \sqrt{|\mathcal{D}_{i,H}|}\, L_{k,i,H}$ and $r_{k,i} = r_{k,i,H}/\sqrt{|\mathcal{D}_{i,H}|}$, the condition $h_{\min} \le r_{k,i}$ implies $\sqrt{|\mathcal{D}_{i,H}|}\, h_{\min} \le r_{k,i,H}$. Therefore,

$$\mu_{\mathrm{low,s,X}} \le L_{k,i} h_{\min}, \qquad \forall s \le k.$$

□

C Approximation: proof of Theorem 3 and Lemma 1

In this section, we will prove a formal version of Theorem 3. We first introduce some definitions and concepts used in this section.

For each simplex $\mathcal{I}_i = \mathrm{conv}\{v_{i0}, v_{i1}, \cdots, v_{in}\}$ in the triangulation $\mathcal{T}$, define

$$\delta v_{ij} = v_{ij} - v_{i0}, \qquad h_{ij} = \|\delta v_{ij}\|, \qquad \Delta v_i = (\delta v_{i1}, \delta v_{i2}, \cdots, \delta v_{in}),$$

and define the normalized edge matrix

$$\widehat{\Delta v}_i = \left(\frac{\delta v_{i1}}{h_{i1}}, \frac{\delta v_{i2}}{h_{i2}}, \cdots, \frac{\delta v_{in}}{h_{in}}\right).$$

We also define

$$h_{i,\min} = \min_{1 \le j \le n} h_{ij}, \qquad h_{i,\max} = \max_{1 \le j \le n} h_{ij}, \qquad h_{\min} = \min_{\mathcal{I}_i \in \mathcal{T}} h_{i,\min}, \qquad h_{\max_{n+1}} = \max_{\mathcal{I}_i \in \mathcal{T}} h_{i,\max}.$$

We consider the approximation in a piecewise Sobolev space. For an integer $r \geq 0$ and a finite triangulation $\mathcal{T}$ of $\mathcal{I}$, define the following piecewise Sobolev space

$$W^{r,p}(\mathcal{T};\mathbb{R}^q) = \left\{ \boldsymbol{u} \in L^p(\mathcal{I}^\circ;\mathbb{R}^q) : \boldsymbol{u}|_{\mathcal{I}_i^\circ} \in W^{r,p}(\mathcal{I}_i^\circ;\mathbb{R}^q) \text{ for every } \mathcal{I}_i \in \mathcal{T} \right\}.$$

where the function is in the sobolev space on every simplex $\mathcal{I}_i$ but not necessary the whole set $\mathcal{I}$.

For $\boldsymbol{u} \in W^{r,p}(\mathcal{T};\mathbb{R}^q)$, define its norm to be

$$\|\boldsymbol{u}\|_{r,p,\mathcal{T}} = \left(\sum_{\mathcal{I}_i \in \mathcal{T}} \|\boldsymbol{u}\|_{r,p,\mathcal{I}_i^\circ}^p \right)^{1/p} = \left(\sum_{\mathcal{I}_i \in \mathcal{T}} \sum_{|\alpha| \leq r} \left\| D^\alpha \left(\boldsymbol{u}|_{\mathcal{I}_i^\circ} \right) \right\|_{L^p(\mathcal{I}_i^\circ;\mathbb{R}^q)}^p \right)^{1/p}.$$

We first prove the relationship between the singular values of the matrix Δx_i and $h_{\min}$.

Proof of Lemma 1. Since the columns of Δx_i are edge vectors of $\mathcal{I}_i$,

$$\sigma_{\max}(\Delta x_i) \leq \|\Delta x_i\|_F = \left(\sum_{j=1}^d h_{i,j}^2 \right)^{1/2} \leq \sqrt{d}\, h_{\max_{d+1}}.$$

Taking the maximum over $\mathcal{I}_i \in \mathcal{T}$ gives the first inequality in equation 3. Moreover,

$$\Delta x_i = \widehat{\Delta x}_i \operatorname{diag}(h_{i,1}, \ldots, h_{i,d}).$$

By $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$, we have

$$\sigma_{\min}(\Delta x_i) \geq \sigma_{\min}(\widehat{\Delta x}_i) \min_{1 \leq j \leq d} h_{i,j} \geq \sigma_{\min}(\widehat{\Delta x}_i) h_{\min}.$$

Taking the minimum over $\mathcal{T}$ proves the second inequality. By Lemma 2, the sampled points are in general position with probability 1. Conditioning on this event, every simplex in $\mathcal{T}$ is then nondegenerate, so $s_{\mathcal{T}} > 0$. Hence $\hat{s}_{\mathcal{T}} > 0$.

□

Below is the formal version of Theorem 3.

Theorem 4 (Approximation by analytic neural networks)**.** *Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$ under Assumption 1, with $N \geq d+1$. Assume the ground truth function $g(x) \in W^{r,p}(\Omega)$ for $1 \leq p < \infty$ and $r \geq 1$, and $y_i = g(x_i) \in \mathbb{R}^d$. Let $1 \leq p < \infty$ and $m > \max\{r, d/p\}$. Assume that the analytic network class in equation 2 uses a non-polynomial activation $\sigma : \mathbb{R} \to \mathbb{R}$ admitting a power series with infinite radius of convergence. Then, with probability 1 over the joint distribution of $x_1, \cdots, x_N$, there exist a vector-valued function $\phi \in W^{m,p}(\mathcal{I}^\circ;\mathbb{R}^d)$ and a finite-width analytic network $f : \mathbb{R}^d \to \mathbb{R}^d$ such that*

$$\|f - g\|_{p,\mathcal{I}^\circ} \leq 3C \max\{H_{\mathcal{T}}, 1\}^m \min\{s_{\mathcal{T}}, 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ}. \tag{8}$$

Moreover, we have

$$\|f - g\|_{p,\mathcal{I}^\circ} \leq 3C_{\mathcal{T}} \max\{h_{\max_{d+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ}, \tag{9}$$

where $C_{\mathcal{T}} > 0$ depends on m, p, d and the finite triangulation $\mathcal{T}$, but not on g or the network approximation tolerance.

Proof of Theorem 4. By Lemma 2, the sampled points are in general position with probability 1. We now condition on this event. Every simplex in $\mathcal{T}$ is then nondegenerate, so $s_{\mathcal{T}} > 0$. Define

$$A_{\mathcal{T}} = C \max\{H_{\mathcal{T}}, 1\}^m \min\{s_{\mathcal{T}}, 1\}^{-r} > 0.$$

Since $\mathcal{I}^\circ$ is a bounded convex domain and $m > r$, $W^{m,p}(\mathcal{I}^\circ;\mathbb{R}^d)$ is dense in $W^{r,p}(\mathcal{I}^\circ;\mathbb{R}^d)$. Choose $\phi \in W^{m,p}(\mathcal{I}^\circ;\mathbb{R}^d)$ such that

$$\|g - \phi\|_{r,p,\mathcal{I}^\circ} \leq \frac{1}{2} \min\{1, A_{\mathcal{T}}\} \|g\|_{r,p,\mathcal{I}^\circ}. \tag{10}$$

The triangle inequality and monotonicity of Sobolev norms give

$$\|\phi\|_{m,p,\mathcal{I}^\circ} \geq \|\phi\|_{r,p,\mathcal{I}^\circ} \geq \|g\|_{r,p,\mathcal{I}^\circ} - \|g-\phi\|_{r,p,\mathcal{I}^\circ} \geq \frac{1}{2}\|g\|_{r,p,\mathcal{I}^\circ}.$$

Consequently, equation 10 implies

$$\|g-\phi\|_{r,p,\mathcal{I}^\circ} \leq A_{\mathcal{T}}\|\phi\|_{m,p,\mathcal{I}^\circ}. \tag{11}$$

By Theorem 7, thre exists a vector-valued piecewise polynomial ψ over $\mathcal{T}$ such that

$$\begin{aligned}\|\phi-\psi\|_{p,\mathcal{I}^\circ} &= \|\phi-\psi\|_{0,p,\mathcal{T}} \\ &\leq \|\phi-\psi\|_{r,p,\mathcal{T}} \\ &\leq A_{\mathcal{T}}\|\phi\|_{m,p,\mathcal{I}^\circ}.\end{aligned} \tag{12}$$

By Corollary 1 with $q = p$ and $\eta = A_{\mathcal{T}}\|\phi\|_{m,p,\mathcal{I}^\circ}$, there exists a finite-width analytic network f of the form equation 2 satisfying

$$\|f-\psi\|_{p,\mathcal{I}^\circ} \leq A_{\mathcal{T}}\|\phi\|_{m,p,\mathcal{I}^\circ}. \tag{13}$$

Combining equation 11, equation 12, and equation 13 with the triangle inequality yields

$$\begin{aligned}\|f-g\|_{p,\mathcal{I}^\circ} &\leq \|f-\psi\|_{p,\mathcal{I}^\circ} + \|\psi-\phi\|_{p,\mathcal{I}^\circ} + \|\phi-g\|_{p,\mathcal{I}^\circ} \\ &\leq 3A_{\mathcal{T}}\|\phi\|_{m,p,\mathcal{I}^\circ},\end{aligned}$$

which proves the first result.

For the second result, Lemma 1 gives $H_{\mathcal{T}} \leq \sqrt{d}\, h_{\max_{d+1}}$ and $s_{\mathcal{T}} \geq \hat{s}_{\mathcal{T}} h_{\min}$. Hence, using the same derivation as in the second conclusion of Theorem 7,

$$\begin{aligned}A_{\mathcal{T}} &\leq Cd^{m/2}\max\{\hat{s}_{\mathcal{T}}^{-r},1\}\max\{h_{\max_{d+1}},1\}^m \min\{h_{\min},1\}^{-r} \\ &= C_{\mathcal{T}}\max\{h_{\max_{d+1}},1\}^m \min\{h_{\min},1\}^{-r},\end{aligned}$$

where $C_{\mathcal{T}} = Cd^{m/2}\max\{\hat{s}_{\mathcal{T}}^{-r},1\}$. This proves the second result for the same ϕ and f. □

C.1 Polynomial approximation by analytic neural networks

In this section, we prove that the analytic network class can approximate polynomials. More precisely, we consider the following analytic network architecture equation 2:

$$u_0 = x, \qquad u_{\ell+1} = u_\ell + \varphi_\ell(\theta_\ell; u_\ell), \qquad f(\theta; x) = \varphi_L(\theta_L; u_L),$$

where φ_ℓ is analytic for $\ell = 0, \cdots, L$.

The following theorem adapts Theorem 1.1 of Park & Wojtowytsch (2022), and the corollary extends the result to finite piecewise polynomials.

Theorem 5 (Polynomial approximation by analytic neural networks)**.** *Let $K \subset \mathbb{R}^d$ be compact. Assume that the activation $\sigma : \mathbb{R} \to \mathbb{R}$ is not a polynomial and admits a power series with infinite radius of convergence. Then, for any polynomial map $p = (p_1, \cdots, p_d) : \mathbb{R}^d \to \mathbb{R}^d$ with $p_k \in P_s(\mathbb{R}^d)$ and $\eta > 0$, there exists a finite-width analytic network $f : \mathbb{R}^d \to \mathbb{R}^d$ of the form equation 2 such that*

$$\|f(\theta;\cdot) - p\|_{L^\infty(K;\mathbb{R}^d)} := \sup_{x\in K}\|f(\theta;x) - p(x)\|_2 < \eta.$$

Consequently, if $K = \overline{\mathcal{I}^\circ}$, then for every $1 \leq q < \infty$,

$$\|f(\theta;\cdot) - p\|_{L^q(\mathcal{I}^\circ;\mathbb{R}^d)} \leq |\mathcal{I}^\circ|^{1/q}\eta.$$

Proof. Choose a bounded open set $D \subset \mathbb{R}^d$ containing K, and let e_k be the kth standard basis vector of $\mathbb{R}^d$. For each polynomial p_k, Theorem 1.1 in Park & Wojtowytsch (2022) implies that there exists a finite-width shallow network

$$G_k(x) = \sum_{j=1}^{M_k} a_{kj}\sigma(w_{kj}^\top x + b_{kj})$$

such that

$$\sup_{x\in\overline{D}} |G_k(x) - p_k(x)| < \frac{\eta}{\sqrt{d}}, \qquad k = 1, \cdots, d. \tag{14}$$

We can then combine these scalar networks into the vector-valued shallow network

$$G(x) = \sum_{k=1}^{d} e_k G_k(x) = \sum_{k=1}^{d} \sum_{j=1}^{M_k} a_{kj} e_k \sigma(w_{kj}^{\top} x + b_{kj}).$$

Then, by equation 14, there is a single finite-width network of width $M = \sum_{k=1}^{d} M_k$ whose output weights $a_{kj}e_k$ lie in $\mathbb{R}^d$, such that, for every $x \in K$,

$$\|G(x) - p(x)\|_2 = \left(\sum_{k=1}^{d} |G_k(x) - p_k(x)|^2 \right)^{1/2} < \left(d \frac{\eta^2}{d} \right)^{1/2} = \eta.$$

We then align G inside equation 2 by setting all residual blocks to zero, so that $u_L = x$, and choosing the final layer to be

$$\varphi_L(\theta_L; z) = \sum_{k=1}^{d} \sum_{j=1}^{M_k} a_{kj} e_k \sigma(w_{kj}^{\top} z + b_{kj}).$$

Then $f(\theta; x) = \varphi_L(\theta_L; x) = G(x)$. Finally,

$$\|f(\theta; \cdot) - p\|_{L^q(\mathcal{I}^o; \mathbb{R}^d)} \le |\mathcal{I}^o|^{1/q} \|f(\theta; \cdot) - p\|_{L^\infty(K; \mathbb{R}^d)},$$

which gives the stated L^q bound. □

We then extend it to piecewise polynomials in the following corollary.

Corollary 1. *Let $\mathcal{I} \subset \mathbb{R}^d$ be a bounded polytope with a finite triangulation $\mathcal{T}$, and let $\boldsymbol{\psi}|_{\mathcal{I}_i} \in [P_{s_i}(\mathcal{I}_i)]^{d_{\text{out}}}$ be the piecewise polynomial, where $\mathcal{I}_i \in \mathcal{T}$. Then, under the same assumptions as Theorem 5, for $1 \le q < \infty$ and $\eta > 0$, there exists a finite-width analytic neural network $f : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$ of the form equation 2 such that*

$$\|f - \boldsymbol{\psi}\|_{L^q(\mathcal{I}^o; \mathbb{R}^{d_{\text{out}}})} < \eta.$$

Proof. Since $\mathcal{T}$ is finite and each restriction $\boldsymbol{\psi}|_{\mathcal{I}_i}$ is polynomial, $\boldsymbol{\psi} \in L^q(\mathcal{I}^o; \mathbb{R}^{d_{\text{out}}})$. Moreover, global polynomial maps are dense in $L^q(\mathcal{I}^o; \mathbb{R}^{d_{\text{out}}})$. Indeed, $C(\overline{\mathcal{I}}; \mathbb{R}^{d_{\text{out}}})$ is dense in this space, while the Stone–Weierstrass theorem shows that polynomial maps are uniformly dense in $C(\overline{\mathcal{I}}; \mathbb{R}^{d_{\text{out}}})$.

Hence, there exists a global polynomial map $\boldsymbol{P} : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$ such that

$$\|\boldsymbol{\psi} - \boldsymbol{P}\|_{L^q(\mathcal{I}^o; \mathbb{R}^{d_{\text{out}}})} < \frac{\eta}{2}.$$

Applying Theorem 5 to $\boldsymbol{P}$ on $K = \overline{\mathcal{I}}$ gives a finite-width analytic network f satisfying

$$\|f - \boldsymbol{P}\|_{L^\infty(\overline{\mathcal{I}}; \mathbb{R}^{d_{\text{out}}})} < \frac{\eta}{2|\mathcal{I}^o|^{1/q}}.$$

Consequently,

$$\|f - \boldsymbol{P}\|_{L^q(\mathcal{I}^o; \mathbb{R}^{d_{\text{out}}})} < \frac{\eta}{2}.$$

The triangle inequality yields

$$\|f - \boldsymbol{\psi}\|_{L^q} \le \|f - \boldsymbol{P}\|_{L^q} + \|\boldsymbol{P} - \boldsymbol{\psi}\|_{L^q} < \eta.$$

□

C.2 Definitions and lemmas related to computational geometry

In this section, we introduce some basic definitions and concepts, mainly based on De Berg (2000).

Definition 1 (Triangulation)**.** *A triangulation of the polygon is defined as a decomposition of a polygon into triangles by a maximal set of non-intersecting diagonals, or equivalently, no vertex of any triangle lies in the interior of an edge of another triangle.*

Definition 2 (n-simplex)**.** *A $n-$simplex in n-dimensional space is the convex hull of $n+1$ affinely independent points in $\mathbb{R}^n$.*

Proposition 1. *Let $\{v_0, v_1, \ldots, v_d\} \subset \mathbb{R}^n$ be the vertices of an n-simplex. Let $u_i = v_i - v_0, \quad$ for $i = 1, \ldots, n$.*

- *$n+1$ vertices, $\binom{n+1}{2}$ edges, $\binom{n+1}{k+1}$$k$-dimensional faces*
- *volumn $V = \frac{1}{n!}\left|\det\left([u_1\, u_2 \cdots u_n]\right)\right|$*

Definition 3 (General position)**.** *The set of points $\{x_i\}_{i=1}^N$ in $\mathbb{R}^n$ is in general position if any $n+1$ points are affinely independent.*

Theorem 6 (Proposition 2.2.4, in De Loera et al. (2010))**.** *Assume the set $\{x_i\}_{i=1}^N$ is in general position. Then there exists a triangulation, which consists of nondegenerate d-simplices.*

We then prove the following lemma, showing that i.i.d. samples obeying absolutely continuous density are in general position with probability one.

Lemma 2. *Suppose Assumption 1 holds and $N \geq d+1$. Then with probability 1 over the joint distribution, the set $\{x_i\}_{i=1}^N$ is in general position, i.e., any $d+1$ points are affinely independent.*

Proof. First, a set of $n+1$ points $x_0, \cdots, x_n$ is affinely independent if and only if the following vectors $x_1 - x_0, \cdots, x_n - x_0$ are linearly independent. This is equivalent to

$$\det([x_1 - x_0, \cdots, x_n - x_0]) \neq 0.$$

Let $F(x_0, \cdots, x_n) = \det([x_1 - x_0, \cdots, x_n - x_0])$. Then F is a polynomial and is nonzero. Let

$$\mathcal{N} = \{(x_0, \cdots, x_n) | F(x_0, \cdots, x_n) = 0\}.$$

Then the zero set of the polynomial is a null set in $\mathbb{R}^{(n+1)d}$, i.e., $\mathcal{N}$ has Lebesgue measure zero.

By Assumption 1, the joint law of $(x_1, \ldots, x_N)$ is absolutely continuous with respect to Leb_{Nd}, and hence $\mathbb{P}(\mathcal{N}_I) = 0$.

There are $\binom{N}{n+1}$ such set of $n+1$ points, and the finite union of measure zero set is still measure zero. Therefore, the set $\{x_i\}_{i=1}^N$ is in general position with probability one. □

C.3 Data-dependent polynomial approximation in Sobolev space

In this section, we prove the main theorem of our approximation result Theorem 3.

Theorem 7 (adaptive Bramble-Hilbert)**.** *Let $m > n/p$, $1 \leq p < \infty$, $0 \leq r \leq m$, and $q \geq 1$. Given N data points $\{(x_i, y_i)\}_{i=1}^N$ in general position, consider the convex hull of all the x_i, i.e., $\mathcal{I} = \mathrm{conv}\{x_1, \cdots, x_N\} \subseteq \Omega$. Let $\mathcal{T} = \{\mathcal{I}_i\}_{i=1}^{N_\mathcal{T}}$ be a finite triangulation of $\mathcal{I}$ into nondegenerate n-simplices. For any vector-valued function $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q) \in W^{m,p}(\mathcal{I}^\circ; \mathbb{R}^q)$, there exists a vector-valued piecewise polynomial $\boldsymbol{\psi}$, where $\boldsymbol{\psi}|_{\mathcal{I}_i^\circ} \in [P_{s_i}(\mathcal{I}_i)]^q$ for some $s_i \in \mathbb{N}_0$, s.t.*

$$\|\boldsymbol{\phi} - \boldsymbol{\psi}\|_{r,p,\mathcal{T}} \leq C \max\{H_\mathcal{T}, 1\}^m \min\{s_\mathcal{T}, 1\}^{-r} \|\boldsymbol{\phi}\|_{m,p,\mathcal{I}^\circ},$$

where $C > 0$ depends only on n, m, r, p. Moreover,

$$\|\boldsymbol{\phi} - \boldsymbol{\psi}\|_{r,p,\mathcal{T}} \leq C_\mathcal{T} \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\boldsymbol{\phi}\|_{m,p,\mathcal{I}^\circ},$$

where $C_\mathcal{T} = C\, n^{m/2} \max\{\hat{s}_\mathcal{T}^{-r}, 1\}$.

Note that for the vector-valued piecewise polynomial over $\mathcal{T}$, no continuity across simplex interfaces is required. The proof is an adaptation of Watkins (1979).

Proof of Theorem 7. We first prove the result for scalar-valued functions $\phi : \mathcal{I}^{\circ} \to \mathbb{R}$. Use the notation introduced above, and for each simplex define

$$F_i(\xi) = v_{i0} + \Delta v_i\, \xi, \qquad J_i = \det(\Delta v_i),$$

where $\xi \in \mathcal{D} = \{\xi \in \mathbb{R}^n : \xi_j \geq 0,\ \sum_{j=1}^n \xi_j \leq 1\}$. Since the vertices of $\mathcal{I}_i$ are affinely independent, Δv_i is invertible. For $\bar{\phi} = \phi \circ F_i$, by chain rule, for every multi-index α with $|\alpha| = k$,

$$D_x^{\alpha}\phi(F_i(\xi)) = \sum_{|\beta|=k} c_{\alpha\beta}\big((\Delta v_i)^{-1}\big) D_{\xi}^{\beta}\bar{\phi}(\xi),$$

where each coefficient $c_{\alpha\beta}((\Delta v_i)^{-1})$ is a finite sum of products of k entries of $(\Delta v_i)^{-1}$. Hence

$$|D_x^{\alpha}\phi(F_i(\xi))| \leq C_{n,k}\|(\Delta v_i)^{-1}\|_2^k \sum_{|\beta|=k} |D_{\xi}^{\beta}\bar{\phi}(\xi)|.$$

Using the change of variables $x = F_i(\xi)$, $dx = |J_i| d\xi$, we get

$$\begin{aligned}\|D_x^{\alpha}\phi\|_{0,p,\mathcal{I}_i^{\circ}} &= |J_i|^{1/p}\|D_x^{\alpha}\phi(F_i(\xi))\|_{0,p,\mathcal{D}^{\circ}} \\ &\leq C_{n,k}|J_i|^{1/p}\|(\Delta v_i)^{-1}\|_2^k \sum_{|\beta|=k} \|D_{\xi}^{\beta}\bar{\phi}\|_{0,p,\mathcal{D}^{\circ}}.\end{aligned}$$

Summing over all $|\alpha| = k$ gives

$$\begin{aligned}|\phi|_{k,p,\mathcal{I}_i^{\circ}} &\leq C_{n,k}|J_i|^{1/p}\|(\Delta v_i)^{-1}\|_2^k |\bar{\phi}|_{k,p,\mathcal{D}^{\circ}} \\ &= C_{n,k}|J_i|^{1/p}\sigma_{\min}(\Delta v_i)^{-k}|\bar{\phi}|_{k,p,\mathcal{D}^{\circ}},\end{aligned} \tag{15}$$

Similarly, the same argument can be applied to $D_{\xi}^{\alpha}\bar{\phi}(\xi)$

$$|\bar{\phi}|_{k,p,\mathcal{D}^{\circ}} \leq C_{n,k}|J_i|^{-1/p}\sigma_{\max}(\Delta v_i)^k |\phi|_{k,p,\mathcal{I}_i^{\circ}}, \tag{16}$$

Let $u \in W^{m,p}(\mathcal{D}^{\circ};\mathbb{R})$ be scalar valued, and let $\Pi : L^p(\mathcal{D}^{\circ};\mathbb{R}) \to P_{m-1}(\mathcal{D})$ be the scalar projector characterized by

$$\int_{\mathcal{D}} (u - \Pi u) q\, d\xi = 0, \qquad \text{for } q \in P_{m-1}(\mathcal{D}).$$

This projector is well defined because the Gram matrix of any basis of $P_{m-1}(\mathcal{D})$ is positive definite. Moreover, $\Pi q = q$ for every $q \in P_{m-1}(\mathcal{D})$, and finite-dimensional norm equivalence gives $\|\Pi u\|_{r,p,\mathcal{D}^{\circ}} \leq C\|u\|_{0,p,\mathcal{D}^{\circ}}$ for $0 \leq r \leq m$.

Let $\lambda_0, \ldots, \lambda_n$ be the barycentric coordinates associated with the vertices $\bar{v}_0, \ldots, \bar{v}_n$ of $\mathcal{D}$. We define

$$\bar{B}u = \Pi u + \sum_{j=0}^{n} \big(u(\bar{v}_j) - \Pi u(\bar{v}_j)\big)\lambda_j.$$

Then $\bar{B}u(\bar{v}_j) = u(\bar{v}_j)$ and $\bar{B}q = q$ for every $q \in P_{m-1}(\mathcal{D})$. Since $m > n/p$, point evaluation is bounded on $W^{m,p}(\mathcal{D}^{\circ})$. Together with the boundedness of Π and finite-dimensional norm equivalence for the fixed barycentric coordinates, this shows that $\bar{B} : W^{m,p}(\mathcal{D}^{\circ}) \to W^{r,p}(\mathcal{D}^{\circ})$ is bounded for $0 \leq r \leq m$.

Since $\mathcal{D}$ is a fixed simplex, by Theorem 1 in Bramble & Hilbert (1970)

$$\inf_{q \in P_{m-1}(\mathcal{D})} \|u - q\|_{m,p,\mathcal{D}^{\circ}} \leq C_{\mathcal{D},m,p}|u|_{m,p,\mathcal{D}^{\circ}}.$$

For every $q \in P_{m-1}(\mathcal{D})$, we have $(I - \bar{B})q = 0$. Therefore, by the boundedness of $I - \bar{B}$,

$$\begin{aligned}\|u - \bar{B}u\|_{r,p,\mathcal{D}^{\circ}} &\leq \|I - \bar{B}\|_{\mathcal{L}(W^{m,p},W^{r,p})} \inf_{q \in P_{m-1}(\mathcal{D})} \|u - q\|_{m,p,\mathcal{D}^{\circ}} \\ &\leq C_{\mathcal{D},m,r,p}|u|_{m,p,\mathcal{D}^{\circ}},\end{aligned}$$

where $\|\cdot\|_{\mathcal{L}(W^{m,p},W^{r,p})}$ is the bounded linear operator norm. Taking $u = \bar{\phi} = \phi \circ F_i$ gives

$$\|\bar{\phi} - \bar{B}\bar{\phi}\|_{r,p,\mathcal{D}^{\circ}} \leq C_{\mathcal{D},m,r,p}|\bar{\phi}|_{m,p,\mathcal{D}^{\circ}}. \tag{17}$$

On each simplex $\mathcal{I}_i$, define the scalar polynomial

$$\psi_i(x) = (\bar{B}\bar{\phi})(F_i^{-1}(x)), \qquad x \in \mathcal{I}_i,$$

where $\psi|_{\mathcal{I}_i^\circ} = \psi_i$. Combining equation 15, equation 17, and equation 16, we obtain

$$\begin{aligned}
\|\phi - \psi_i\|_{r,p,\mathcal{I}_i^\circ} &\le C|J_i|^{1/p} \max\{\|(\Delta v_i)^{-1}\|_2^r, 1\} \|\bar{\phi} - \bar{B}\bar{\phi}\|_{r,p,\mathcal{D}^\circ} \\
&\le C|J_i|^{1/p} \max\{\|(\Delta v_i)^{-1}\|_2^r, 1\} |\bar{\phi}|_{m,p,\mathcal{D}^\circ} \\
&\le C \max\{\sigma_{\max}(\Delta v_i), 1\}^m \max\{\sigma_{\min}(\Delta v_i)^{-r}, 1\} \|\phi\|_{m,p,\mathcal{I}_i^\circ} \\
&= C \max\{\sigma_{\max}(\Delta v_i), 1\}^m \min\{\sigma_{\min}(\Delta v_i), 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}_i^\circ}.
\end{aligned}$$

Summing over the simplices gives

$$\|\phi - \psi\|_{r,p,\mathcal{T}} \le C \max_i \max\{\sigma_{\max}(\Delta v_i), 1\}^m (\min_i \min\{\sigma_{\min}(\Delta v_i), 1\})^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ}, \tag{18}$$

where the constant depends only on n, m, r, p and the reference simplex, and ψ is piecewise polynomial.

By Lemma 1, we have

$$\max_i \max\{\sigma_{\max}(\Delta v_i), 1\}^m \le n^{m/2} \max\{h_{\max_{n+1}}, 1\}^m,$$

and

$$\begin{aligned}
(\min_i \min\{\sigma_{\min}(\Delta v_i), 1\})^{-r} &\le \min\{\hat{s}_{\mathcal{T}} h_{\min}, 1\}^{-r} \\
&\le \max\{\hat{s}_{\mathcal{T}}^{-r}, 1\} \min\{h_{\min}, 1\}^{-r}.
\end{aligned}$$

Therefore equation 18 implies

$$\|\phi - \psi\|_{r,p,\mathcal{T}} \le C\, n^{m/2} \max\{\hat{s}_{\mathcal{T}}^{-r}, 1\} \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p,\mathcal{I}^\circ},$$

where $C_{\mathcal{T}} = C\, n^{m/2} \max\{\hat{s}_{\mathcal{T}}^{-r}, 1\}$.

The above results for scalar-valued functions extend naturally to vector-valued functions $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q)$ and polynomial $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)$ through applying it componentwisely, which concludes the proof.

□

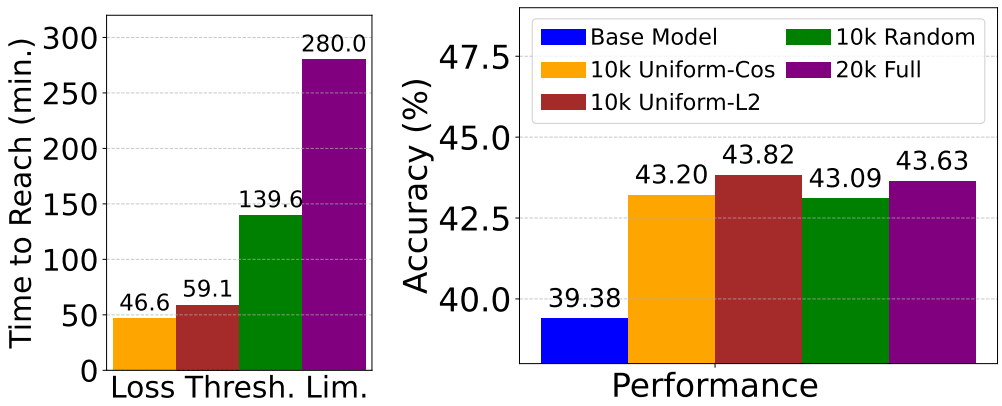
D Ablation Experiments and Data Visualization

D.1 Effect of the distance measure.

To assess the sensitivity of uniform selection to the distance measure, we compare Euclidean (ℓ_2) distance and cosine dissimilarity within the farthest-first procedure in Section 3.3, using the same Word2Vec representations and subset size. As shown in Figure 6, On WizardLM with LLaMA-1-13B, the cosine- and ℓ_2-based 10k subsets reach the training-loss threshold of 0.6 in 46.6 and 59.1 minutes, respectively, compared with 139.6 minutes for the random 10k subset and 280.0 minutes for the full 20k dataset. At matched-loss checkpoints, their downstream accuracies are 43.20% and 43.82%, respectively, compared with 43.09% for random selection and 43.63% for full-data training. Cosine-based selection reaches the target loss faster, whereas ℓ_2-based selection achieves higher downstream accuracy. These results show that the efficiency gains from uniform selection persist across both distance measures in this setting, while the choice of metric affects the resulting balance between training speed and downstream performance.

D.2 Data Visualization via PCA

We visualize data geometry using two-dimensional PCA projections of sentence embeddings (Maćkiewicz & Ratajczak, 1993). For each dataset, PCA is fitted on the full pool and applied to its uniform and random subsets, with a shared MinMax transformation rescaling the projected coordinates to $[0, 1]$. Figure 7(a)–(c) compares the full WizardLM 20k dataset (Xu et al., 2024) with



(a) Efficiency comparison (The lower, the better).

(b) Performance comparison (The higher, the better).

Figure 6: **Training efficiency and downstream performance on WizardLM with LLaMA-1-13B. (a)** Wall-clock time to first reach a training-loss threshold of 0.6 using cross-entropy loss and AdamW. Both 10k uniform subsets converge substantially faster than the 10k random subset and the full 20k dataset. **(b)** Downstream performance (ARC (Clark et al., 2018) and TruthfulQA (Lin et al., 2022)) at matched-loss checkpoints. Despite using only half of the data, uniform selection achieves performance comparable to or better than the full dataset, demonstrating improved training efficiency without sacrificing model quality.

10k subsets obtained by farthest-first selection and random sampling. The uniform subset exhibits less local crowding while retaining broad coverage of the full dataset in the displayed projection, whereas the random subset preserves more of its concentrated regions. The nearest-neighbor distance coloring and the larger minimum separation $h_{\min}$ in panel (d) further illustrate the reduction in near-collisions. These observations qualitatively support our geometric motivation for uniform selection; the PCA projections alone do not establish semantic coverage or preservation of the full high-dimensional geometry.

D.3 Subset Size, Data Geometry, and Learning Performance

To further validate our insights, we examine how uniform subset size affects data geometry, training efficiency, and downstream performance on the WizardLM 20k pool. This analysis provides empirical context for the separation–coverage trade-off discussed in Section 3.3.

Geometric measurements. We compute geometric diagnostics in a fixed 100-dimensional Word2Vec representation space using Euclidean distance. The TwoNN estimator (Facco et al., 2017) gives an intrinsic dimension estimate of 17.7, and we use $\hat{d} = 18$ neighbors. To characterize local neighborhood sufficiency while reducing sensitivity to isolated samples, we measure

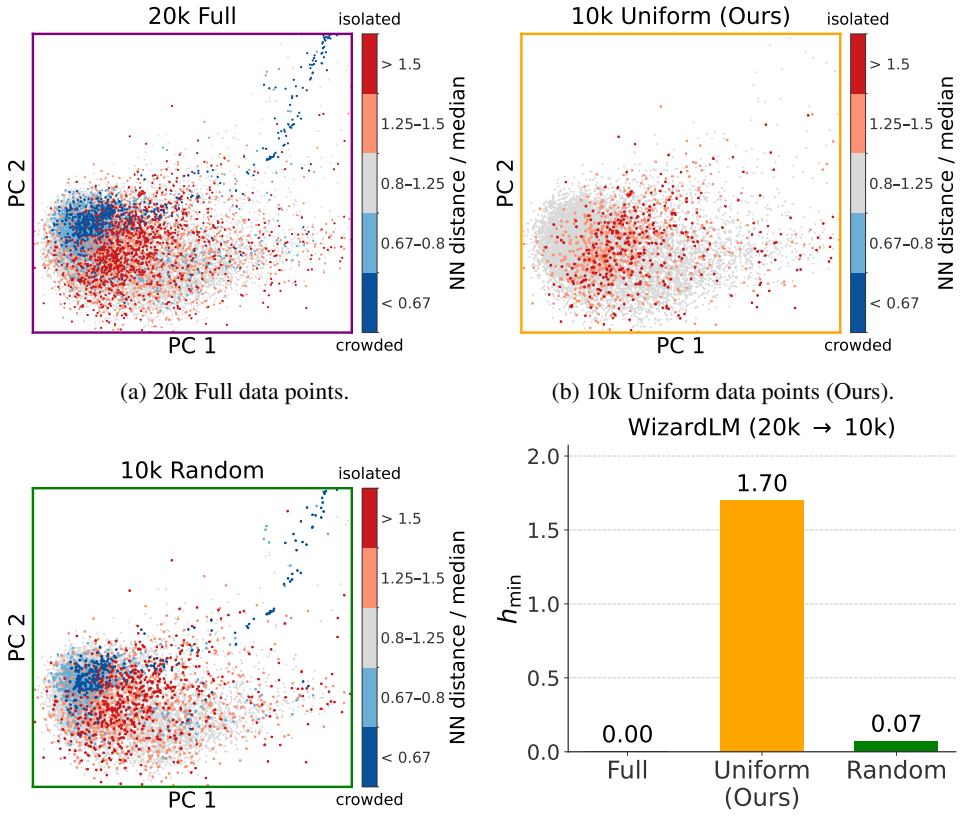
$$r_{95}(S) = \text{Quantile}_{0.95}\left(\{d_{\hat{d}}(x;S)\}_{x\in S}\right), \tag{19}$$

$$m = \text{median}_{x\in\mathcal{D}}\, d_1(x;\mathcal{D}), \tag{20}$$

where $d_k(x;S)$ denotes the distance to the k-th nearest neighbor in $S \setminus \{x\}$. Here, $m = 1.562$ measures typical local spacing in the full pool, and r_{95}/m is invariant to uniform rescaling of the embeddings. The radius r_{95} characterizes local neighborhood sufficiency; it does not guarantee global coverage or well-conditioned simplices.

Subset size and geometric trade-offs. Table 1 shows that smaller uniform subsets have greater minimum separation $h_{\min}$ but larger neighborhood radii r_{95}, reflecting sparser local neighborhoods. The 10k subset achieves the downstream accuracy of 45.65%. These results are consistent with a trade-off between sample separation and local neighborhood sufficiency: neither quantity alone predicts the best downstream performance.

Training efficiency and downstream performance. Figure 8 complements these geometric measurements with downstream accuracy and time to the training-loss threshold of 0.65. All evaluated



(a) 20k Full data points.

(b) 10k Uniform data points (Ours).

(c) 10k Random data points.

(d) h_{min} distance comparison.

Figure 7: **Data geometry of uniform and random subsets on WizardLM. (a)–(c)** Two-dimensional PCA visualizations of the full 20k dataset, the 10k uniform subset, and an equal-size random subset. Colors indicate nearest-neighbor distances normalized by the median: blue denotes locally crowded points, while red denotes relatively isolated points. Uniform selection reduces local crowding while retaining broad coverage of the distribution in the displayed projection. **(d)** Minimum pairwise separation $h_{\min}$ for each dataset. The uniform subset achieves substantially greater separation (1.70) than the random subset (0.07) and the full dataset (0.00, rounded). Together, these comparisons illustrate how geometry-guided uniform selection reduces near-collisions and improves sample separation while retaining only half of the training data.

uniform subsets reach this threshold faster than the random 10k baseline. The uniform 10k subset combines the highest observed accuracy with a training time of 23.5 minutes, compared with 68.6 minutes for the random baseline. The 5k subset reaches the threshold fastest but achieves lower accuracy, showing that faster attainment of the training-loss threshold does not necessarily imply better downstream performance.

E Experiment Settings

We summarize the key configurations [2] used in our experiments to ensure reproducibility and clarity. Table 2 details the training hyperparameters. Based on these and building on Llama-X[3], we train LLaMA-1-7B (one A100 GPU) and 13B (two A100 GPUs) models on datasets of varying sizes: WizardLM[4] (10k uniform datasets, 10k random datasets, and 20k full datasets). The uniform subsets

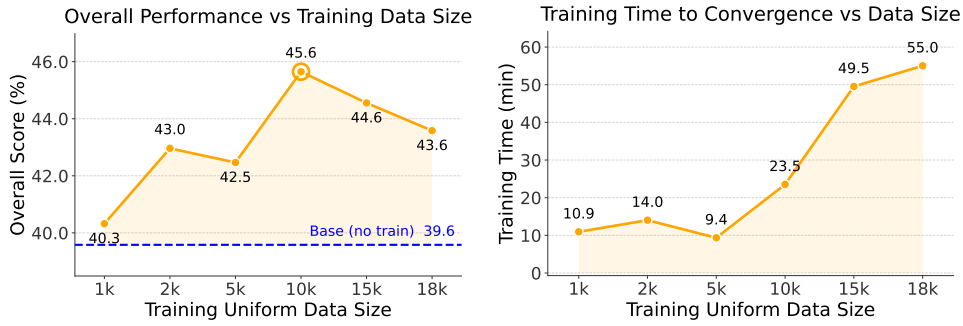
[2] https://anonymous.4open.science/r/data-uniformity-1A5C

[3] https://github.com/AetherCortex/Llama-X

[4] https://huggingface.co/datasets/WizardLMTeam/WizardLM_evol_instruct_V2_196k

Table 1: Separation, neighborhood radius, and downstream accuracy across uniform subset sizes on the WizardLM 20k pool ($\hat{d}$ = 18, m = 1.562). Overall accuracy is the mean of ARC-Challenge and TruthfulQA-MC accuracy at checkpoints reaching a training-loss threshold of 0.65, using LLaMA-1-7B.

Subset	Ratio	$h_{\min}$	r_{95}	r_{95}/m	Overall (%)
1k	0.05	3.004	5.64	3.61	40.32
2k	0.10	2.582	4.77	3.06	42.96
5k	0.25	2.036	3.95	2.53	42.47
10k	**0.50**	**1.700**	**3.39**	**2.17**	**45.65**
15k	0.75	1.431	3.10	1.99	44.55
18k	0.90	1.235	2.99	1.91	43.58
Full pool	1.00	0.000	2.91	1.86	–



(a) Overall accuracy (higher is better).

(b) Time to target loss (lower is better).

Figure 8: **Effect of uniform subset size on WizardLM with LLaMA-1-7B. (a)** Mean ARC-Challenge and TruthfulQA-MC accuracy at checkpoints reaching a training-loss threshold of 0.65, compared with the base model and a random 10k subset. The uniform 10k subset achieves the highest observed accuracy. **(b)** Wall-clock time to first reach the same loss threshold; all evaluated uniform subsets reach it faster than the random 10k baseline.

from WizardLM are constructed using Word2Vec embeddings and a greedy selection strategy that maximizes uniformity. Specifically, one data point is randomly selected to initialize the set, and subsequent points are chosen to maximize the minimum cosine distance to all previously selected points.

Table 3 shows the DeepSpeed configuration using cross-entropy loss with the Adam optimizer. Table 4 presents the DeepSpeed configuration used for ℓ^2-regularized SGD training with Zero Stage 3 optimization and CPU offloading. Finally, we leverage lm-evaluation-harness (Gao et al., 2024) for our evaluation. Table 5 outlines the evaluation setup, including tasks, few-shot settings, and batch size settings used in the LM Evaluation Harness. These configurations collectively support our experiments across diverse datasets and model scales.

F Limitations and Broader Impact.

Our findings may not generalize to all learning settings, the stopping radius may require recalibration in other settings. Despite these limitations, our results demonstrate that geometry-guided uniform selection can reduce data requirements and training costs while improving downstream performance in the evaluated settings.

Table 2: Key experimental settings used in our study, including training with ℓ^2 loss and cross-entropy loss using SGD and Adam optimizers.

Parameters	value	Parameters	value
GPUs	[1, 2]	model max length	512
per device train batch size	64	per device eval batch size	1
gradient accumulation steps	1	evaluation strategy	no
learning rate	2e-5	warmup steps	2
logging steps	1	lr scheduler type	constant
gradient checkpointing	True	fp16	True

Table 3: Key DeepSpeed configuration parameters used in our experiments (Zero Stage 3 with CPU offloading, FP16, AdamW optimizer).

Parameters	Value	Parameters	Value
zero optimization stage	3	overlap communication	True
offload optimizer device	cpu	pin memory (optimizer)	True
offload param device	cpu	pin memory (param)	True
contiguous gradients	True	sub group size	0
reduce bucket size	auto	prefetch bucket size	auto
param persistence threshold	auto	max live parameters	0
max reuse distance	0	gather 16bit weights on save	True
fp16 enabled	True	auto cast	False
loss scale	0	initial scale power	32
loss scale window	1000	hysteresis	2
optimizer type	AdamW	learning rate	2e-5
beta1	0.9	beta2	0.999
epsilon	1e-8	weight decay	0
train batch size	auto	micro batch size per GPU	auto
gradient accumulation steps	auto	wall clock breakdown	False

Table 4: Key DeepSpeed configuration parameters used in the experiments (Zero Stage 3 with CPU offloading, FP16, SGD optimizer).

Parameters	Value	Parameters	Value
zero optimization stage	3	overlap communication	True
offload optimizer device	cpu	pin memory (optimizer)	True
offload param device	cpu	pin memory (param)	True
contiguous gradients	True	sub group size	0
reduce bucket size	auto	prefetch bucket size	auto
param persistence threshold	auto	max live parameters	0
max reuse distance	0	gather 16bit weights on save	True
allow untested optimizer	True	min loss scale	1
fp16 enabled	True	auto cast	False
loss scale	0	initial scale power	32
loss scale window	1000	hysteresis	2
optimizer type	SGD	learning rate	2e-5
momentum	0.0	weight decay	0
train batch size	auto	micro batch size per GPU	auto
gradient accumulation steps	auto	wall clock breakdown	False

Table 5: Evaluation settings for baseline models using the LM Evaluation Harness.

Parameter	Value	Parameter	Value
Model type	hf-causal-experimental	Batch size	2
Few-shot for `arc_challenge`	25	Few-shot for `truthfulqa_mc`	0

Table 6: Training, data-selection, and evaluation settings for Qwen3-1.7B-Base on Tulu-3.

Parameter	**Value**	**Parameter**	**Value**
Model	Qwen3-1.7B-Base	Dataset	Tulu-3
Optimizer	AdamW	Loss function	Cross-entropy
Learning rate	5×10^{-6}	Training epochs	2
Candidate pool	Stratified 40k	Full training set	40k
Uniform subset size	20k	Random subset size	20k
Selection method	Greedy farthest-first	Selection representation	Word2Vec
Training-loss threshold	0.786	Evaluation checkpoints	Matched loss
ARC-Challenge	25-shot	GSM8K	5-shot
IFEval	0-shot	TruthfulQA-mc2	0-shot
Reference model	Base model (no fine-tuning)	Overall score	Mean accuracy over four tasks
Efficiency metric	Wall-clock time to first reach the training-loss threshold		