

DATA UNIFORMITY IMPROVES TRAINING EFFICIENCY AND MORE, WITH A CONVERGENCE FRAMEWORK BEYOND THE NTK REGIME

Anonymous authors

Paper under double-blind review

ABSTRACT

Data selection plays a crucial role in data-driven decision-making, including in large language models (LLMs), and is typically task-dependent. Properties such as data quality and diversity have been extensively studied and are known to enhance model performance. However, it remains unclear whether there exist other quantitative and general principles of data selection that can consistently improve performance, especially for complicated tasks. In this paper, we demonstrate that selecting more uniformly distributed data can improve training efficiency while enhancing performance. Specifically, we establish that more uniform (less biased) distribution leads to a larger minimum pairwise distance between data points, denoted by $h_{\min}$, and prove that a smaller $h_{\min}$ can slow down the training dynamics of gradient descent (GD). Moreover, we theoretically show that the approximation error of neural networks decreases as $h_{\min}$ increases. Our analysis introduces a convergence framework for GD beyond the Neural Tangent Kernel (NTK) regime, applicable to a broad class of architectures, including transformers, without requiring Lipschitz smoothness. This framework further provides theoretical justification for the use of residual connection and function composition in deep neural architectures. In the end, we conduct comprehensive experiments for supervised fine-tuning across various settings, including different optimization strategies, model sizes, and training datasets. The results consistently demonstrate that selecting data by maximizing pairwise distance significantly accelerates training and achieves comparable or better performance in LLMs across diverse datasets. Code and Datasets are available at the link: <https://anonymous.4open.science/r/data-uniformity-1A5C>.

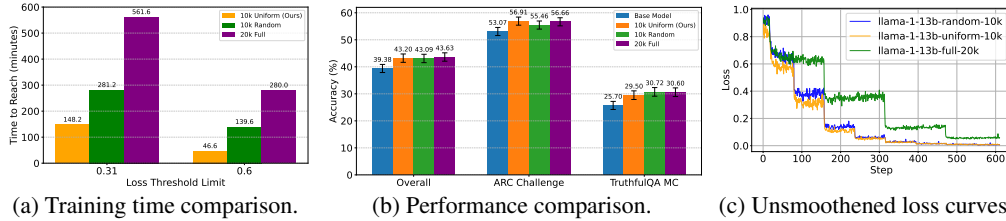


Figure 1: Comparison of LLaMA-1-13B training efficiency and evaluation performance across different data selection strategies. (a) **Time to reach loss thresholds** (0.31 and 0.6), reported based on wall-clock time using a single A100 GPU. The y-axis shows the time (in minutes) required to reach each target loss for the first time. Uniform selection consistently leads to faster convergence, indicating improved training efficiency. (b) **Evaluation performance on ARC Challenge and TruthfulQA MC** at matched loss checkpoints, showing that the 10k Uniform subset outperforms the 10k Random subset and matches the full 20k dataset. (c) **Training loss curves**, 10k uniform subset is faster than 10k random subset and full dataset.

1 INTRODUCTION

Data selection is fundamental to a lot of applications including large language models (LLMs) (Al-balak et al., 2024; Zhao et al., 2023; Chang et al., 2024), such as TeaMs-RL (Gu et al., 2024) and

WizardLM (Xu et al., 2024). One main difficulty of data selection is that it is often tailored to specific tasks, which highlights the need for general and widely applicable rules to guide the data selection process. Existing rules include improving data quality and data diversity (Muennighoff et al., 2025; Albalak et al., 2024; Gu et al., 2024), which have been widely studied and are recognized for their positive impact on model performance (Gu et al., 2024; Ba et al., 2024; Gong et al., 2019, etc.). However, data selection is still a challenging problem that calls for more quantitative and broadly applicable principles.

Regarding the role of data, existing results from approximation theory suggest that uniform sampling exhibits near-optimal behavior under many circumstances. Krieg et al. (2022) showed that i.i.d. random sampling achieves optimal convergence rates, and Krieg & Sonleitner (2024) proved that uniform sampling is asymptotically as effective as optimal deterministic schemes for Sobolev functions. In the context of learning, Dong et al. (2024) further demonstrated that uniform sampling yields near-optimal generalization in low-dimensional linear probing. These results collectively emphasize the efficiency of uniform sampling, despite its simplicity. Motivated by these results, we wonder whether selecting data to be more uniformly distributed can benefit training and lead to improved generalization performance for complicated tasks.

In this paper, we demonstrate that increased data uniformity accelerates training and achieves performance comparable to using a larger dataset. More precisely, we characterize the effect of data uniformity through the minimum distance between input data points, denoted by $h_{\min}$, both theoretically and experimentally:

- Theoretically, we first establish the relationship between data uniformity and the minimum distance $h_{\min}$ (Theorem 1): a more uniform (less biased) distribution results in larger $h_{\min}$.
- On the optimization side, we first develop a general convergence framework for gradient descent (GD) beyond the neural tangent kernel (NTK) regime (Theorem 2; Remark 1). This framework applies to a broad family of complicated architectures with residual connections, including deep transformers with feedforward and attention layers using softmax, tanh, GELU, SiLU, and other analytic activations (Definition 1). It builds on tools from measure theory and differential topology, combined with polynomial generalized smoothness (Definition 4) and local relaxed dissipativity (Definition 5), and is also applicable to general nonconvex optimization without Lipschitz smoothness. Based on this convergence analysis, we show that smaller $h_{\min}$ can slow down training (Corollary 3), and provide theoretical insights into the benefits of residual connection and function composition in neural networks (Remark 2).
- On the approximation side, we prove that larger $h_{\min}$ and smaller local maximum distance $h_{\max_{d+1}}$ lead to smaller approximation error between the neural network and ground truth (Theorem 3). This is achieved by establishing a data-dependent Bramble-Hilbert lemma, combined with ideas in computational geometry.
- Experimentally, we demonstrate the benefits of selecting more uniformly distributed data by maximizing pairwise distances between data points in supervised fine-tuning. Specifically, we employ our approach into state-of-the-art (SOTA) data distillation and data selection baselines, including TeamMs-RL (Gu et al., 2024), WizardLM (Xu et al., 2024), Zcore (Griffin et al., 2024), and LESS (Xia et al., 2024). Comparative experiments show that this uniformity-driven sampling strategy consistently leads to faster convergence and comparable or improved performance across both ℓ^2 -SGD and cross-entropy loss with Adam optimization (e.g., as shown in Figure 1). Our results generalize across multiple data sources, optimization strategies, dataset sizes, and model scales, including LLaMA-1 7B and 13B models.

2 RELATED WORK

Convergence of Neural Networks. In the overparameterized regime, the neural tangent kernel (NTK) framework, introduced by Jacot et al. (2018), has led to a series of convergence results showing that when the network width is sufficiently large—typically a high-order polynomial in the number of samples N , network depth L , and inverse of angle between data—(S)GD with infinitesimal learning rates can drive the training loss to zero exponentially (Allen-Zhu et al., 2019; Du et al., 2019; Lee et al., 2019; Zou et al., 2020; Zou & Gu, 2019; Ji & Telgarsky, 2019; Chen et al., 2020b; Song & Yang, 2019; Oymak & Soltanolkotabi, 2020). However, in the NTK regime, networks essentially behave as linear models and do not exhibit meaningful feature learning. To address this limitation, recent work has moved towards analyzing feature learning, using either gradient flow or

finite number of large steps Yang & Hu (2020); Chen et al. (2022; 2025); Ba et al. (2022); Allen-Zhu & Li (2020; 2022); Cao et al. (2022); Shi et al. (2021); Telgarsky (2022). There is also a series of works that analyze SGD convergence under various regularity assumptions, including the PL condition, quadratic growth, the aiming condition, and Lipschitz smoothness (Liu et al., 2023; Ma et al., 2018, etc.). Another line of work adopts a mean-field perspective (Song et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Wei et al., 2019; Chen et al., 2020a; Sirignano & Spiliopoulos, 2020; Fang et al., 2021), dynamical mean-field theory perspective (Bordelon & Pehlevan, 2022; 2023), and distributional perspective Han & Imaizumi (2025). In contrast to these approaches, which typically assume fixed architectures and/or infinitesimal learning rates, our work introduces tools applicable to a broad class of networks and remains effective under larger learning rates, offering a theoretical foundation for convergence in more practical and complex settings.

See more related work in Appendix A.

3 PRELIMINARY

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 $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$, 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 $|\alpha| = |(\alpha_1, \dots, \alpha_d)| = \sum_{i=1}^d \alpha_i$; for a set Ω , we use $|\Omega|$ to denote the area of the set. The closure and interior of a set Ω are denoted as $\bar{\Omega}$ and Ω° . 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 $\text{conv}\{v_1, \dots, v_n\}$ to denote the convex hull over $v_1, \dots, v_n$. We denote $B_x(r)$ to be the open ball of radius r centered at x under Euclidean distance. We use $\mathbb{R}_{\geq 0}$ to denote all the non-negative real values.

We introduce the Sobolev space $W^{m,p}(\Omega)$, which is defined as $W^{m,p}(\Omega) = \{\phi \in L^p(\Omega) : D^\alpha \phi \in L^p(\Omega), \forall |\alpha| \leq k\}$, with Sobolev norm $\|\phi\|_{m,p} = (\sum_{|\alpha| \leq m} \|D^\alpha \phi\|_p^p)^{1/p}$ for $1 \leq p < \infty$, and $\max_{|\alpha| \leq m} \|D^\alpha \phi\|_\infty$ for $p = \infty$ where $D^\alpha \phi = \frac{\partial^{|\alpha|} \phi}{\partial^{\alpha_1} x_1 \dots \partial^{\alpha_d} x_d}$.

Next, we introduce analyticity, polynomial generalized continuity and smoothness, and local relaxed dissipativity as key properties for proving the convergence of neural networks.

3.1 ANALYTIC FUNCTION

Analyticity is a useful property in both measure theory and differential topology. For a compact domain $\mathcal{D}$, real-analytic functions are dense in the space of continuous function $C(\mathcal{D})$ under Whitney C^0 -topology (Grauert, 1958), i.e., with respect to uniform convergence on compact sets. Therefore, assuming analyticity is reasonable in many settings. It is defined as follows.

Definition 1 (Analytic function). *A function $f(x) = (f_1(x), \dots, f_m(x))^T : U \rightarrow \mathbb{R}^m$ with an open subset $U \subseteq \mathbb{R}^m$, is called real-analytic on U , if for each $x_0 \in U$, the function $f_i(x)$ can be represented by a convergent power series in some neighborhood of x_0 , for all $i = 1, \dots, m$.*

Many activation functions and components of neural networks are real-analytic. For example:

Corollary 1. *Any product, sum, and composition of softmax, tanh, sigmoid, GELU, SiLU, polynomial, normalization $\tau_\epsilon(x) = \frac{x}{\sqrt{\|x\|^2 + \epsilon^2}}$ for some $\epsilon > 0$, and exponential functions, is real-analytic.*

The above corollary implies that feedforward layers using tanh, GELU, or SiLU, as well as attention layers with softmax or linear activation, are real-analytic. Consequently, architectures such as transformers, residual networks, and feedforward networks built with real-analytic activations and normalizations are themselves real-analytic.

3.2 POLYNOMIAL GENERALIZED CONTINUITY AND SMOOTHNESS

Classical non-convex optimization theory typically assumes Lipschitz smoothness of the objective function, a condition that is rarely satisfied in neural network training. One direction of the generalization of Lipschitz smoothness stems from the concept (L_0, L_1) -smoothness (Zhang et al., 2019). Building on this, Li et al. (2023) introduced the generalized smoothness for a broader class of non-Lipschitz smooth functions. Motivated by these developments, we propose a new formulation in nonconvex optimization tailored to neural network architectures, based on the following concepts.

Definition 2 (Poly-boundedness). *A function $f(x) : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^d$ is polynomially bounded if*

$$\|f(x)\| \leq S(\|x_1\|, \dots, \|x_{n_x}\|), \quad \forall x \in \Omega,$$

where $x_i \in \mathbb{R}^{n_{i,1} \times n_{i,2}}$, and $\dim x = n = \sum_{i=1}^{n_x} n_{i,1} n_{i,2}$; $S(\cdot)$ is some polynomial whose coefficients are all positive.

Definition 3 (Poly-continuity). A function $f(x) : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^d$ satisfies polynomial generalized continuity if

$$\|f(x) - f(x')\| \leq S(\|x_{\max,1}\|, \dots, \|x_{\max,n_x}\|) \|x - x'\|, \quad \forall x, x' \in \Omega,$$

where $\|x_{\max,i}\| = \max\{\|x_i\|, \|x'_i\|\}$, and $S(\cdot)$ is some polynomial whose coefficients are all positive.

Definition 4 (Poly-smoothness). A function $f(x) : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^d$ in C^1 satisfies polynomial generalized smoothness if ∇f satisfies polynomial generalized continuity.

See Appendix C.2.1 for comparisons between poly-smoothness and generalized smoothness (Li et al., 2023). The three definitions are easily satisfied in typical neural network settings. Specifically:

Corollary 2. Any product, sum, and composition of polynomial, softmax, tanh, sigmoid, normalization τ_e , GELU, SiLU, satisfies polynomial boundedness for both the functions and their gradients, polynomial generalized continuity, and polynomial generalized smoothness.

The corollary above implies that many neural networks, including transformers, residual networks, and feedforward networks employing the aforementioned functions, satisfy the three properties.

3.3 LOCAL RELAXED DISSIPATIVE CONDITION

Besides poly-boundedness, continuity, and smoothness, additional conditions are needed to guarantee convergence. Inspired by some weak dissipative conditions from operator theory (for example, hypomonotonicity (Moudafi, 2004)), we define the following local relaxed dissipative condition:

Definition 5 ($(x, x^*, r, \rho, \epsilon)$ -dissipativity). A function $f(x) \in C^1$ satisfies the $(x, x^*, r, \rho, \epsilon)$ -dissipative condition near x if there exists some stationary point x^* and some constant $r > \|x - x^*\|$, s.t., $\forall y \in B_{x^*}(r) \setminus \{x : \|\nabla f(x)\| \leq \epsilon\}$, we have

$$\nabla f(y)^\top (y - x^*) \geq -\rho \|\nabla f(y)\|^2,$$

where $\epsilon \geq 0$, and $\rho \in \mathbb{R}$ is a constant depending on x, x^*, ϵ , but independent of y .

The above condition characterizes the quality of the local landscape near a stationary point by comparing the gradient direction and the stationary point direction. It holds trivially with $\rho = \frac{1}{\epsilon} \max \|y - x^*\| = \frac{r}{\epsilon}$ by Cauchy-Schwartz inequality. Note that ρ can be negative, and a function f with more favorable dissipative properties leads to a smaller ρ . For example, if f is locally convex near x^* with some $r > 0$, then $\rho = 0$.

4 THEORY

In this section, we consider the data $\{(x_i, y_i)\}_{i=1}^N$ where $x_i, y_i \in \mathbb{R}^d$, and define $h_{ij} = \|x_i - x_j\|$, and $h_{\min} = \min_{i,j} h_{ij}$. We mainly use $h_{\min}$ to characterize the effect of data distributions. More precisely, we show that an increase in $h_{\min}$ corresponds to a more uniform (less biased) distributions (Section 4.1). We then prove that a larger $h_{\min}$ leads to faster training (Section 4.2) and smaller approximation error (Section 4.3). Additionally, we present a convergence framework for a family of neural networks under GD beyond the NTK regime (Remark 1). This framework provides theoretical support for the practical effectiveness of structures such as residual connection and function composition (Remark 2).

We assume that the data satisfy the following properties:

Assumption 1 (Data). Assume $x_1, \dots, x_N \stackrel{i.i.d.}{\sim} \pi(x)dx$, where $\pi(x) \in L^\infty$ is supported on an open set $\Omega \subseteq \mathbb{R}^d$ with $|\Omega| < \infty$, and $\pi \ll \text{Leb}$ on Ω . Assume the ground truth function $g(x) \in W^{r,p}(\Omega)$ for $1 \leq p < \infty$ and $r \geq 1$, and $y_i = g(x_i) \in \mathbb{R}^d$.

In the above assumption, the input data are sampled i.i.d. from some density, which is absolutely continuous w.r.t. Lebesgue measure on its support, and lies in L^∞ to exclude degenerate cases like point masses. For the ground truth function g , we assume that it is relatively smooth.

Unlike analyses in the NTK regime (Allen-Zhu et al., 2019; Zou et al., 2020; Zou & Gu, 2019; Chen et al., 2020b; Oymak & Soltanolkotabi, 2020, etc.), which typically assume normalized input data so that $h_{\min}$ reflects only angular separation between points, we do not assume normalized input, and therefore, $h_{\min}$ encodes both the angular separation and norm difference between input data.

4.1 MINIMUM DISTANCE $h_{\min}$ VS MORE UNIFORM DISTRIBUTION

In this section, we discuss the relationship between $h_{\min}$ and the sampling density function of input data $\pi(\cdot)$: less biased density results in larger $h_{\min}$, and consequently admits faster convergence (Theorem 2) and smaller approximation error (Theorem 3).

Before stating the result, we introduce the following notations. Let $\pi_{\max} = \|\pi(x)\|_{\infty}$. For some $0 < \bar{\pi}_{\max} \leq \pi_{\max}$, we define $\Omega_{\max} = \{x \in \Omega \mid \pi(x) \geq \bar{\pi}_{\max}\}$. Below is the main theorem showing the relationship between $h_{\min}$ and $\pi_{\max}, \bar{\pi}_{\max}$.

Theorem 1. *Suppose Assumption 1 holds. For any biased distribution such that there exists a ball $B\left(C\left(\frac{-\log \delta}{\pi_{\max}(N-1)V_d}\right)^{1/d}\right) \subseteq \Omega_{\max}$, and $\frac{\pi_{\max}}{\bar{\pi}_{\max}} = \mathcal{O}(1)$, we have with probability at least $1 - 2\delta$,*

$$\left(\frac{2\delta}{\pi_{\max}N(N-1)V_d}\right)^{1/d} \leq h_{\min} \leq C\left(\frac{-\log \delta}{\bar{\pi}_{\max}(N-1)V_d}\right)^{1/d},$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volume of d -dimensional unit ball, and $C > 0$ is some universal constant.

The above theorem provides both the lower and upper bounds on $h_{\min}$ in terms of $\pi_{\max}$ and $\bar{\pi}_{\max}$. The specific assumption on the biased distribution requires the density to remain high over a sufficiently large region, avoiding narrow peaks with near-zero support. When $\pi(\cdot)$ is biased, both $\pi_{\max}$ and $\bar{\pi}_{\max}$ can be large, leading to smaller upper and lower bounds for $h_{\min}$. Conversely, if $\pi(\cdot)$ becomes more “flattened”, then $\pi_{\max}$ decreases, and $\bar{\pi}_{\max}$ also tends to decrease when the flattening is significant. Thus, a more uniform (i.e., less biased) density yields a larger $h_{\min}$.

4.2 OPTIMIZATION: CONVERGENCE OF NEURAL NETWORKS BEYOND NTK REGIME

In this section, we establish the convergence of a broad class of neural networks under GD beyond NTK regime (Theorem 2). Moreover, smaller $h_{\min}$ tends to slow down convergence (Corollary 3).

Neural networks, Loss, and Algorithm. We consider a family of neural networks as follows:

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

where $\theta = (\theta_0, \dots, \theta_L)$ is the collection of weights with the θ_i corresponding to the i th layer. For the rest of the paper, we sometimes omit the subscript of θ for simplicity. We use the following 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 k th iteration.

We then introduce the following assumptions before stating the main theorem.

Assumption 2 (Real-analyticity). *Assume $\varphi_{\ell}(\theta; u)$ are real-analytic (Definition 1) for all θ, u , and $\ell = 1, \dots, L$, and $\varphi_0(\theta; u)$ is real-analytic for all θ, u in some bounded open neighborhood of $\bar{\Omega}$.*

Assumption 3 (Poly-boundedness, continuity, smoothness). *Assume for all $\ell = 0, \dots, L$, $\varphi_{\ell}(\theta; u)$ satisfies polynomial generalized continuity (Definition 3) and smoothness (Definition 4) for both θ and u . In addition, $\varphi_{\ell}(\theta; u), \nabla_{\theta} \varphi_{\ell}(\theta; u), \nabla_u \varphi_{\ell}(\theta; u)$ are polynomially bounded (Definition 2).*

Assumption 2 and 3 hold true for many practical structures, including transformers, feedforward networks, and residual networks. See more discussions in Section 3.1 and 3.2.

Assumption 4 (Nonlinear architecture). *Let $\tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell},i}) = f(\theta; x_i)$. Assume there exists some $\bar{\ell} \in \{0, \dots, L-1\}$, s.t., for any $\theta_{\bar{\ell}}, \dots, \theta_L$ except for a measure-zero set, there exists one tuple $(u_1, \dots, u_N)$,*

$$\text{s.t., } \begin{pmatrix} \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_N) \end{pmatrix} \text{ has full row rank } Nd, \text{ and } \dim \theta_{\bar{\ell}} > Nd.$$

The above assumption is mild, and guarantees nonlinearity of the architecture through Theorem 7 (Mityagin, 2015). The $\hat{f}_\ell(\theta; u_{\ell,i})$ is the composition of layers from the $(\ell + 1)$ th to the L th layer. It requires the row spans of the Jacobian of $\hat{f}_\ell(\theta; u_i)$'s do not intersect. The following is one example.

Lemma 1. *The following structure satisfies Assumption 4 with $\ell = L - 1$: $\varphi_L(\theta; u) = \theta_L u$ and $\varphi_{L-1}(\theta; u) = \theta_{L-1,2}\sigma(\theta_{L-1,1}u)$, where $\sigma(\cdot)$ is GELU, $\theta_L \in \mathbb{R}^{d \times d}$, $\theta_{L-1,1} \in \mathbb{R}^{m_{L-1} \times d}$, $\theta_{L-1,2} \in \mathbb{R}^{d \times m_{L-1}}$, and $m_{L-1} > Nd$.*

The above lemma considers an architecture in which the $(L - 1)$ th layer is a two-layer feedforward network, and the final layer is linear. It further shows that the width $m_{L-1} > Nd$ is sufficient to satisfy Assumption 4. Indeed, this lower bound of m_{L-1} can potentially be improved in the proof of Lemma 1, since Assumption 4 only requires $m_{L-1} > N/2$. See Appendix C.3 for detailed proofs.

Below is the main theorem showing the convergence of neural networks (1) under GD.

Theorem 2. *Suppose Assumptions 1 to 4 hold. Consider the initialization θ^0 in $\mathbb{R}^{\dim \theta}$ except for a measure-zero set. Let $R = \sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho+2}{\delta}\mathcal{L}(\theta^0)}$ for some stationary point θ^* . Let η be in some dense set of $(0, \min\{\frac{2-\delta}{L}, 1\})$ for some small constant $0 < \delta < 2$, and $L = S(\cdots, \|\theta_i^*\| + R + \max_{\theta \in B_{\theta^*}(R)} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \cdots)$, where $S(\cdot)$ is some polynomial whose coefficients are all positive. Assume $\mathcal{L}(\theta)$ satisfies the $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition in Definition 5. Then, under some arbitrarily small adjustment on the scale of φ_ℓ for $\ell = 0, \dots, L - 1$, with probability 1 over the joint distribution of the input data $x_1, \dots, x_N$, GD converges to $\|\nabla \mathcal{L}(\theta)\| \leq \epsilon_{\mathcal{L}}$, and*

$$\mathcal{L}(\theta^k) \leq \prod_{s=0}^{k-1} \left(1 - \eta \left(1 - \frac{\eta L}{2}\right) \frac{\mu_{\text{low},s,X}}{N}\right) \mathcal{L}(\theta^0), \quad \forall k \geq 1, \text{ and } \|\nabla \mathcal{L}(\theta^k)\| > \epsilon_{\mathcal{L}},$$

where $\mu_{\text{low},s,X} > 0$ depends on θ^s and $x_1, \dots, x_N$.

In the above theorem, the loss of GD is proved to decay until it converges to the neighborhood of a stationary point, under practical assumptions. A complete version of the above theorem can be found in Theorem 5. Please see Appendix E for detailed proofs.

The L is the Lipschitz smoothness constant along the GD trajectory, and is proved to be bounded instead of assumed. It depends on the poly-smoothness function $S(\cdot)$ of $\mathcal{L}(\theta)$ (Lemma 7), the initial condition θ^0 , initial loss $\mathcal{L}(\theta^0)$, maximum $\|\nabla \mathcal{L}(\theta)\|$ in some bounded region, and some stationary point θ^* and its landscape nearby determined solely by θ^0 . The θ^* can be any stationary point that satisfies the dissipativity condition (Definition 5), and is not necessarily the limit of GD iterations. The radius R of the bounded region for local relaxed dissipativity is determined by the initialization θ^0 and does not assume any information about the subsequent GD iterations. The constant $\mu_{\text{low},s,X}$ is the lower frame bound of the derivatives of $(f(x_1)^\top, \dots, f(x_N)^\top)^\top$ which is proven to form a frame with probability 1 for almost all θ , and consequently $\mu_{\text{low},s,X} > 0$ (see more details in Appendix E.1).

The unusual density requirement on the learning rate η arises from the parametric transversality theorem (Theorem 6 (Hirsch, 2012), Lemma 11). The excluded “bad” values of η may cause convergence to a degenerate map with non-full-rank Jacobian, reducing expressivity by restricting the image to a lower-dimensional subspace. Fortunately, such cases can be avoided by a small perturbation of η , which suffices to yield a converging η^1 . The small adjustment to the scale of φ_ℓ follows the same reasoning to prevent degeneracy, and only needs to be done once before GD iterations.

Based on the convergence of the neural network, we derive the following corollary, which highlights the relationship between data pairwise distance, particularly $h_{\min}$, and the convergence rate.

Corollary 3. *Under the same assumptions as Theorem 2, fix some x_i . Let $X = (x_1, \dots, x_N)$, and $\mathcal{D}_{i,H} = \{j \mid h_{ij} = \|x_i - x_j\| \leq H, \forall j \neq i\}$. Then for any $k \geq 1$, there exists $r_{k,i,H} > 0$ and $L_{k,i,H} > 0$, s.t., when $\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2} \leq r_{k,i,H}$, we have*

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

¹The set of η in Theorem 2 is residual, i.e., roughly a “large” set in $(0, \min\{\frac{2-\delta}{L}, 1\})$, compared to its complement (see Appendix E.3).

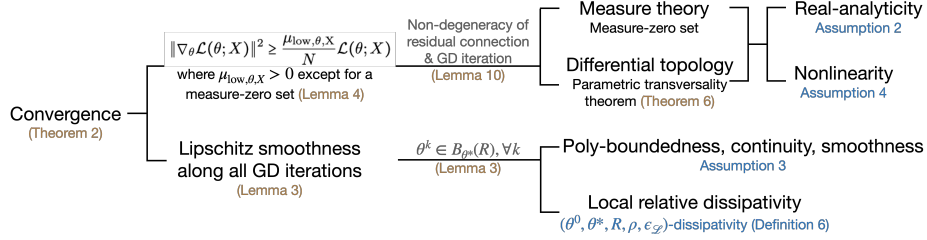


Figure 2: Proof sketch of the convergence framework beyond the NTK regime. Assumptions are highlighted in blue, key lemmas in brown, and the statements or implication of lemmas in grey.

Specifically, if $H = h_{\min}$ and $\mathcal{D}_{i,H} \neq \emptyset$, there exists $L_{k,i}, r_{k,i} > 0$, s.t., when $h_{\min} \leq r_{k,i}$, we have $\mu_{\text{low},s,X} \leq L_{k,i} h_{\min}$, $\forall s \leq k$.

The above corollary 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 5. 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 μ_{low,s,X_t} also monotonically decreases.

Consequently, the value of $1 - \eta \left(1 - \frac{\eta L}{2}\right) \frac{\mu_{\text{low},s,X_t}}{N}$ increases, and thus leads to slower convergence. The dependence on $h_{\min}$ is a special case of this phenomenon and follows the same reasoning.

Remark 1 (Convergence beyond the NTK regime). *Theorem 2 provides a general framework of convergence for a family of neural networks beyond the NTK regime in the following two senses, both of which concern the local relaxed dissipativity and the learning rate. (1) When the loss landscape is unfavorable around the initial condition θ^0 (e.g. $\rho = \frac{R}{\epsilon_{\mathcal{L}}}$ which holds for any landscape; see Section 3.3), which may indicate insufficient overparameterization, this setting naturally lies outside the NTK regime. (2) When the loss landscape is favorable near θ^0 , i.e., ρ is small, this may imply a large width of the neural network. However, the learning rate η in Theorem 2 can be much larger than the infinitesimal (gradient flow) regime assumed in the NTK analysis, due to a small ρ and the independence of width if θ^0 is properly scaled. This means that the NTK regime can be covered by our analysis, i.e., when the neural network is in the NTK regime, our analysis then guarantees convergence to a neighborhood of a global minimum. See Appendix B for more detailed discussions.*

Remark 2 (Theoretical insights on architectures: residual connection, and function composition). *When constructing a neural network, it is important to ensure that the model is expressive enough. Theoretically, this corresponds to requiring non-degeneracy of the neural network f , i.e., its Jacobian ∇f should be full rank, avoiding loss of full expressivity. In fact, non-degeneracy is a generic property among real-analytic functions. If a function is degenerate, we can typically restore non-degeneracy by adding a small identity component, i.e., by incorporating a residual connection (see Lemma 11). This helps explain the empirical success of residual architectures in practice. Moreover, since neural networks are compositions of layers, if each layer has a Jacobian with full rank (e.g., via residual connections), then by chain rule the Jacobian of the overall composition remains full rank, preserving non-degeneracy. See Appendix B for more detailed discussions.*

The proof sketch of Theorem 2 is shown in Figure 2. The proof of Corollary 3 mainly relies on the local Lipschitz continuity of eigenvalues for real-analytic matrix functions (see Theorem 4.1 in Kurdyka & Paunescu (2008)). Detailed proofs are in Appendix E.

4.3 APPROXIMATION: DATA-DEPENDENT ERROR IN THE INTERPOLATION REGIME

In this section, we show that a larger $h_{\min}$ and smaller local maximal distances lead to smaller approximation error in the interpolation regime $\mathcal{I} = \text{conv}\{x_1, \dots, x_N\}$.

The following theorem characterizes the data-dependent approximation error.

Theorem 3. *Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$ under Assumption 1. Let $\mathcal{I}_i = \text{conv}\{x_{i_1}, \dots, x_{i_{d+1}}\}$ with $x_j \notin \mathcal{I}_i, \forall j \neq i_1, \dots, i_{d+1}$, and define $h_{\max_{d+1}} = \max_i \max_{x_q, x_j \in \mathcal{I}_i} h_{qj}$. Assume $h_{\min} \leq 1 \leq h_{\max_{d+1}}$. Suppose there is a family of neural networks that can approximate polynomials of degree $n_d = n_d(N, d)$, i.e., for any $\epsilon_2 > 0$ and $\psi \in P_{n_d}$, there exists some f s.t. $\|f - \psi\|_p \leq \epsilon_2$. Then, for some $C_1 > 0$ and $m > \max\{r, n_d\}$, there exists some $\phi \in W^{m,p}(\Omega)$ satisfying $\|g - \phi\|_{r,p} \leq C_2 h_{\max_{d+1}}^m h_{\min}^{-r} \|\phi\|_{m,p,\mathcal{I}^0}$, s.t., with probability 1 over the joint distribution of $x_1, \dots, x_N$, we have*

$$\|f - g\|_{p,\mathcal{I}^0} \leq C_1 h_{\max_{d+1}}^m h_{\min}^{-r} \|\phi\|_{m,p,\mathcal{I}^0}.$$

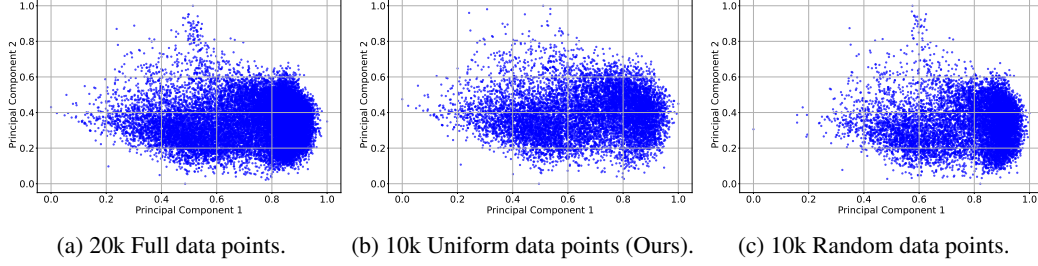


Figure 3: Visualization of data point distributions for datasets selected using different methods from WizardLM (Xu et al., 2024).

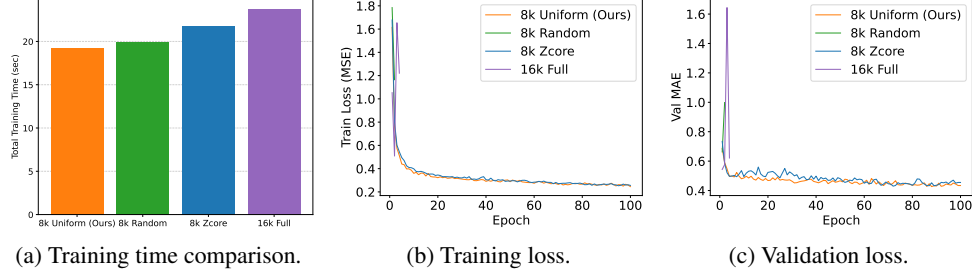


Figure 4: L2-SGD training on the California Housing dataset (Pace & Barry, 1997). Training time is measured as the total time required to complete the same number of epochs (e.g., 100). Uniform subsets (ours) achieve faster, smoother, and more stable convergence with fewer samples, matching or surpassing Zcore (Griffin et al., 2024) and mitigating gradient instabilities seen in random and full datasets.

The above theorem differs from classical results in approximation theory. It fixes a specific approximation function and examines how the distances between data points affect its error. More precisely, the error bound is determined mainly by two quantities: a local maximum distance $h_{\max_{d+1}}$, and a minimum distance $h_{\min}$. It then indicates that using fewer but more uniformly distributed data points, i.e., increasing $h_{\min}^2$, 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 5).

The proof mainly involves establishing a data-dependent Bramble-Hilbert lemma (Bramble & Hilbert, 1970) for each d -simplex $\mathcal{T}_i$ in the Delaunay triangulation. See a formal version of Theorem 3 and detailed proofs in Appendix F.

5 EXPERIMENTS

In our experiments, we introduce a data uniformity strategy that iteratively selects points maximizing the minimum pairwise distance $h_{\min}$ (Algorithm 1), and visualize differences between uniform subsets, random subsets of equal size, and the full datasets. Comparison to random subsets isolates the effect of $h_{\min}$ while holding dataset size fixed. We then perform supervised fine-tuning with two circumstances: ℓ^2 loss with SGD for theoretical validation and cross-entropy loss with Adam for practical evaluation. This design enables systematic assessment of performance and training efficiency across datasets, scales, and model sizes. We further deploy our method and compare with SOTA baselines, including TeaMs-RL (Gu et al., 2024), WizardLM (Xu et al., 2024), Zcore (Griffin et al., 2024), and LESS (Xia et al., 2024). Results show that uniform subsets accelerate convergence (Corollary 3) and maintain comparable or superior performance relative to both full datasets and strong baselines (Theorem 3). Full experimental settings appear in Appendix J.

Algorithm 1: Data uniformity strategy

Input: Full dataset $\{x_i\}_{i=1}^N$, target size $\tilde{N}$
Output: Subset S with $|S| = \tilde{N}$

- 1: // initialize selected set
- 2: $i \leftarrow 1$; randomly pick x^* from $\{x_i\}_{i=1}^N$
- 3: $S \leftarrow \{x^*\}$
- 4: **while** $i < \tilde{N}$ **do**
- 5: // maximize min distance
- 6: $\tilde{x}_i \leftarrow \arg \max_{x \in \{x_j\}_{j=1}^N \setminus S} \min_{y \in S} \text{dist}(x, y)$
- 7: $S \leftarrow S \cup \{\tilde{x}_i\}$
- 8: $i \leftarrow i + 1$
- 9: **end while**

²Larger $h_{\min}$ also contributes to smoother approximation as r specifies the order of smoothness in $W^{r,p}(\Omega)$.

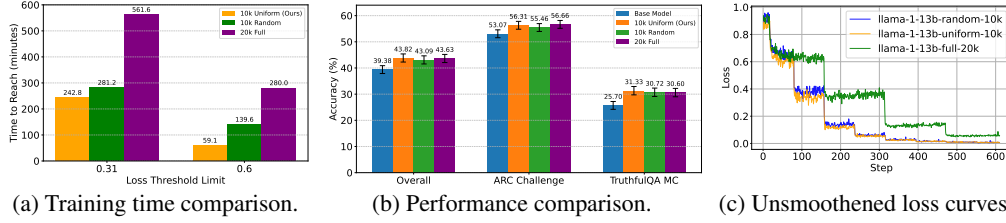


Figure 5: Cross-entropy training with AdamW on the WizardLM dataset (Xu et al., 2024) using LLaMA-1-13B (Touvron et al., 2023). Uniform subsets (ours, via ℓ^2 distance) (a) reach target loss thresholds in shorter run time, (b) achieve comparable or superior ARC Challenge and TruthfulQA MC accuracies, and (c) have faster loss decay, showing efficiency of uniformity-aware selection.

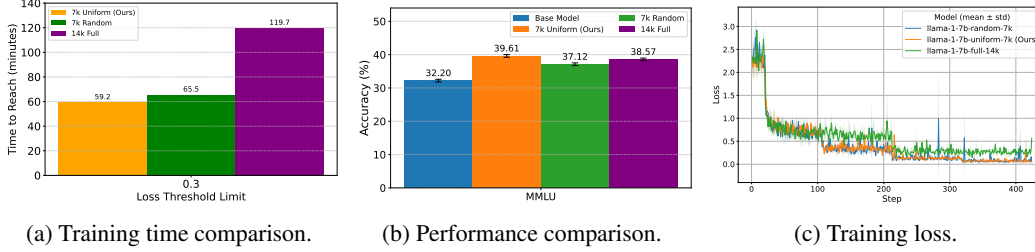


Figure 6: Cross-entropy training with AdamW on the LESS dataset (Xia et al., 2024) using LLaMA-1-7B (Touvron et al., 2023). Uniform subsets (ours, via cosine distance) (a) reach target loss thresholds faster, (b) achieve comparable or superior MMLU accuracy, and (c) show smoother, more stable convergence than random or full datasets, demonstrating efficiency of uniformity-aware selection.

5.1 DATA SELECTION AND DATA VISUALIZATION

We propose a greedy distance-maximization strategy to build compact yet representative instruction-tuning datasets (Algorithm 1; see also Tan et al. (2023) for a related use of uniformity not using $h_{\min}$, and Ethayarajh (2019) for discussions on uniformity). Each data point is encoded via Word2Vec embeddings, and points are iteratively chosen to maximize the minimum Euclidean or cosine distance, ensuring broad coverage. Applied to TeaMs-RL (Gu et al., 2024), WizardLM (Xu et al., 2024), Zcore (Griffin et al., 2024), and LESS (Xia et al., 2024), we construct balanced subsets (e.g., 4.5k from 9k TeaMs-RL, 10k from 20k WizardLM) for fair comparison with random baselines. PCA visualizations (Maćkiewicz & Ratajczak, 1993) reveal that uniform subsets achieve broader, less redundant coverage than random or full datasets, highlighting their effectiveness in preserving semantic diversity with fewer samples (see Appendix H.1).

5.2 ℓ^2 -SGD TRAINING EXPERIMENTS

We validate our theoretical insights under ℓ^2 loss with SGD on the California Housing dataset (Pace & Barry, 1997), comparing our uniform subsets with random, Zcore (Griffin et al., 2024), and full datasets. Zcore is an SOTA coreset selection method that leverages foundation model embeddings to compute importance scores, producing compact and representative subsets that have been shown to outperform several strong baselines (Griffin et al., 2024). As shown in Figure 4, random and full datasets exhibit divergence, uniform subsets achieve comparable or better performance than Zcore with shorter training time, measured as the total time to complete the same number of epochs (e.g., 100), while avoiding gradient instabilities. Consistent results across other datasets, model sizes, and baselines such as TeaMs-RL and WizardLM (Appendix H.2) confirm the efficiency of uniformity-based selection, supporting theory (Section 4) and demonstrating improved efficiency and stability.

5.3 CROSS ENTROPY LOSS ADAM TRAINING EXPERIMENTS

Beyond ℓ^2 -SGD, we also conduct experiments using cross-entropy loss with Adam, the standard optimization setup for LLMs (Gunel et al., 2021; Mao et al., 2023), and observe consistent improvements in the two settings: LLaMA-13B on the WizardLM dataset (Figure 1 where data is selected using cosine distance, and 5 with Euclidean distance) and LLaMA-1-7B on the LESS dataset (Xia et al., 2024) (Figure 6). The LESS dataset was originally constructed via gradient-based selection for MMLU (Xia et al., 2024), and we further reapply our method on this preselected data. Results show that uniformity-aware selection can further enhance the performance of LESS. Figure 1(a), 5(a),

and 6(a) demonstrates that the 10k and 7k uniform subsets reach the loss threshold more than $2\times$ faster than the corresponding full datasets, and also faster than the corresponding random subsets of the same sizes, improving training efficiency. Figure 1(b), 5(b), and 6(b) show that, at matched loss checkpoints (ensuring fair evaluation across methods), the uniform subset achieves accuracy comparable to or exceeding random and full datasets while requiring less training time. Additionally, Figure 6 illustrates smoother and more stable convergence (less spikes) across seeds. Consistent findings are also observed on TeaMs-RL and WizardLM across different dataset scales, data selection distances (e.g., cosine distance), model sizes, and sources (see Appendix H.3).

6 DISCUSSION AND CONCLUSION

We remark that using different distances in Algorithm 1 offer distinct benefits on WizardLM. The training speed of using cosine distance is faster than using Euclidean distance (Figure 1(a) vs Figure 5(a)), while the generalization of using Euclidean distance is better than cosine distance (Figure 5(b) vs (Figure 1(b))). We will leave it for future exploration. In sum, we demonstrate the importance of data uniformity, selecting more uniformly distributed data, through both theory and experiments. Theoretically, we show that higher uniformity, i.e., larger minimum pairwise distance $h_{\min}$, can accelerate training and reduce approximation error. We also develop a general convergence framework for complicated neural networks beyond the NTK regime, and it also highlights how residual connection and function composition preserve expressivity. Empirically, across optimizers (ℓ^2 -SGD, Adam), model sizes (LLaMA-1 7B, 13B), and datasets (TeaMs-RL, WizardLM, ZCore, LESS), uniform selection consistently accelerates convergence and achieves accuracy comparable to or better than full datasets with fewer training time, improving both training loss and downstream tasks such as ARC and TruthfulQA. These findings emphasize data uniformity as an effective strategy for efficient and scalable LLM training.

REFERENCES

- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 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.
- Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.
- Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.
- Paul Eric Ammann and John C Knight. Data diversity: An approach to software fault tolerance. *Ieee transactions on computers*, 37(4):418–425, 2002.
- 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.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.
- Yang Ba, Michelle V Mancenido, and Rong Pan. How does data diversity shape the weight landscape of neural networks? *arXiv preprint arXiv:2410.14602*, 2024.
- Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

- Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36:9707–9750, 2023.
- Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.
- 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.
- Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. On feature learning in neural networks with global convergence guarantees. *arXiv preprint arXiv:2204.10782*, 2022.
- Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33: 13363–13373, 2020a.
- Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representations*, 2020b.
- Zixiang Chen, Greg Yang, Qingyue Zhao, and Quanquan Gu. Global convergence and rich feature learning in l -layer infinite-width neural networks under μp parametrization. *arXiv preprint arXiv:2503.09565*, 2025.
- Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.
- 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.
- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. *arXiv preprint arXiv:1205.2653*, 2012.
- Mark De Berg. *Computational geometry: algorithms and applications*. Springer Science & Business Media, 2000.
- Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.
- Yijun Dong, Viet Hoang Phan, Xiang Pan, and Qi Lei. Sketchy moment matching: Toward fast and provable data selection for finetuning. *Advances in Neural Information Processing Systems*, 37: 43367–43402, 2024.
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

- 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.
- Etienne Emmrich. *Discrete versions of Gronwall’s lemma and their application to the numerical analysis of parabolic problems*. Techn. Univ., 1999.
- 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>.
- Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=SJx63jRqFm>.
- Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pp. 1887–1936. PMLR, 2021.
- 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.
- Hans Grauert. On levi’s problem and the imbedding of real-analytic manifolds. *Annals of Mathematics*, 68(2):460–472, 1958.
- 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>.
- Qiyang Han and Masaaki Imaizumi. Precise gradient descent training dynamics for finite-width multi-layer neural networks. *arXiv preprint arXiv:2505.04898*, 2025.
- Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.
- Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.
- Morris W Hirsch. *Differential topology*, volume 33. Springer Science & Business Media, 2012.
- Johannes Hofmanninger, Florian 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.
- Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in neural information processing systems*, 33:2698–2709, 2020.

- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.
- 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.
- 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.
- David Krieg and Mathias Sonnleitner. Random points are optimal for the approximation of sobolev functions. *IMA Journal of Numerical Analysis*, 44(3):1346–1371, 2024.
- David Krieg, Erich Novak, and Mathias Sonnleitner. Recovery of sobolev functions restricted to iid sampling. *Mathematics of Computation*, 91(338):2715–2738, 2022.
- Krzysztof Kurdyka and Laurentiu Paunescu. Hyperbolic polynomials and multiparameter real-analytic perturbation theory. 2008.
- Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.
- Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023.
- 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. *arXiv preprint arXiv:2109.07958*, 2021.
- Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023.
- Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. *Advances in neural information processing systems*, 32, 2019.
- Jianfeng Lu, Zuwei 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.
- Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

- Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 1-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.
- 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.
- Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.
- Abdellatif Moudafi. A remark on the convergence of the tikhonov regularization without monotonicity. *MATHEMATICAL INEQUALITIES AND APPLICATIONS*, 7:283–288, 2004.
- 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.
- Anh Nguyen-Tuong, David Evans, John C Knight, Benjamin Cox, and Jack W Davidson. Security through redundant data diversity. In *2008 IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN)*, pp. 187–196. IEEE, 2008.
- 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.
- Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.
- R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.
- Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. *arXiv preprint arXiv:2006.08859*, 2020.
- Zhen Qin, Jinxin Zhou, and Zhihui Zhu. On the convergence of gradient descent on learning transformers with residual connections. *arXiv preprint arXiv:2506.05249*, 2025.
- Trout Rader. Nice demand functions. *Econometrica: Journal of the Econometric Society*, pp. 913–935, 1973.
- Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.
- 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.
- Kamron Saniee. A simple expression for multivariate lagrange interpolation. *SIAM undergraduate research online*, 1(1):1–9, 2008.
- Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

- 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.
- Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2021.
- Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.
- 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.
- Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.
- Paulo Tabuada and Bahman Ghahsifard. 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.
- Matus Telgarsky. Feature selection with gradient descent on two-layer networks in low-rotation regimes. *arXiv preprint arXiv:2208.02789*, 2022.
- 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, 2022a.
- 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, 2024.
- 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, 2023a.
- 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.
- Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022b. URL <https://openreview.net/forum?id=3tbDrs77LJ5>.
- Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *arXiv preprint arXiv:2310.17087*, 2023b.
- David S Watkins. A generalization of the bramble-hilbert lemma and applications to multivariate interpolation. *Journal of approximation theory*, 26(3):219–231, 1979.
- Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

- 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>.
- Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.
- 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.
- Wing Ho Yuen, Roy D Yates, and S-C Mau. Exploiting data diversity and multiuser diversity in non-cooperative mobile infostation networks. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 3, pp. 2218–2228. IEEE, 2003.
- 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.
- Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.
- Difan Zou and Quanguan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019.
- Difan Zou, Yuan Cao, Dongruo Zhou, and Quanguan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

APPENDIX

Contents

A	More related work	18
B	More discussions on the convergence results	18
C	Preliminary properties and assumption justification	19
C.1	Analytic function	19
C.2	Polynomial boundedness, polynomial generalized continuity and smoothness	20
C.2.1	Comparison between poly-smoothness and generalized smoothness	21
C.3	Example of Assumption 4	22
D	Minimum distance and biased distribution: proof of Theorem 1	23
E	Optimization: proof of Theorem 2 and Corollary 3	25
E.1	Lower bound of gradient	30
E.2	Polynomial generalized smoothness of the loss	32
E.3	Additional lemmas in measure theory and differential topology	37
E.3.1	Basic definitions and theorems	37
E.3.2	Supplementary lemmas of the proof	38
F	Approximation: proof of Theorem 3	42
F.1	Definitions and lemmas related to computational geometry	42
F.2	Data-dependent polynomial approximation in Sobolev space	43
G	Additional lemmas	45
H	Extended Experiments	45
H.1	Data Selection and Visualization	45
H.2	ℓ^2 -SGD Training Experiments	46
H.3	Cross Entropy Loss Adam Training Experiments	47
H.4	Performance and Training Time Comparison	48
I	Supplementary Experiments	50
I.1	Additional Experiments Using ℓ^2 -SGD	50
I.2	Effectiveness of Maximized-Distance Based Sampling on Smaller Scale Datasets.	50
I.3	Effectiveness of Maximized-Distance Based Sampling on Larger Scale Datasets.	51
I.4	Full Step Training: 7b and 13b models	51
I.5	Ablation Experiment	52
J	Experiment Settings	55

A MORE RELATED WORK

Data Diversity. Data diversity has been explored across a wide range of domains for enhancing generalization, efficiency, and robustness. For example, Muennighoff et al. (2025) investigated the role of data diversity in improving test-time scaling generalization. In WizardLM (Xu et al., 2024), diverse queries are obtained by sampling inputs randomly for distillation from larger LLMs. TeaMs-RL (Gu et al., 2024) leveraged reinforcement learning (RL) to teach LLMs for diverse instructional data generalization, improving downstream model generalization. Beyond language modeling, data diversity has shown benefits for generalization in supervised learning (Yu et al., 2022), software security (Nguyen-Tuong et al., 2008), fault tolerance in software systems (Ammann & Knight, 2002), and medical image segmentation (Hofmanninger et al., 2020). Particularly, in RL, diverse data supports more efficient offline training (Nguyen-Tang & Arora, 2023) and improves exploration and policy learning (Eysenbach et al., 2019). Other domains, such as mobile infostation networks (Yuen et al., 2003) and privacy protection (Machanavajjhala et al., 2007), also leverage diversity for improved robustness. More broadly, Gong et al. (2019) demonstrated that data diversity enhances machine learning performance across various stages, such as model diversification and inference diversification.

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

Residual Connections. Residual connections have been widely studied for their theoretical and practical benefits in deep learning. Hardt & Ma (2016) showed that deep linear residual networks exhibit no spurious local optima and maintain strong expressivity, a result further supported by Liu et al. (2019), who demonstrated the absence of spurious minima in residual architectures. Huang et al. (2020) highlighted the role of residual connections in preserving learnability across depth, while Scholkemper et al. (2024) found that they help mitigate oversmoothing in graph neural networks. In the context of transformers, Qin et al. (2025) showed that residual connections improve the conditioning of output matrices in single-layer architectures with feedforward network and attention. In contrast to these works, our study analyzes general and more complex architectures and provides a theoretical explanation based on measure theory and differential topology for the intrinsic non-degeneracy of the layerwise mapping induced by residual connections throughout training.

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 $1 \leq p \leq \infty$, there are works studying, for example, transformers 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 & Ghahesifard (2022), feedforward networks (Hanin & Sellke, 2017; Kidger & Lyons, 2020; Park et al., 2020; Cai, 2022, etc.).

B MORE DISCUSSIONS ON THE CONVERGENCE RESULTS

Below are more detailed versions of Remark 1 and 2.

Remark 3 (Convergence beyond the NTK regime). *Theorem 2 provides a general framework of convergence for a family of neural networks beyond the NTK regime in the following two senses, both of which concern the local relaxed dissipativity and the learning rate:*

(1) *When the loss landscape is unfavorable around the initial condition (for example, in the worst case $\rho = \frac{R}{\epsilon}$ for some $\epsilon > 0$, which holds for any landscape; see Section 3.3), it indicates insufficient overparameterization, meaning the width of the neural network is too small, and therefore lies outside the NTK regime. Moreover, a large ρ leads to large L , which may imply small learning rate η . However, different from the NTK regime, η does not necessarily decrease as the width increases, since it depends mainly on the initialization, and with proper scaling, it can be independent of width.*

(2) *When the loss landscape is favorable near the initial condition, i.e., ρ is small, this may imply a large width of the neural network. However, the learning rate η in Theorem 2 can be much larger than that required by the infinitesimal (gradient flow) regime assumed in the NTK analysis, due to a small ρ and the independence of width as discussed in (1).*

Apart from the local relaxed dissipativity condition, the only width requirement appears in Assumption 4. As an example of the Assumption 4, Lemma 1 shows that it suffices for the width of the $(L - 1)$ th feedforward layer to be greater than Nd ; no constraints are imposed on the widths of other layers. This lower bound Nd on a single layer, which can potentially be improved (see more discussions below Lemma 1), is much milder than the NTK-type convergence conditions, which typically require the widths of all layers to scale as high-degree polynomials in d, N, L , etc. (see Section 2). We also remark that the NTK regime is covered by our analysis. More precisely, if the width is large enough, i.e., the neural network is in the NTK regime, our analysis then guarantees convergence to a neighborhood of a global minimum.

Remark 4 (Theoretical insights on architectures: residual connection, composition, and normalization). *When constructing a neural network, it is important to ensure that the model is expressive enough. Theoretically, this corresponds to requiring the neural network mapping f to be non-degenerate, i.e., its Jacobian ∇f should be full rank in some region. Otherwise, the mapping may collapse onto a lower-dimensional subspace, and thus lose full expressivity.*

In fact, non-degeneracy is a generic property among real-analytic functions. If a function is degenerate, we can typically restore non-degeneracy by adding a small identity component, i.e., by incorporating a residual connection, and then the modified mapping is non-degenerate (see Lemma 11). This helps explain the empirical success of residual architectures in practice.

Moreover, neural networks are constructed as compositions of layer-wise transformations. If each layer has a Jacobian with full rank (e.g., ensured via residual connections), then the Jacobian of the overall composition remains full rank due to the chain rule: $\nabla(f \circ g)(x) = \nabla f(g(x))\nabla g(x)$. This shows that the compositional (iterative) structure of neural networks preserves non-degeneracy. In contrast, other combinations, such as summations of functions, do not necessarily preserve this property, even if each individual function is non-degenerate. For example, consider $f_1(x) = x^\top x$ and $f_2(x) = -x^\top x$; while each is non-degenerate, their sum $f_1 + f_2 = 0$ is degenerate.

Regarding the normalization used in (1), it differs from practical schemes such as batch normalization and layer normalization. Nonetheless, they share a common goal, which is to reduce the magnitude of neurons and restore desirable regularity properties in the network. Such normalization helps flatten the optimization landscape, thereby facilitating more stable and efficient training. See more discussions in Wang et al. (2023b).

C PRELIMINARY PROPERTIES AND ASSUMPTION JUSTIFICATION

C.1 ANALYTIC FUNCTION

Real-analyticity is preserved under many operations. Specifically:

Proposition 1. *Real-analytic functions have the following properties:*

1. *The sums, products, divisions where the denominators are not zero, and compositions of real-analytic functions are real-analytic.*
2. *The derivative and integral of real-analytic functions are real-analytic.*

3. Real-analytic function is C^∞ .

We then prove that a lot of neural network architectures are real-analytic.

Proof of Corollary 1. Since exponential function admits Taylor series in some neighbourhood of any point in $\mathbb{R}^d$, and the denominators of softmax and sigmoid functions are strictly positive at any point, by Proposition 1, we have that softmax, sigmoid, GELU, SiLU are analytic functions on $\mathbb{R}^d$.

Similarly, polynomial functions admit Taylor series in some neighbourhood of any point in $\mathbb{R}^d$, and the denominator of τ_ϵ is strictly positive at any point, we have τ_ϵ and polynomial functions are analytic.

Also, tanh can be Taylor expanded at any point and thus analytic.

By Proposition 1, any product, sum, compositions of the above functions are still analytic. $\square$

C.2 POLYNOMIAL BOUNDEDNESS, POLYNOMIAL GENERALIZED CONTINUITY AND SMOOTHNESS

Polynomial boundedness, polynomial generalized continuity and smoothness are preserved under the following operations:

Proposition 2. Assume f and g satisfies polynomial boundedness for both the functions and their gradients, polynomial generalized continuity, and polynomial generalized smoothness. Then

1. $f(x) + g(x)$, where $f, g : \mathbb{R}^d \rightarrow \mathbb{R}^m$,
2. $f(x)g(x)$, where $f \in \mathbb{R}^{m \times d}$, and $g \in \mathbb{R}^{d \times n}$,
3. $(f \circ g)(x)$, where $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$, and $g : \mathbb{R}^n \rightarrow \mathbb{R}^d$,

satisfy the same properties under different polynomials.

Proof. Let $S_{f,0}, S_{g,0}$ be the corresponding polynomials in polynomial generalized continuity of f and g , $S_{f,1}, S_{g,1}$ be the ones in polynomial generalized smoothness, S_f, S_g be the corresponding polynomials in polynomial bound, and $S_{\nabla f}, S_{\nabla g}$ be the corresponding polynomials in polynomial bound of their gradients. For notational simplicity, we use $S(x)$ to denote $S(\|x_1\|, \dots, \|x_d\|)$, and $S(x_{\max})$ to denote $S(\|x_{\max,1}\|, \dots, \|x_{\max,d}\|)$.

Then

$$\|f(x) + g(x) - f(x') - g(x')\| \leq \|f(x) - f(x')\| + \|g(x) - g(x')\| \leq (S_{f,0} + S_{g,0})(x_{\max})\|x - x'\|.$$

Similarly, $\nabla f + \nabla g$ follows the same idea with $(S_{f,1} + S_{g,1})(x_{\max})$. Moreover, $f + g$ is upper bounded by $S_f(x_{\max}) + S_g(x_{\max})$, and $\nabla f + \nabla g$ is upper bounded by $S_{\nabla f}(x_{\max}) + S_{\nabla g}(x_{\max})$.

Also,

$$\begin{aligned} \|f(x)g(x) - f(x')g(x')\| &= \|f(x)g(x) - f(x)g(x') + f(x)g(x') - f(x')g(x')\| \\ &\leq \|f(x)\| \|g(x) - g(x')\| + \|f(x) - f(x')\| \|g(x')\| \\ &\leq (S_f(x)S_{g,0}(x_{\max}) + S_{f,0}(x_{\max})S_g(x'))\|x - x'\| \\ &\leq (S_f(x_{\max})S_{g,0}(x_{\max}) + S_{f,0}(x_{\max})S_g(x_{\max}))\|x - x'\|. \end{aligned}$$

Obviously, $(S_f(x_{\max})S_{g,0}(x_{\max}) + S_{f,0}(x_{\max})S_g(x_{\max}))$ is a polynomial with positive coefficients, and therefore fg satisfies polynomial generalized continuity.

Based on the above derivation, $\nabla(fg) = \nabla f \cdot g + f \nabla g$ is a sum of two product, and thus satisfies polynomial continuity with $(S_{\nabla f}S_{g,0} + S_{f,1}S_g + S_{\nabla g}S_{f,0} + S_{g,1}S_f)$.

Also, fg is upper bounded by S_fS_g , and $\nabla(fg) = \nabla f \cdot g + f \nabla g$ is upper bounded by $S_{\nabla f}S_g + S_{\nabla g}S_f$.

Next, since all the S have positive coefficients, we have

$$\begin{aligned} \|f(g(x)) - f(g(x'))\| &\leq S_{f,0}(\dots, \|g(x)_{\max,i}\|, \dots)\|g(x) - g(x')\| \\ &\leq S_{f,0}(S_g(x_{\max}), \dots, S_g(x_{\max}))S_{g,0}(x_{\max})\|x - x'\|. \end{aligned}$$

Also, $\nabla_x f(g(x)) = \nabla_g f(g(x)) \nabla g(x)$. Similar to the above derivations, we have $\nabla_x f(g(x))$ is polynomial continuous with $S_{f,1}(S_g(x_{\max}), \dots, S_g(x_{\max})) S_{g,0}(x_{\max}) S_{\nabla g}(x_{\max}) + S_{\nabla f}(S_f(x_{\max}), \dots, S_f(x_{\max})) S_{g,1}(x_{\max})$.

In the end, it can be shown that $f(g(x))$ is upper bounded by $S_f(S_g(x_{\max}), \dots, S_g(x_{\max}))$, and $\nabla f(g(x))$ is upper bounded by $S_{\nabla g}(x_{\max}) S_{\nabla f}(S_g(x_{\max}), \dots, S_g(x_{\max}))$. $\square$

Corollary 4. *The functions, polynomials, softmax, tanh, sigmoid, τ_ϵ , GELU, SiLU, and the product, sum, compositions of them, satisfies polynomial boundedness for both the functions and their gradients, polynomial generalized continuity, and polynomial generalized smoothness.*

Proof. By Proposition 2, we only need to show that polynomials, softmax, tanh, sigmoid, τ_ϵ , GELU, SiLU satisfy all the properties.

First, polynomials obviously satisfies all the properties. It can be shown by simple calculations that softmax, tanh, sigmoid, τ_ϵ , and their derivatives are all upper bounded by constant, and therefore satisfies all the above properties.

For SiLU, it is the product of x and sigmoid, and therefore by Proposition 2 satisfies all the above properties.

For GELU, $\sigma(x) = x\Phi(x)$, where $\Phi(x) = \int_{-\infty}^x \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds$. Also, $\Phi'(x) = \phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. Then $\Phi(x)$ and $\Phi'(x)$ are both upper bounded by constants, and thus satisfies all the above properties. By Proposition 2, GELU, the product of x and $\Phi(x)$, also satisfies these properties. $\square$

Similar to Lipschitz smoothness and generalized smoothness (Li et al., 2023), the polynomial generalized smoothness has another interpretation as follows:

Lemma 2. *If the function $f(x) \in C^1$ satisfies the polynomial generalized smoothness, then the following inequalities hold*

$$f(x') \leq f(x) + \nabla f(x)^\top (x' - x) + \frac{S(\|x_{\max,1}\|, \dots, \|x_{\max,d}\|)}{2} \|x - x'\|^2 \quad (2)$$

Proof. Since $f(x) \in C^1$ satisfies the polynomial generalized smoothness, we have

$$\begin{aligned} \|\nabla f(x) - \nabla f((1-t)x + tx')\| &\leq S(\dots, \max\{\|x_i\|, \|(1-t)x_i + tx'_i\|, \dots\}) \|x - ((1-t)x + tx')\| \\ &\leq t S(\dots, \max\{\|x_i\|, \|x'_i\|\}, \dots) \|x - x'\| \\ &= t S(\|x_{\max,1}\|, \dots, \|x_{\max,d}\|) \|x - x'\| \end{aligned}$$

Then

$$\begin{aligned} f(x') - f(x) &= \int_0^1 \nabla f((1-t)x + tx')^\top (x' - x) dt \\ &= \int_0^1 \nabla f(x)^\top (x' - x) dt + \int_0^1 (\nabla f((1-t)x + tx') - \nabla f(x))^\top (x' - x) dt \\ &\leq \nabla f(x)^\top (x' - x) + \|x - x'\| \int_0^1 \|\nabla f((1-t)x + tx') - \nabla f(x)\| dt \\ &\leq \nabla f(x)^\top (x' - x) + S(\|x_{\max,1}\|, \dots, \|x_{\max,d}\|) \|x - x'\|^2 \int_0^1 t dt \\ &= \nabla f(x)^\top (x' - x) + \frac{S(\|x_{\max,1}\|, \dots, \|x_{\max,d}\|)}{2} \|x - x'\|^2 \end{aligned}$$

$\square$

C.2.1 COMPARISON BETWEEN POLY-SMOOTHNESS AND GENERALIZED SMOOTHNESS

The differences between the above poly-smoothness and generalized smoothness (Li et al., 2023) lies in two parts. (1) Li et al. (2023) defined the bound of Hessian first, and therefore evaluate the smoothness function at one fixed point for any two points in its neighbourhood, while we directly consider the continuity of the gradient and evaluate the smoothness function $S(\cdot)$ at the two points. (2) Regarding the smoothness function, Li et al. (2023) employed a function of $\|\nabla f(x)\|$, while we use a polynomial of $\|x_i\|, \|x'_i\|$. Note such $S(\cdot)$ is monotonically increasing in $\mathbb{R}_{\geq 0} \times \dots \times \mathbb{R}_{\geq 0}$.

C.3 EXAMPLE OF ASSUMPTION 4

Assumption 4 is used in the convergence of neural networks. Before introducing the convergence proof, we first verify that the Assumption 4 can be easily satisfied:

Proof of Lemma 1. First, we compute the Jacobian of the neural network w.r.t. $\theta_{L-1,1}$

$$\begin{aligned} & \nabla_{\theta_{L-1,1}} f(\theta; x_i) \\ &= \nabla_{\tau_\epsilon} \varphi_L(\tau_\epsilon(u_{L,i})) \nabla_u \tau_\epsilon(u_{L,i}) \nabla_{\theta_{L-1,2}} \bar{\varphi}(u_{L-1,i}) \\ &= \theta_L \underbrace{\left(\frac{1}{\sqrt{\|u_{L,i}\|^2 + \epsilon^2}} I - \frac{1}{(\sqrt{\|u_{L,i}\|^2 + \epsilon^2})^3} u_{L,i} u_{L,i}^\top \right)}_{P_i} \underbrace{\begin{pmatrix} \sigma(\theta_{L-1,1} u_{L-1,i})^\top & & \\ & \ddots & \\ & & \sigma(\theta_{L-1,1} u_{L-1,i})^\top \end{pmatrix}}_{Q_i} \end{aligned}$$

where in $\nabla_{\theta_{L-1,2}} \bar{\varphi}(u_{L-1,i})$, we vectorize $\theta_{L-1,2}$ in row.

Also, $\sigma(x) = x\Phi(x)$, where $\Phi(x) = \int_{-\infty}^x \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds$. Let $\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. Then we have

$$\begin{aligned} \sigma'(x) &= \Phi(x) + x\phi(x) \\ \sigma^{(n)}(x) &= 2\phi^{(n-2)}(x) + \phi^{(n-1)}(x) \\ &= (2(-1)^{n-2}H_{n-2}(x) + (-1)^{n-1}H_{n-1}(x))\phi(x) \end{aligned}$$

where $H_n(x)$ is the probabilist's Hermite polynomial.

By Taylor expansion,

$$\sigma(v^\top u + a) = \sum_{n=0}^{\infty} \frac{\sigma^{(n)}(a)}{n!} (v^\top u)^n$$

Consider $\sigma(t\theta u)$ near $t = 0$. Choose $0 < t_1 < t_2 < \dots < t_N \ll 1$. For probabilist's Hermite polynomials,

$$H_{2n}(0) \neq 0, H_{2n-1}(0) = 0, \forall n \geq 1.$$

Thus,

$$\sigma(0) = 0, \sigma^{(n)}(0) \neq 0, \forall n \geq 1.$$

Let $\theta_{L-1,1,j}$ be the j th row of $\theta_{L-1,1}$. Then

$$\begin{pmatrix} \sigma(t_1 \theta_{L-1,1,j} u) \\ \vdots \\ \sigma(t_N \theta_{L-1,1,j} u) \end{pmatrix} = \begin{pmatrix} \sum_{n=0}^{\infty} \frac{\sigma^{(n)}(0)}{n!} (t_1 \theta_{L-1,1,j} u)^n \\ \vdots \\ \sum_{n=0}^{\infty} \frac{\sigma^{(n)}(0)}{n!} (t_N \theta_{L-1,1,j} u)^n \end{pmatrix} = \sum_{n=1}^N \frac{\sigma^{(n)}(0)}{n!} (\theta_{L-1,1,j} u)^n \begin{pmatrix} t_1^n \\ \vdots \\ t_N^n \end{pmatrix} + \mathcal{O} \left(\begin{pmatrix} t_1^{N+1} \\ \vdots \\ t_N^{N+1} \end{pmatrix} \right)$$

Note

$$\det \begin{pmatrix} t_1 & \dots & t_1^N \\ \vdots & & \vdots \\ t_N & \dots & t_N^N \end{pmatrix} = \prod_{i=1}^N t_i \prod_{1 \leq j < k \leq N} (t_k - t_j) > 0.$$

Thus the N vectors $\begin{pmatrix} t_1^n \\ \vdots \\ t_N^n \end{pmatrix}$ are linearly independent.

We then choose u s.t. $\theta_{L-1,1} u$ has at least N distinct non-zero elements. We claim that the set of all $\theta_{L-1,1}$ such that there is no u s.t. $\theta_{L-1,1} u$ has at least N distinct non-zero elements, is measure-zero in $\mathbb{R}^{|\theta_{L-1,1}|}$. We denote the set to be $\mathcal{B}_{\theta_{L-1,1}}$.

Consider $v \in \mathbb{R}^m$ in the column space of $\theta_{L-1,1}$. Fix a partition of m elements into at most $N-1$ blocks, where in each block of v is constant. Such set of vectors is of dimensional $N-1$. Moreover, the number of such partition is finite. Therefore, $\text{Leb}_{|\theta_{L-1,1}|}(\mathcal{B}_{\theta_{L-1,1}}) = 0$.

WOLG, assume the first N elements are distinct. Then, similarly, we have that the N vectors $\begin{pmatrix} (\theta_{L-1,1,1}u)^n \\ \vdots \\ (\theta_{L-1,1,N}u)^n \end{pmatrix}$ are linearly independent. Thus

$$\begin{aligned} & \det \begin{pmatrix} \frac{\sigma^{(1)}(0)}{1!}(\theta_{L-1,1,1}u) & \cdots & \frac{\sigma^{(N)}(0)}{N!}(\theta_{L-1,1,1}u)^N \\ \vdots & & \vdots \\ \frac{\sigma^{(1)}(0)}{1!}(\theta_{L-1,1,N}u) & \cdots & \frac{\sigma^{(N)}(0)}{N!}(\theta_{L-1,1,N}u)^N \end{pmatrix} \\ &= \prod_{i=1}^N \frac{\sigma^{(i)}(0)}{i!} \det \begin{pmatrix} (\theta_{L-1,1,1}u) & \cdots & (\theta_{L-1,1,1}u)^N \\ \vdots & & \vdots \\ (\theta_{L-1,1,N}u) & \cdots & (\theta_{L-1,1,N}u)^N \end{pmatrix} \neq 0. \end{aligned}$$

Therefore, $\begin{pmatrix} \sigma(t_1\theta_{L-1,1}u)^\top \\ \vdots \\ \sigma(t_N\theta_{L-1,1}u)^\top \end{pmatrix}$ is full rank N , and then $\begin{pmatrix} Q_1 \\ \vdots \\ Q_N \end{pmatrix}$ is full rank Nd .

Also, since P_i is invertible for any $u_{L,i}$, we have $\begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_N \end{pmatrix}$ is invertible. Thus

$$\begin{pmatrix} \nabla_{\tau_\epsilon} \varphi_L(\tau_\epsilon(u_{L,1})) \nabla_u \tau_\epsilon(u_{L,1}) \nabla_{\theta_{L-1,2}} \bar{\varphi}(u_{L-1,1}) \\ \vdots \\ \nabla_{\tau_\epsilon} \varphi_L(\tau_\epsilon(u_{L,N})) \nabla_u \tau_\epsilon(u_{L,N}) \nabla_{\theta_{L-1,2}} \bar{\varphi}(u_{L-1,N}) \end{pmatrix} = \begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_N \end{pmatrix} \begin{pmatrix} Q_1 \\ \vdots \\ Q_N \end{pmatrix}$$

is full rank Nd , where $u_{L-1,i} = t_i u$. $\square$

D MINIMUM DISTANCE AND BIASED DISTRIBUTION: PROOF OF THEOREM 1

We now present the complete version of Theorem 1.

Theorem 4. Suppose Assumption 1 holds.

1. For any biased distribution such that there exists a ball $B\left(C\left(\frac{-\log \delta}{\pi_{\max}(N-1)V_d}\right)^{1/d}\right) \subseteq \Omega_{\max}$, and $\frac{\pi_{\max}}{\pi_{\min}} = \mathcal{O}(1)$, we have with probability at least $1 - 2\delta$,

$$\left(\frac{2\delta}{\pi_{\max}N(N-1)V_d}\right)^{1/d} \leq h_{\min} \leq C\left(\frac{-\log \delta}{\pi_{\max}(N-1)V_d}\right)^{1/d}$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volumn of d -dimensional unit ball, and $C > 0$ is some universal constant.

2. For general distribution, we have with probability at least $1 - 2\delta$,

$$\left(\frac{2\delta}{\pi_{\max}N(N-1)V_d}\right)^{1/d} \leq h_{\min} \leq \left(\frac{-\log \delta}{\pi_{\min}(N-1)V_d}\right)^{1/d}.$$

Proof. First fix any point $x_i = x$. For this proof, we consider some small enough $r \in (0, 1)$, and then

$$\pi_{\min} V_d r^d \leq \mathbb{P}(\|x_j - x_i\| < r | x_i = x) = \int_{B(x,r)} \pi(z) dz \leq \pi_{\max} V_d r^d \quad (3)$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volumn of d -dimensional unit ball.

We would like to lower bound

$$\begin{aligned} \mathbb{P}(h_{\min} \geq r) &= \mathbb{P}(\|x_i - x_j\| \geq r, \forall 1 \leq i < j \leq N) \\ &= \mathbb{E}[\mathbb{1}_{\{\|x_i - x_j\| \geq r, \forall 1 \leq i < j \leq N\}}] \\ &= \mathbb{E}[\mathbb{1}_{\{\|x_1 - x_2\| \geq r\}} \mathbb{1}_{\{\|x_3 - x_i\| \geq r, \forall i=1,2\}} \cdots \mathbb{1}_{\{\|x_N - x_i\| \geq r, \forall i=1, \dots, N-1\}}] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\|x_1 - x_2\| \geq r\}} | x_1] \cdot \mathbb{E}[\mathbb{1}_{\{\|x_3 - x_i\| \geq r, \forall i=1,2\}} | x_1, x_2] \cdots \mathbb{E}[\mathbb{1}_{\{\|x_N - x_i\| \geq r, \forall i=1, \dots, N-1\}} | x_1, \dots, x_{N-1}]] \end{aligned}$$

where the last inequality follows from the properties of conditional expectation.

Also, we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\{\|x_1 - x_2\| \geq r\}} | x_1] &= \mathbb{P}(\|x_1 - x_2\| \geq r | x_1) \\ &\geq 1 - \pi_{\max} V_d r^d\end{aligned}$$

where the inequality follows from (3), and similarly,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\{\|x_s - x_i\| \geq r, \forall i=1, \dots, s-1\}} | x_1, \dots, x_{s-1}] &= \mathbb{P}(\|x_s - x_i\| \geq r, \forall i=1, \dots, s-1 | x_1, \dots, x_{s-1}) \\ &\geq 1 - (s-1)\pi_{\max} V_d r^d.\end{aligned}$$

Therefore, for some small r , we have

$$\begin{aligned}\mathbb{P}(h_{\min} \geq r) &\geq \mathbb{E}\left[\prod_{i=1}^{N-1} (1 - i\pi_{\max} V_d r^d)\right] = \prod_{i=1}^{N-1} (1 - i\pi_{\max} V_d r^d) \\ &\geq 1 - \sum_{i=1}^{N-1} i\pi_{\max} V_d r^d = 1 - \frac{N(N-1)}{2} \pi_{\max} V_d r^d.\end{aligned}$$

Thus with probability at least $1 - \delta$, we have

$$h_{\min} \geq \left(\frac{2\delta}{\pi_{\max} N(N-1) V_d}\right)^{1/d}.$$

For the other side, for biased distribution satisfying the assumptions stated in the theorem, we consider

$$r \leq 2C \left(\frac{-\log \delta}{\bar{\pi}_{\max} (N-1) V_d}\right)^{1/d}, \text{ and } \Omega_{\max, B} = B\left(\frac{r}{2}\right) \subseteq \Omega_{\max}.$$

Then,

$$\begin{aligned}\mathbb{P}(h_{\min} \leq r) &= \mathbb{P}(\exists i, j, s.t. \|x_i - x_j\| \leq r) \\ &\geq \mathbb{P}(\exists \text{ at least two points } x_i, x_j \in \Omega_{\max, B}, i.e., \|x_i - x_j\| \leq r) \\ &= 1 - \mathbb{P}(x_i \notin \Omega_{\max, B}, \forall i) - \mathbb{P}(\exists i, s.t. x_i \in \Omega_{\max, B}, x_j \notin \Omega_{\max, B}, \forall j \neq i) \\ &\geq 1 - \left((1 - \bar{\pi}_{\max} |\Omega_{\max, B}|)^N + N(1 - \bar{\pi}_{\max} |\Omega_{\max, B}|)^{N-1} \pi_{\max} |\Omega_{\max, B}|\right) \\ &= 1 - (1 - \bar{\pi}_{\max} |\Omega_{\max, B}|)^{N-1} (1 - \bar{\pi}_{\max} |\Omega_{\max, B}| + N\pi_{\max} |\Omega_{\max, B}|) \\ &\geq 1 - e^{-(N-1)\bar{\pi}_{\max} |\Omega_{\max, B}|} (1 - \bar{\pi}_{\max} |\Omega_{\max, B}| + N\pi_{\max} |\Omega_{\max, B}|)\end{aligned}$$

where the last inequality follows from $(1-x)^N \leq e^{-Nx}$.

Then, we have with probability at least $1 - \delta$,

$$h_{\min} \leq C \left(\frac{-\log \delta}{\bar{\pi}_{\max} (N-1) V_d}\right)^{1/d},$$

where this inequality follows from $|\Omega_{\max, B}| = V_d \left(\frac{r}{2}\right)^d$ and $C > 0$ is some universal constant.

The last conclusion follows from taking union bound of the above two results.

For general distribution,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\{\|x_1 - x_2\| \geq r\}} | x_1] &= \mathbb{P}(\|x_1 - x_2\| \geq r | x_1) \\ &\leq 1 - \pi_{\min} V_d r^d\end{aligned}$$

where the inequality follows from (3), and similarly,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\{\|x_s - x_i\| \geq r, \forall i=1, \dots, s-1\}} | x_1, \dots, x_{s-1}] &= \mathbb{P}(\|x_s - x_i\| \geq r, \forall i=1, \dots, s-1 | x_1, \dots, x_{s-1}) \\ &\leq 1 - \pi_{\min} V_d r^d.\end{aligned}$$

Therefore, for some small r , we have

$$\mathbb{P}(h_{\min} \geq r) \leq (1 - \pi_{\min} V_d r^d)^{N-1}$$

Thus with probability at least $1 - \delta$, we have

$$h_{\min} < \left(\frac{1 - \delta^{\frac{1}{N-1}}}{\pi_{\min} V_d} \right)^{1/d} \leq \left(\frac{-\log \delta}{\pi_{\min} V_d (N-1)} \right)^{1/d}$$

where the last inequality follows from $1 - e^{-x} \leq x$ for $x > 0$.

□

E OPTIMIZATION: PROOF OF THEOREM 2 AND COROLLARY 3

We recall that the architecture of neural network is defined as follows

$$\begin{aligned} u_{0,i} &= x_i \\ u_{\ell+1,i} &= u_{\ell,i} + \varphi_{\ell}(\theta; u_{\ell,i}), \ell = 0, \dots, L-1 \\ f(x_i) &= \varphi_L(\theta; u_{L,i}). \end{aligned}$$

The loss function of each data point is

$$l(f(x_i), y_i) = \frac{1}{2} \|y_i - f(x_i)\|^2$$

and the total loss is

$$\mathcal{L}(f(x_i), y_i) = \frac{1}{N} \sum_{i=1}^N l(f(x_i), y_i).$$

We then compute the Jacobian of the following maps that will be used in the proof:

$$\nabla_f l(x_i) = (f(x_i) - y_i)^\top$$

$$\nabla_{\theta_j} f(x_i) = \sum_{\ell \in \mathcal{J}_j} \nabla_u \bar{\varphi}_L(u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_{\ell}(u_{\ell,i}; \theta)$$

where $\nabla_u \bar{\varphi}_{\ell}(u_{\ell})$ is the Jacobian of the map $\bar{\varphi}_{\ell}(u_{\ell})$ w.r.t. u_{ℓ} , $\nabla_{\theta_{\ell}} f(x_i)$ is also the Jacobian; the set $\mathcal{J}_j$ denotes the indices of layers in which θ_j appears.

Then

$$\begin{aligned} \nabla_{\theta_j} l(x_i) &= \nabla_f l(x_i) \nabla_{\theta_{\ell}} f(x_i) \\ &= \sum_{\ell \in \mathcal{J}_j} \nabla_f l(x_i) \nabla_u \bar{\varphi}_L(u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_{\ell}(u_{\ell,i}; \theta) \end{aligned}$$

and we have

$$\nabla_{\theta_j} \mathcal{L} = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta_j} l(x_i).$$

As shown in equation (1), we consider the case where the weights at each layer are distinct, i.e., $\mathcal{J}_j = \{j\}$.

In this section, we also define the following function for any matrix function $M(x) \in \mathbb{R}^{n \times m}$ with $m \geq n$

$$\det_r(M(x)) = \sum_{i=1}^{\binom{m}{n}} (\det M_i(x))^2,$$

where $M_i(x) \in \mathbb{R}^{n \times n}$ is the matrix with n columns of $M(x)$. Then $\det_r(M(x)) \neq 0$ if and only if $M(x)$ is full rank at x .

The following is a more complete version of the convergence result in Theorem 2, combining Corollary 3. We also specify that the dense set of η is indeed residual (see more details in Appendix E.3).

Theorem 5. Suppose Assumptions 1 to 4 hold. Consider the initialization θ^0 in $\mathbb{R}^{\dim \theta}$ except for a measure-zero set. Then, under some arbitrarily small adjustment on the scale of φ_ℓ for $\ell = 0, \dots, L-1$, with probability 1 over the joint distribution $\pi \times \dots \times \pi$ of the input data $x_1, \dots, x_N$, we have the following results:

1. When η is in some residual set of $(0, \min\{\frac{2-\delta}{L_1}, 1\})$, where

$$L_1 = S\left(\dots, \|\theta_i^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)} + \eta \max_{\|\theta\| \leq \|\theta^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)}} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \dots\right)$$

for some polynomial $S(\cdot)$ with positive coefficients, we have

$$\mathcal{L}(\theta^k) \leq \prod_{s=0}^{k-1} \left(1 - \eta \left(1 - \frac{\eta L_1}{2}\right) \frac{\mu_{\text{low},s,X}}{N}\right) \mathcal{L}(\theta^0), \quad \forall k \leq T = \frac{C}{\eta},$$

where $\mu_{\text{low},s,X} > 0$ is some strictly positive constant depending on θ^s and $x_1, \dots, x_N$, and $C > 0$ is universal constant.

2. Furthermore, let $R = \sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho+2}{\delta} \mathcal{L}(\theta^0)}$ for some stationary point θ^* . Assume $\mathcal{L}(\theta)$ satisfies the $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition in Definition 5. When η is in some residual set of $(0, \min\{\frac{2-\delta}{L_2}, 1\})$, where

$$L_2 = S\left(\dots, \|\theta_i^*\| + R + \max_{\theta \in B_{\theta^*}(R)} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \dots\right),$$

GD will converge to $\|\nabla \mathcal{L}(\theta)\| \leq \epsilon_{\mathcal{L}}$, and

$$\mathcal{L}(\theta^k) \leq \prod_{s=0}^{k-1} \left(1 - \eta \left(1 - \frac{\eta L_2}{2}\right) \frac{\mu_{\text{low},s,X}}{N}\right) \mathcal{L}(\theta^0), \quad \forall k \geq 1 \text{ and } \|\nabla \mathcal{L}(\theta^k)\| > \epsilon_{\mathcal{L}}.$$

3. Fix x_i , and let $\mathcal{D}_{i,H} = \{j \mid h_{ij} = \|x_i - x_j\| \leq H, \forall j \neq i\}$. Then for both the above cases, for any $k \geq 1$, there exists $r_{k,i,H} > 0$ and $L_{k,i,H} > 0$, s.t., when $\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2} \leq r_{k,i,H}$, we have

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

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

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

Proof. Define the GD iteration map to be

$$\Psi(\theta) = \theta - \eta \nabla_{\theta} \mathcal{L}(\theta), \text{ and } \Psi^k(\theta) = \underbrace{\Psi \circ \dots \circ \Psi}_{k}(\theta).$$

First, consider fixed $x_1, \dots, x_N$. Let

$$\mathcal{B}_{\theta, \text{rank}} = \{\theta : \nabla_{\theta} \Psi(\theta) \text{ is not full rank}\}$$

By Lemma 11, we know that for $\eta \in \mathcal{D}_{\eta,0}$, where $\mathcal{D}_{\eta,0}$ is a residual set of $(0, \min\{\frac{2}{L}, 1\})$, we have

$$\text{Leb}_{\dim \theta}(\mathcal{B}_{\theta, \text{rank}}) = 0.$$

Next we would like to show that along the trajectory, i.e., for any k , the union of “bad” sets of the initial condition θ leading to the degeneracy of the map Ψ^k is a measure-zero set.

Now consider any measure-zero set $\mathcal{B}_{\theta,0}$ in $\mathbb{R}^{\dim \theta}$. By Theorem 8,

$$\text{Leb}_{\dim \theta}(\Psi^{-1}(\mathcal{B}_{\theta,0})) = 0.$$

We then would like to prove that $(\Psi^k)^{-1}(\mathcal{B}_{\theta,0})$ is measure-zero for any $k \geq 1$. Suppose $(\Psi^k)^{-1}(\mathcal{B}_{\theta,0})$ is measure-zero. We would like to show that $(\Psi^{k+1})^{-1}(\mathcal{B}_{\theta,0})$ is measure zero. This is obvious since $(\Psi^{k+1})^{-1}(\mathcal{B}_{\theta,0}) = \Psi^{-1}((\Psi^k)^{-1}(\mathcal{B}_{\theta,0}))$.

Let $\mathcal{B}_{\theta,0}$ be the union of all the bad measure-zero sets of θ in Assumption 4 and Lemmas 4, including $\mathcal{B}_{\theta,\text{rank}}$, and there are finite union of such sets, which implies $\text{Leb}_{\dim \theta}(\mathcal{B}_{\theta,0}) = 0$. Let

$$\mathcal{B}_\theta = \bigcup_{k=1}^{\infty} (\Psi^k)^{-1}(\mathcal{B}_{\theta,0}).$$

Then

$$\text{Leb}_{\dim \theta}(\mathcal{B}_\theta) = \text{Leb}_{\dim \theta} \left(\bigcup_{k=1}^{\infty} (\Psi^k)^{-1}(\mathcal{B}_{\theta,0}) \right) \leq \sum_{k=1}^{\infty} 0 = 0,$$

namely, the initial condition set of θ where there exists some iteration k s.t. $\theta^k \in \mathcal{B}_{\theta,0}$, is measure-zero. In the rest of the proof, we just consider $\theta^0 \in \mathbb{R}^{\dim \theta} \setminus \mathcal{B}_\theta$.

Next we consider the bad set of $(x_1, \dots, x_N)$. By Lemma 4, for any θ^0 , the bad set of $(x_1, \dots, x_N)$, denoted as $\mathcal{B}_{X,0}$, is measure-zero,

$$\text{Leb}_{Nd}(\mathcal{B}_{X,0}) = 0.$$

Also, there are at most countable θ^k in the GD trajectory, and thus all such bad sets $\bigcup_{k=0}^{\infty} \mathcal{B}_{X,0}$ satisfy

$$\text{Leb}_{Nd} \left(\bigcup_{k=0}^{\infty} \mathcal{B}_{X,k} \right) \leq \sum_{k=0}^{\infty} 0 = 0.$$

Since $(x_1, \dots, x_N) \sim \pi \times \dots \times \pi$, and $\pi \ll \text{Leb}$, we have

$$(\pi \times \dots \times \pi)(\mathcal{B}_X) = 0,$$

namely, with probability 0, $(x_1, \dots, x_N)$ will be in this set.

Then, consider some initial conditions θ^0 except for a measure-zero set, and consider some dataset $\{(x_i, y_i)\}_{i=1}^N$ chosen with probability 1 over the joint distribution. By Lemma 7 and Lemma 2, we have for $i = 1, 2$,

$$\begin{aligned} \mathcal{L}(\theta^{k+1}) &\leq \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^\top (\theta^{k+1} - \theta^k) + \frac{S_k}{2} \|\theta^k - \theta^{k+1}\|^2 \\ &= \mathcal{L}(\theta^k) - \eta \left(1 - \frac{\eta S_k}{2} \right) \|\nabla \mathcal{L}(\theta^k)\|^2 \\ &\leq \left(1 - \eta \left(1 - \frac{\eta S_k}{2} \right) \frac{\mu_{\text{low},k,X}}{N} \right) \mathcal{L}(\theta^k) \\ &\leq \left(1 - \eta \left(1 - \frac{\eta L_i}{2} \right) \frac{\mu_{\text{low},k,X}}{N} \right) \mathcal{L}(\theta^k) \\ &\leq \prod_{s=0}^k \left(1 - \eta \left(1 - \frac{\eta L_i}{2} \right) \frac{\mu_{\text{low},s,X}}{N} \right) \mathcal{L}(\theta^0) \end{aligned}$$

where the second inequality follows from Lemma 4, the third inequality follows from Lemma 3 and $S(a_1, \dots, a_{n_\theta}) \leq L$ by the monotonicity of $S(\cdot)$ in $\mathbb{R}_{\geq 0} \times \dots \times \mathbb{R}_{\geq 0}$; $\mu_{\text{low},k,X} > 0$ is some strictly positive constant depending on θ^k and $x_1, \dots, x_N$.

By Lemma 3, when $i = 1$, the above inequality holds for $k + 1 \leq T = \frac{C}{\eta}$. When $i = 2$ with $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition, it holds for any $k \geq 1$.

Then, under $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition, GD will converge to the region $\|\nabla \mathcal{L}(\theta)\| \leq \epsilon_{\mathcal{L}}$. If not, we have $\mu_{\text{low},k,X} > 0$ because by Assumption 4 and Lemma 4, the θ s.t. $\mu_{\text{low},\theta,X} = 0$ lies in some measure-zero set and we remove any finite-step convergence to this set at the beginning of this proof. Thus $\left(1 - \eta \left(1 - \frac{\eta L_2}{2} \right) \frac{\mu_{\text{low},k,X}}{N} \right) < 1$, which means $\mathcal{L}(\theta^k)$ will keep decreasing. Contradiction.

The last statement follows from Lemma 4, and take $L_{k,i,H} = \max_{s \leq k} L_{s,i,H}$ and $r_{k,i,H} = \min_{s \leq k} r_{s,i,H}$.

□

The following lemma shows that the loss function is Lipschitz smooth along the GD iteration.

Lemma 3. Assume $\mathcal{L}(\theta)$ satisfies polynomial generalized smoothness.

1. Without $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition, let $\eta \leq \min\{\frac{2-\delta}{L_2}, 1\}$, where

$$L_2 = S \left(\dots, \|\theta_i^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)} + \max_{\|\theta\| \leq \|\theta^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)}} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \dots \right).$$

Then

$$S_k \leq L_2, \quad \forall k \leq T = \frac{C}{\eta}.$$

2. With $(\theta^0, \theta^*, R, \rho, \epsilon_{\mathcal{L}})$ -dissipative condition, let $\eta \leq \min\{\frac{2-\delta}{L_2}, 1\}$, where

$$L_2 = S \left(\dots, \|\theta_i^*\| + \sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho+2}{\delta} \mathcal{L}(\theta^0)} + \max_{\theta \in B_{\theta^*} \left(\sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho+2}{\delta} \mathcal{L}(\theta^0)} \right)} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \dots \right).$$

Then

$$S_k \leq L_2, \quad \forall k \geq 0 \text{ and } \|\nabla \mathcal{L}(\theta^k)\| \geq \epsilon_{\mathcal{L}}.$$

Proof. First, note

$$\max\{\|\theta_i^k\|, \|\theta_i^{k+1}\|\} \leq \|\theta_i^k\| + \eta \|\nabla_{\theta_i} \mathcal{L}(\theta^k)\|.$$

By the monotonicity of $S(\cdot)$, we only need to show $\|\theta_i^k\|$ is bounded under the two cases.

Without the dissipative assumption, we would like to show that

$$\|\theta_i^k\| \leq \|\theta_i^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)}, \quad \forall k \leq T = \frac{C}{\eta}.$$

Suppose the above inequality holds for all θ_i^j , where $j \leq k$ and $i = 1, \dots, n_{\theta}$. Then by $\eta \leq 1$, we have

$$\begin{aligned} S_k &\leq \bar{S}_k = S(\dots, \|\theta_i^k\| + \eta \|\nabla_{\theta_i} \mathcal{L}(\theta^k)\|, \dots) \\ &\leq L_2 = S(\dots, \|\theta_i^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)} + \max_{\|\theta\| \leq \|\theta^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)}} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \dots) \end{aligned}$$

namely,

$$\eta \leq \frac{2-\delta}{L_2} \leq \frac{2-\delta}{\bar{S}_k}.$$

By Lemma 7 and Lemma 2, we have

$$\begin{aligned} 0 &\leq \mathcal{L}(\theta^{k+1}) \leq \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^\top (\theta^{k+1} - \theta^k) + \frac{S_k}{2} \|\theta^k - \theta^{k+1}\|^2 \\ &= \mathcal{L}(\theta^k) - \eta \left(1 - \frac{\eta S_k}{2}\right) \|\nabla \mathcal{L}(\theta^k)\|^2 \\ &\leq \mathcal{L}(\theta^0) - \eta \sum_{j=0}^k \left(1 - \frac{\eta S_j}{2}\right) \|\nabla \mathcal{L}(\theta^j)\|^2 \\ &\leq \mathcal{L}(\theta^0) - \frac{\delta}{2} \eta \sum_{j=0}^k \|\nabla \mathcal{L}(\theta^j)\|^2, \end{aligned}$$

namely,

$$\sum_{j=0}^k \|\nabla \mathcal{L}(\theta^j)\|^2 \leq \frac{2}{\delta\eta} (\mathcal{L}(\theta^0) - \mathcal{L}(\theta^{k+1})) \leq \frac{2}{\delta\eta} \mathcal{L}(\theta^0). \quad (4)$$

Then

$$\begin{aligned} \|\theta_i^{k+1}\| &= \|\theta_i^k - \eta \nabla_{\theta_i} \mathcal{L}(\theta^k)\| = \left\| \theta_i^0 - \eta \sum_{j=0}^k \nabla_{\theta_i} \mathcal{L}(\theta^j) \right\| \\ &\leq \|\theta_i^0\| + \eta \sum_{j=0}^k \|\nabla \mathcal{L}(\theta^j)\| \\ &\leq \|\theta_i^0\| + \eta \sqrt{k \sum_{j=0}^k \|\nabla \mathcal{L}(\theta^j)\|^2} \\ &\leq \|\theta_i^0\| + \sqrt{\frac{2}{\delta} k \eta \mathcal{L}(\theta^0)} \\ &\leq \|\theta_i^0\| + \sqrt{\frac{2C}{\delta} \mathcal{L}(\theta^0)} \end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality, the third inequality follows from (4), and last inequality follows from $k \leq T = \frac{C}{\eta}$.

Then the above derivation for θ^k also holds for θ^{k+1} . Therefore $S_{k+1} \leq L_2$.

Next, under the dissipative condition, we would like to show that

$$\theta_i^k \in B_{\theta_i^*} \left(\sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho + 2}{\delta} \mathcal{L}(\theta^0)} \right), \text{ i.e., } \|\theta_i^k - \theta_i^*\|^2 \leq \|\theta^0 - \theta^*\|^2 + \frac{4\rho + 2}{\delta} \mathcal{L}(\theta^0)$$

Suppose the above inequality holds for all $j \leq k$. Then we have

$$\begin{aligned} S_k &\leq \bar{S}_k = S(\cdots, \|\theta_i^k\| + \eta \|\nabla_{\theta_i} \mathcal{L}(\theta^k)\|, \cdots) \\ &\leq L_2 = S(\cdots, \|\theta_i^*\| + \sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho + 2}{\delta} \mathcal{L}(\theta^0)} + \max_{\theta \in B_{\theta^*}(\sqrt{\|\theta^0 - \theta^*\|^2 + \frac{4\rho + 2}{\delta} \mathcal{L}(\theta^0)})} \|\nabla_{\theta_i} \mathcal{L}(\theta)\|, \cdots) \end{aligned}$$

namely,

$$\eta \leq \frac{2 - \delta}{L_2} \leq \frac{2 - \delta}{\bar{S}_k}.$$

Then, consider the $k + 1$ th iteration

$$\begin{aligned} \|\theta^{k+1} - \theta^*\|^2 &= \|\theta^k - \theta^* - \eta \nabla \mathcal{L}(\theta^k)\|^2 \\ &= \|\theta^k - \theta^*\|^2 - 2\eta \nabla \mathcal{L}(\theta^k)^\top (\theta^k - \theta^*) + \eta^2 \|\nabla \mathcal{L}(\theta^k)\|^2 \\ &\leq \|\theta^k - \theta^*\|^2 + \eta(2\rho + \eta) \|\nabla \mathcal{L}(\theta^k)\|^2 \\ &\leq \|\theta^0 - \theta^*\|^2 + \eta(2\rho + \eta) \sum_{j=0}^k \|\nabla \mathcal{L}(\theta^j)\|^2 \\ &\leq \|\theta^0 - \theta^*\|^2 + \frac{4\rho + 2\eta}{\delta} \mathcal{L}(\theta^0) \end{aligned}$$

where the first inequality follows from Definition 5, and the last inequality follows from (4), which also holds true when L_2 is replaced by L_2 .

Then together with $\eta \leq 1$ and

$$\|\theta_i^k\| \leq \|\theta_i^*\| + \|\theta_i^k - \theta_i^*\| \leq \|\theta_i^*\| + \|\theta^k - \theta^*\|,$$

we obtain $S_k \leq L_2$ for all k . □

E.1 LOWER BOUND OF GRADIENT

In this section, we use θ_i and $\theta_{\bar{\ell},i}$ to represent the scalar elements of θ and $\theta_{\bar{\ell}}$ respectively. The main lemma in this section is as follows.

Lemma 4. *Under Assumption 4, and some small adjustment of the scale of φ_ℓ for $\ell = 0, \dots, L-1$, given fixed θ except for a measure-zero set in $\mathbb{R}^{\dim \theta}$, and $X = (x_1, \dots, x_N)$ except for a measure-zero set in $\mathbb{R}^{Nd}$ that depends on θ , the gradient satisfies the following inequality*

$$\|\nabla_\theta \mathcal{L}(\theta; X)\|^2 \geq \frac{\mu_{\text{low},\theta,X}}{N} \mathcal{L}(\theta; X)$$

where $\mu_{\text{low},\theta,X} > 0$ is a strictly positive constant depending on θ and X .

Moreover, fix x_i , and let $\mathcal{D}_{i,H} = \{j \mid h_{ij} = \|x_i - x_j\| \leq H, \forall j \neq i\}$. Then there exists $r_{\theta,i,H} > 0$ and $L_{\theta,i,H} > 0$, s.t., when $\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2} \leq r_{\theta,i,H}$, we have

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

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

$$\mu_{\text{low},\theta,X} \leq L_{\theta,i} h_{\min}.$$

Proof. Consider

$$\begin{aligned} \|\nabla_\theta \mathcal{L}(\theta; X)\|^2 &= \left\| \frac{1}{N} \sum_{i=1}^N \nabla_f l(x_i) \nabla_\theta f(x_i) \right\|^2 \\ &\geq \frac{1}{N^2} \mu_{\text{low},\theta,X} \sum_{i=1}^N \|\nabla_f l(x_i)\|^2 = \frac{\mu_{\text{low},\theta,X}}{N} \mathcal{L}(\theta; X) \end{aligned}$$

where the first inequality follows from Lemma 6, and $\mu_{\text{low},\theta,X} > 0$ is a constant depending on θ and X . The second statement follows from Lemma 6. $\square$

Before presenting other lemmas, we first introduce the definition of a frame as follows.

Definition 6 (Frame). *The set of vectors $\{e_k\}$ in an inner-product vector space V is a frame of V if there exists constants $0 < \mu_{\text{low}} \leq \mu_{\text{up}} < \infty$, s.t.,*

$$\mu_{\text{low}} \|v\|^2 \leq \sum_k |\langle v, e_k \rangle|^2 \leq \mu_{\text{up}} \|v\|^2, \quad \forall v \in V.$$

The frame operator $T : V \rightarrow V$ is defined as

$$Tv = \sum_k \langle v, e_k \rangle e_k = \left(\sum_k e_k e_k^\top \right) v.$$

Then the gradient of the network f at all data points forms a frame in $\mathbb{R}^{Nd}$.

Lemma 5. *Under Assumption 4 and some small adjustment of the scale of φ_ℓ for $\ell = 0, \dots, L-1$, for any fixed θ except for a measure-zero set, $\left\{ \begin{pmatrix} \nabla_{\theta_i} f(x_1) \\ \vdots \\ \nabla_{\theta_i} f(x_N) \end{pmatrix} \right\}_{i=1}^{\dim \theta}$ form a frame except for a measure-zero set of $(x_1, \dots, x_N)$ in $\mathbb{R}^{Nd}$.*

Proof. By Assumption 4, there exists $\bar{\ell} \in \{0, \dots, L-1\}$ and $U = (u_1, \dots, u_N) \in \mathbb{R}^{Nd}$, θ , s.t.

$$\det_r \begin{pmatrix} \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_N) \end{pmatrix} \neq 0,$$

where $\tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell}}) = f(\theta; x)$.

Then by Theorem 7,

$$\text{Leb}_{Nd} \left(\left\{ U \in \mathbb{R}^{Nd} : \det_r \begin{pmatrix} \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_N) \end{pmatrix} = 0 \right\} \right) = 0,$$

namely,

$$\text{Leb}_{Nd} \left(\left\{ U \in \mathbb{R}^{Nd} : \begin{pmatrix} \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u_N) \end{pmatrix} \text{ is not full rank} \right\} \right) = 0$$

Next consider the map $U = U(X) = \begin{pmatrix} u_{\bar{\ell}}(x_1) \\ \vdots \\ u_{\bar{\ell}}(x_N) \end{pmatrix} : \mathbb{R}^{Nd} \rightarrow \mathbb{R}^{Nd}$. By Lemma 10,

$$\text{Leb}_{Nd} \left(\left\{ (x_1, \dots, x_N) \in \mathbb{R}^{Nd} : \begin{pmatrix} \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u(x_1)) \\ \vdots \\ \nabla_{\theta_{\bar{\ell}}} \tilde{f}_{\bar{\ell}}(\theta; u(x_N)) \end{pmatrix} \text{ is not full rank} \right\} \right) = 0$$

Therefore, for any fixed θ except for a measure-zero set, $\left\{ \begin{pmatrix} \nabla_{\theta_i} f(x_1) \\ \vdots \\ \nabla_{\theta_i} f(x_N) \end{pmatrix} \right\}_{i=1}^{\dim \theta}$ forms a frame except for a measure-zero set of $(x_1, \dots, x_N)$. $\square$

Lemma 6. Under Assumption 4 and some small adjustment of the scale of φ_{ℓ} for $\ell = 0, \dots, L-1$, consider fixed θ except for a measure-zero set. The lower frame bound $\mu_{\text{low}, \theta, X}$ is strictly positive except for a measure-zero set of $(x_1, \dots, x_N)$ in $\mathbb{R}^{Nd}$.

Moreover, fix x_i , and let $\mathcal{D}_{i,H} = \{j \mid h_{ij} = \|x_i - x_j\| \leq H, \forall j \neq i\}$. Then there exists $r_{\theta,i,H} > 0$ and $L_{\theta,i,H} > 0$, s.t., when $\sqrt{\sum_{j \in \mathcal{D}_{i,H}} h_{ij}^2} \leq r_{\theta,i,H}$, we have

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

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

$$\mu_{\text{low}, \theta, X} \leq L_{\theta,i} h_{\min}.$$

Proof. Let $A(\theta, X)$ be the frame operator, i.e.,

$$A(\theta, X) = \sum_{i=1}^{\dim \theta} \begin{pmatrix} \nabla_{\theta_i} f(x_1) \\ \vdots \\ \nabla_{\theta_i} f(x_N) \end{pmatrix} \begin{pmatrix} \nabla_{\theta_i} f(x_1) \\ \vdots \\ \nabla_{\theta_i} f(x_N) \end{pmatrix}^{\top} = \begin{pmatrix} \nabla_{\theta} f(x_1) \\ \vdots \\ \nabla_{\theta} f(x_N) \end{pmatrix} \begin{pmatrix} \nabla_{\theta} f(x_1) \\ \vdots \\ \nabla_{\theta} f(x_N) \end{pmatrix}^{\top}$$

We would like to bound $\lambda_{\min}(A)$. Consider any fixed θ except for a measure-zero set.

Let

$$\begin{aligned} B(\theta, X) &= \sum_{i=1}^{\dim \theta_{\bar{\ell}}} \begin{pmatrix} \nabla_{\theta_{\bar{\ell},i}} f(x_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell},i}} f(x_N) \end{pmatrix} \begin{pmatrix} \nabla_{\theta_{\bar{\ell},i}} f(x_1) \\ \vdots \\ \nabla_{\theta_{\bar{\ell},i}} f(x_N) \end{pmatrix}^{\top} \\ &= \sum_{i=1}^{\dim \theta_{\bar{\ell}}} \begin{pmatrix} \nabla_{\theta_{\bar{\ell},i}} \tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell}}(x_1)) \\ \vdots \\ \nabla_{\theta_{\bar{\ell},i}} \tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell}}(x_N)) \end{pmatrix} \begin{pmatrix} \nabla_{\theta_{\bar{\ell},i}} \tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell}}(x_1)) \\ \vdots \\ \nabla_{\theta_{\bar{\ell},i}} \tilde{f}_{\bar{\ell}}(\theta; u_{\bar{\ell}}(x_N)) \end{pmatrix}^{\top} \end{aligned}$$

We also denote

$$\tilde{B}(\theta, U) = \tilde{B}(\theta, (u_{\bar{\ell},1}, \dots, u_{\bar{\ell},N})) \triangleq B(\theta, X)$$

Then obviously, we have

$$\lambda_{\min}(A) \geq \lambda_{\min}(B) = \lambda_{\min}(\tilde{B}).$$

By Lemma 5, $\lambda_{\min}(B) > 0$ except for a measure-zero set of $(x_1, \dots, x_N)$ in $\mathbb{R}^{Nd}$.

In the end, we would like to show the dependence on $h_{\min}$. First, consider $h_{\min} = 0$, i.e., there exists i, j s.t. $x_i = x_j$. Then obviously $A(\theta, X)$ is degenerate, and $\sigma_{\min}(A(\theta, X)) = 0$.

Next, let $\mathcal{D}_{i,H} = \{j \mid h_{ij} = \|x_i - x_j\| \leq H, \forall j \neq i\}$. WLOG, assume that $\mathcal{D}_{N-k,H} = \{N-k+1, \dots, N\}$ for some k , and consider $X_1 = \begin{pmatrix} x_1 \\ \vdots \\ x_{N-k} \end{pmatrix}$, and $X_2 = \begin{pmatrix} x_{N-k+1} \\ \vdots \\ x_N \end{pmatrix}$.

Since $A(\theta, X)$ is symmetric, its eigenvalues are real for any X and θ . By Theorem 4.1 in Kurdyka & Paunescu (2008), we have that the map from X to all the eigenvalues listed as a column is locally Lipschitz. Therefore, there exists $r_{\theta,X,k} > 0$ and $L_{\theta,X,k} > 0$, s.t., for any $X_2 \in B_{X_2}(r_{\theta,X,k})$, where

$$X_2' = \begin{pmatrix} x_{N-k} \\ \vdots \\ x_{N-k} \end{pmatrix}, \text{ i.e., } \sqrt{\sum_{j \in \mathcal{D}_{N-k,H}} h_{N-k,j}^2} \leq r_{\theta,X,k}, \text{ we have}$$

$$\lambda_{\min}(\theta, X_1, X_2) \leq L_{\theta,X,k} \|X_2' - X_2\| = L_{\theta,X,k} \sqrt{\sum_{j \in \mathcal{D}_{N-k,H}} h_{N-k,j}^2} = L_{\theta,X,k} \sqrt{\sum_{j=0}^{k-1} h_{N-k,N-j}^2}.$$

Specifically, if $H = h_{\min}$ and $\mathcal{D}_{N-k,H} \neq \emptyset$, we have

$$\lambda_{\min}(\theta, X_1, X_2) \leq L_{\theta,X} h_{\min},$$

for $h_{\min} \leq r_{\theta,X}$ and $L_{\theta,X}, r_{\theta,X} > 0$. $\square$

E.2 POLYNOMIAL GENERALIZED SMOOTHNESS OF THE LOSS

In this section, we denote $\theta_{\max,i} = \arg \max\{\|\theta\|, \|\theta'\|\}$, and $u_{\max} = \arg \max\{\|u\|, \|u'\|\}$. All the constants C 's are ≥ 0 and independent of $\theta, \bar{\varphi}, u$; all the p 's in the subscripts are in $\mathbb{N}$. We denote $u_{\ell,i} = u_{\ell,i}(\theta)$ to emphasize the dependency on θ , especially when comparing $u_{\ell,i}$ under different θ 's. Also, note that $S(\cdot)$ has positive coefficients. Therefore $S(\cdot)$ is monotonically increasing on $\mathbb{R}_{\geq 0} \times \dots \times \mathbb{R}_{\geq 0}$, and

$$S(a_1, \dots, a_i, \dots, a_d) \leq S(|a_1|, \dots, |a_i|, \dots, |a_d|).$$

Below is the general lemma about the poly-smoothness of the neural network.

Lemma 7. *Under Assumption 3, $l(\theta; x_i)$ satisfies generalized smoothness*

$$\|\nabla_{\theta} l(\theta; x_i) - \nabla_{\theta} l(\theta'; x_i)\| \leq S_l(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_{\theta}}\|) \|\theta - \theta'\|, \quad (5)$$

and consequently,

$$\|\nabla_{\theta} \mathcal{L}(\theta) - \nabla_{\theta} \mathcal{L}(\theta')\| \leq S(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_{\theta}}\|) \|\theta - \theta'\| \quad (6)$$

where $S_l(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_{\theta}}\|), S(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_{\theta}}\|)$ are some polynomials with positive coefficient.

Proof. The proof follows directly from Proposition 2. $\square$

Next, we aim at deriving the polynomial function precisely under the neural networks where there is normalization τ_{ϵ} in each layer

$$\begin{aligned} u_{0,i} &= x_i; \\ u_{\ell+1,i} &= u_{\ell,i} + \bar{\varphi}_{\ell}(\theta_{\ell}; u_{\ell,i}), \forall \ell = 0, \dots, L-1; \\ f(\theta; x_i) &= \bar{\varphi}_L(\theta_L; u_{L,i}) \end{aligned} \quad (7)$$

where $\theta = (\theta_0, \dots, \theta_L)$ is the collection of weights with the θ_i corresponding to the i th layer, and $\bar{\varphi}_{\ell}(\theta_{\ell}; u) = \varphi_{\ell}(\theta_{\ell}; \tau_{\epsilon}(u))$ with $\varphi_{\ell}(\theta_{\ell}; \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and the normalization function τ_{ϵ} (see Section 3) for $\ell = 0, \dots, L$. Before the precise derivation, we provide a more detailed version of Assumption 3 as follows:

Assumption 5. Assume for all $\ell = 0, \dots, L$,

$$\|\varphi_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{0,\ell}} C_{0,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,\ell,t,i}} \|u\|^{p_{0,\ell,t}} \quad (8)$$

$$\|\nabla_{\theta_\ell} \varphi_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{1,\ell}} C_{1,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{1,\ell,t,i}} \|u\|^{p_{1,\ell,t}} \quad (9)$$

$$\|\nabla_u \varphi_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{2,\ell}} C_{2,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{2,\ell,t,i}} \|u\|^{p_{2,\ell,t}} \quad (10)$$

$$\begin{aligned} \|\bar{\varphi}_\ell(\theta; u) - \bar{\varphi}_\ell(\theta'; u')\| &\leq (C_{1,1,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{1,1,\ell,i}} \|u_{\max}\|^{p_{1,1,\ell}} + C_{1,1,\ell,0}) \|\theta - \theta'\| \\ &\quad + (C_{1,2,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{1,2,\ell,i}} \|u_{\max}\|^{p_{1,2,\ell}} + C_{1,2,\ell,0}) \|u - u'\| \end{aligned} \quad (11)$$

$$\begin{aligned} \|\nabla_{\theta_\ell} \bar{\varphi}_\ell(\theta; u) - \nabla_{\theta_\ell} \bar{\varphi}_\ell(\theta'; u')\| &\leq (C_{2,1,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{2,1,\ell,i}} \|u_{\max}\|^{p_{2,1,\ell}} + C_{2,1,\ell,0}) \|\theta - \theta'\| \\ &\quad + (C_{2,2,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{2,2,\ell,i}} \|u_{\max}\|^{p_{2,2,\ell}} + C_{2,2,\ell,0}) \|u - u'\| \end{aligned} \quad (12)$$

$$\begin{aligned} \|\nabla_u \bar{\varphi}_\ell(\theta; u) - \nabla_u \bar{\varphi}_\ell(\theta'; u')\| &\leq (C_{3,1,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{3,1,\ell,i}} \|u_{\max}\|^{p_{3,1,\ell}} + C_{3,1,\ell,0}) \|\theta - \theta'\| \\ &\quad + (C_{3,2,\ell} \prod_{i=1}^{n_\theta} \|\theta_{\max,i}\|^{p_{3,2,\ell,i}} \|u_{\max}\|^{p_{3,2,\ell}} + C_{3,2,\ell,0}) \|u - u'\| \end{aligned} \quad (13)$$

Lemma 8. Under the Assumption 5, we have that, for all $\ell = 0, \dots, L$,

$$\|\bar{\varphi}_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{0,\ell}} C_{0,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,\ell,t,i}} \quad (14)$$

$$\|\nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{1,\ell}} C_{1,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{1,\ell,t,i}} \quad (15)$$

$$\|\nabla_u \bar{\varphi}_\ell(\theta; u)\| \leq \sum_{t=1}^{n_{2,\ell}} C_{2,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{2,\ell,t,i}} \quad (16)$$

$$\|u_{\ell,i}\| \leq \|x_i\| + \sum_{j=0}^{\ell-1} \sum_{t=1}^{n_{0,j}} C_{0,j,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,j,t,i}} \quad (17)$$

$$\|u_{\ell,i}(\theta) - u_{\ell,i}(\theta')\| \leq \sum_{j=0}^{g(\ell)} C_{u,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{u,j,q}} \|\theta - \theta'\| \quad (18)$$

Proof. By the definition of ϵ -normalization, we have

$$\|\tau_\epsilon(u)\| = \left\| \frac{u}{\sqrt{\|u\|^2 + \epsilon^2}} \right\| \leq 1.$$

Then we have

$$\|\bar{\varphi}_\ell(\theta; u)\| = \|\varphi(\theta; \tau_\epsilon(u))\| \leq \sum_{t=1}^{n_{0,\ell}} C_{0,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,\ell,t,i}} \|\tau_\epsilon(u)\|^{p_{0,\ell,t}} \leq \sum_{t=1}^{n_{0,\ell}} C_{0,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,\ell,t,i}}.$$

The inequality of $\nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u)$ follows the same idea. For $\nabla_u \bar{\varphi}_\ell(\theta; u)$, first consider

$$\nabla \tau_\epsilon(u) = \frac{1}{\sqrt{\|u\|^2 + \epsilon^2}} I - \frac{1}{(\sqrt{\|u\|^2 + \epsilon^2})^3} uu^\top$$

Then by Wely's theorem,

$$\|\nabla \tau_\epsilon(u)\| \leq \frac{1}{\sqrt{\|u\|^2 + \epsilon^2}} + \frac{\|u\|^2}{(\sqrt{\|u\|^2 + \epsilon^2})^3} \leq \frac{1}{\epsilon}$$

where the equality of the last inequality is achieved at $\|u\| = 0$. Then

$$\begin{aligned} \|\nabla_u \bar{\varphi}_\ell(\theta; u)\| &= \|\nabla_{\tau_\epsilon(u)} \varphi_\ell(\theta; \tau_\epsilon(u)) \nabla \tau_\epsilon(u)\| \\ &\leq \|\nabla_{\tau_\epsilon(u)} \varphi_\ell(\theta; \tau_\epsilon(u))\| \|\nabla \tau_\epsilon(u)\| \\ &\leq \sum_{t=1}^{n_{2,\ell}} C_{2,\ell,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{2,\ell,t,i}} \end{aligned}$$

where $C_{2,\ell,t}$ depends on ϵ and the last inequality follows the same idea as the previous two inequalities.

For the upper bound of $u_{\ell,i}$, note that

$$u_{\ell+1,i} = u_{\ell,i} + \bar{\varphi}_\ell(\theta; u_{\ell,i}) = \dots = u_{0,i} + \sum_{j=0}^{\ell} \bar{\varphi}_j(\theta; u_{j,i}).$$

Then,

$$\|u_{\ell,i}\| \leq \|x_i\| + \sum_{j=0}^{\ell-1} \|\bar{\varphi}_j(\theta; u_{j,i})\| \leq \|x_i\| + \sum_{j=0}^{\ell-1} \sum_{t=1}^{n_{0,j}} C_{0,j,t} \prod_{i=1}^{n_\theta} \|\theta_i\|^{p_{0,j,t,i}}$$

where the second inequality follows from (14).

Also,

$$\begin{aligned} \|u_{\ell,i}(\theta) - u_{\ell,i}(\theta')\| &= \|x_0 + \sum_{j=0}^{\ell-1} \bar{\varphi}_j(\theta; u_{j,i}(\theta)) - x_0 - \sum_{j=0}^{\ell-1} \bar{\varphi}_j(\theta'; u_{j,i}(\theta'))\| \\ &\leq \sum_{j=0}^{\ell-1} \|\bar{\varphi}_j(\theta; u_{j,i}(\theta)) - \bar{\varphi}_j(\theta'; u_{j,i}(\theta'))\| \\ &\leq \sum_{j=0}^{\ell-1} (C_{1,1,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,j,q}} \|u_{\max}\|^{p_{1,1,j}} + C_{1,1,j,0}) \|\theta - \theta'\| \\ &\quad + \sum_{j=0}^{\ell-1} (C_{1,2,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,j,q}} \|u_{\max}\|^{p_{1,2,j}} + C_{1,2,j,0}) \|u_{j,i}(\theta) - u_{j,i}(\theta')\| \end{aligned}$$

where the first inequality follows from (14).

Then we denote the above inequality as follows

$$\underbrace{\|u_{\ell,i}(\theta) - u_{\ell,i}(\theta')\|}_{a_\ell} \leq \underbrace{\sum_{j=0}^{\ell-1} (C_{1,1,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,j,q}} \|u_{\max}\|^{p_{1,1,j}} + C_{1,1,j,0}) \|\theta - \theta'\|}_{b_\ell} + \underbrace{\sum_{j=0}^{\ell-1} (C_{1,2,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,j,q}} \|u_{\max}\|^{p_{1,2,j}} + C_{1,2,j,0}) \|u_{j,i}(\theta) - u_{j,i}(\theta')\|}_{\lambda_j} \underbrace{\quad}_{a_j}.$$

Since $a_0 = \|x_i - x_i\| = 0$, define $b_0 = 0 \geq a_0$, and then by discrete Gronwall's inequality (Proposition 4.1 in Emmrich (1999) with $\theta = 0, \tau_j = 1$), we have

$$a_\ell \leq b_\ell + \sum_{j=0}^{\ell-1} \lambda_j b_j \prod_{s=j+1}^{\ell-1} (1 + \lambda_s),$$

namely,

$$\begin{aligned} & \|u_{\ell,i}(\theta) - u_{\ell,i}(\theta')\| \\ & \leq \left(\sum_{j=0}^{\ell-1} (C_{1,1,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,j,q}} \|u_{\max,j,i}\|^{p_{1,1,j}} + C_{1,1,j,0}) \right. \\ & \quad + \sum_{j=1}^{\ell-1} (C_{1,2,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,j,q}} \|u_{\max,j,i}\|^{p_{1,2,j}} + C_{1,2,j,0}) \\ & \quad \times \left(\sum_{r=0}^{j-1} C_{1,1,r} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,r,q}} \|u_{\max,r,i}\|^{p_{1,1,r}} + C_{1,1,r,0} \right) \\ & \quad \times \prod_{s=j+1}^{\ell-1} (1 + C_{1,2,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,s,q}} \|u_{\max,s,i}\|^{p_{1,2,s}} + C_{1,2,s,0}) \Big) \|\theta - \theta'\| \\ & \leq \left(\sum_{j=0}^{\ell-1} (C_{1,1,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,j,q}} \left(\sum_{s=0}^{j-1} C_{0,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,s,q}} + \|x_i\| \right)^{p_{1,1,s}} + C_{1,1,s,0}) \right. \\ & \quad + \sum_{j=1}^{\ell-1} (C_{1,2,j} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,j,q}} \left(\sum_{s=0}^{j-1} C_{0,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,s,q}} + \|x_i\| \right)^{p_{1,2,j}} + C_{1,2,j,0}) \\ & \quad \times \left(\sum_{r=0}^{j-1} C_{1,1,r} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,r,q}} \left(\sum_{s=0}^{r-1} C_{0,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,s,q}} + \|x_i\| \right)^{p_{1,1,r}} + C_{1,1,r,0} \right) \\ & \quad \times \prod_{s=j+1}^{\ell-1} (1 + C_{1,2,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,s,q}} \left(\sum_{t=0}^{s-1} C_{0,t} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,t,q}} + \|x_i\| \right)^{p_{1,2,s}} + C_{1,2,s,0}) \Big) \|\theta - \theta'\| \\ & = \sum_{j=0}^{g_3(\ell)} C_{u,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{u,j,q}} \|\theta - \theta'\| \end{aligned}$$

where the second inequality follows from (17), $g_3(\ell)$ is some polynomial of ℓ , and $C_{u,j,i}$ is some constant depending on $\|x_i\|$. $\square$

Corollary 5. Under the same assumptions as Lemma 8, we have that, for all $\ell = 0, \dots, L$,

$$\|\bar{\varphi}_\ell(\theta; u_{\ell,i}(\theta)) - \bar{\varphi}_\ell(\theta'; u_{\ell,i}(\theta'))\| \leq \sum_{j=0}^{\tilde{g}_3(\ell)} C_{\bar{\varphi},\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{\bar{\varphi},\ell,j,q}} \|\theta - \theta'\| \quad (19)$$

$$\|\nabla_{\theta_\ell} \bar{\varphi}_\ell(\theta; u_{\ell,i}(\theta)) - \nabla_{\theta_\ell} \bar{\varphi}_\ell(\theta'; u_{\ell,i}(\theta'))\| \leq \sum_{j=0}^{\tilde{g}_3(\ell)} C_{\nabla_{\theta_\ell} \bar{\varphi},\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{\nabla_{\theta_\ell} \bar{\varphi},\ell,j,q}} \|\theta - \theta'\| \quad (20)$$

$$\|\nabla_u \bar{\varphi}_\ell(\theta; u_{\ell,i}(\theta)) - \nabla_u \bar{\varphi}_\ell(\theta'; u_{\ell,i}(\theta'))\| \leq \sum_{j=0}^{\tilde{g}_3(\ell)} C_{\nabla_u \bar{\varphi},\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{\nabla_u \bar{\varphi},\ell,j,q}} \|\theta - \theta'\| \quad (21)$$

where $\tilde{g}_3(\ell)$ is some polynomial of ℓ , and the C 's depends on $\|x_i\|$.

Proof. For the first inequality,

$$\begin{aligned}
& \|\bar{\varphi}_\ell(\theta; u_{\ell,i}(\theta)) - \bar{\varphi}_\ell(\theta'; u_{\ell,i}(\theta'))\| \\
& \leq \left(C_{1,1,\ell} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,\ell,q}} \|u_{\max}\|^{p_{1,1,\ell}} + C_{1,1,\ell,0} \right) \|\theta - \theta'\| \\
& \quad + \left(C_{1,2,\ell} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,\ell,q}} \|u_{\max}\|^{p_{1,2,\ell}} + C_{1,2,\ell,0} \right) \|u_{\ell,i}(\theta) - u_{\ell,i}(\theta')\| \\
& \leq \left(C_{1,1,\ell} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,1,\ell,q}} \left(\sum_{s=0}^{j-1} C_{0,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,s,q}} + \|x_i\| \right)^{p_{1,1,\ell}} + C_{1,1,\ell,0} \right) \|\theta - \theta'\| \\
& \quad + \left(C_{1,2,\ell} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{1,2,\ell,q}} \left(\sum_{s=0}^{j-1} C_{0,s} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{0,s,q}} + \|x_i\| \right)^{p_{1,2,\ell}} + C_{1,2,\ell,0} \right) \\
& \quad \times \sum_{j=0}^{g_3(\ell)} C_{u,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{u,j,q}} \|\theta - \theta'\| \\
& = \sum_{j=0}^{\tilde{g}_3(\ell)} C_{\bar{\varphi},\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{\bar{\varphi},\ell,j,q}} \|\theta - \theta'\|
\end{aligned}$$

where the second inequality follows from (17) and (18), $\tilde{g}_3(\ell)$ is some polynomial of ℓ , $C_{\bar{\varphi},\ell,j,i}$ depends on $\|x_i\|$. The other two inequalities follow the same idea. $\square$

Lemma 9. Under Assumption 5, $l(\theta; x_i)$ satisfies generalized smoothness

$$\|\nabla_\theta l(\theta; x_i) - \nabla_\theta l(\theta'; x_i)\| \leq \sum_{j=1}^{n_\theta} \sum_{\ell \in \mathcal{J}_j} \sum_{j=0}^{g_4(\ell)} C_{l,\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{l,\ell,j,q}} \|\theta - \theta'\| \quad (22)$$

where $g_4(\ell)$ is some polynomial of ℓ . Consequently,

$$\|\nabla_\theta \mathcal{L}(\theta) - \nabla_\theta \mathcal{L}(\theta')\| \leq S(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_\theta}\|) \|\theta - \theta'\| \quad (23)$$

where $S(\|\theta_{\max,1}\|, \dots, \|\theta_{\max,n_\theta}\|) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{n_\theta} \sum_{\ell \in \mathcal{J}_j} \sum_{j=0}^{g_4(\ell)} C_{l,\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{l,\ell,j,q}}$.

Proof. From the calculations at the beginning of this section, we know

$$\begin{aligned}
\nabla_{\theta_j} l(\theta; x_i) &= \nabla_f l(\theta; x_i) \nabla_{\theta_\ell} f(x_i) \\
&= \sum_{\ell \in \mathcal{J}_j} \nabla_f l(\theta; x_i) \nabla_u \bar{\varphi}_L(\theta; u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(\theta; u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(\theta; u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u_{\ell,i})
\end{aligned}$$

Then

$$\begin{aligned}
& \|\nabla_{\theta_j} l(\theta; x_i) - \nabla_{\theta_j} l(\theta'; x_i)\| \\
& \leq \sum_{\ell \in \mathcal{J}_j} \left\| \nabla_f l(\theta; x_i) \nabla_u \bar{\varphi}_L(\theta; u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(\theta; u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(\theta; u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u_{\ell,i}) \right. \\
& \quad \left. - \nabla_f l(\theta'; x_i) \nabla_u \bar{\varphi}_L(\theta'; u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(\theta'; u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(\theta'; u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_\ell(\theta'; u_{\ell,i}) \right\| \\
& = \sum_{\ell \in \mathcal{J}_j} \left\| (\bar{\varphi}_L(\theta; u_{L,i}) - y_i)^\top \nabla_u \bar{\varphi}_L(\theta; u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(\theta; u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(\theta; u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u_{\ell,i}) \right. \\
& \quad \left. - (\bar{\varphi}_L(\theta'; u_{L,i}) - y_i)^\top \nabla_u \bar{\varphi}_L(\theta'; u_{L,i}) (I + \nabla_u \bar{\varphi}_{L-1}(\theta'; u_{L-1,i})) \cdots (I + \nabla_u \bar{\varphi}_{\ell+1}(\theta'; u_{\ell+1,i})) \nabla_{\theta_j} \bar{\varphi}_\ell(\theta'; u_{\ell,i}) \right\| \\
& \leq \sum_{\ell \in \mathcal{J}_j} (\|\bar{\varphi}_L(\theta; u_{L,i})\| + \|y_i\|) \|\nabla_u \bar{\varphi}_L(\theta; u_{L,i})\| (1 + \|\nabla_u \bar{\varphi}_{L-1}(\theta; u_{L-1,i})\|) \cdots \\
& \quad \times (1 + \|\nabla_u \bar{\varphi}_{\ell+1}(\theta; u_{\ell+1,i})\|) \|\nabla_{\theta_j} \bar{\varphi}_\ell(\theta; u_{\ell,i}) - \nabla_{\theta_j} \bar{\varphi}_\ell(\theta'; u_{\ell,i})\| \\
& \quad + \cdots + (\|\bar{\varphi}_L(\theta; u_{L,i})\| + \|y_i\|) \|\nabla_u \bar{\varphi}_L(\theta; u_{L,i})\| (1 + \|\nabla_u \bar{\varphi}_{L-1}(\theta; u_{L-1,i})\|) \cdots \\
& \quad \times \|\nabla_u \bar{\varphi}_s(\theta; u_{s,i}) - \nabla_u \bar{\varphi}_s(\theta'; u_{s,i})\| \cdots (1 + \|\nabla_u \bar{\varphi}_{\ell+1}(\theta'; u_{\ell+1,i})\|) \|\nabla_{\theta_j} \bar{\varphi}_\ell(\theta'; u_{\ell,i})\| \\
& \quad + \cdots + \|\bar{\varphi}_L(\theta; u_{L,i}) - \bar{\varphi}_L(\theta'; u_{L,i})\| \|\nabla_u \bar{\varphi}_L(\theta'; u_{L,i})\| (1 + \|\nabla_u \bar{\varphi}_{L-1}(\theta'; u_{L-1,i})\|) \cdots \\
& \quad \times (1 + \|\nabla_u \bar{\varphi}_{\ell+1}(\theta'; u_{\ell+1,i})\|) \|\nabla_{\theta_j} \bar{\varphi}_\ell(\theta'; u_{\ell,i})\| \\
& \leq \sum_{\ell \in \mathcal{J}_j} \sum_{j=0}^{g_4(\ell)} C_{l,\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{l,\ell,j,q}} \|\theta - \theta'\|
\end{aligned}$$

where the second inequality follows from triangular inequality, and the last inequality follows from Lemma 8 and Corollary 5; $C_{l,\ell,j,i}$ is a constant that depends on $\|x_i\|$ and $\|y_i\|$, and $g_4(\ell)$ is a polynomial of ℓ .

Then

$$\begin{aligned}
\|\nabla_\theta l(\theta; x_i) - \nabla_\theta l(\theta'; x_i)\| & \leq \sum_{j=1}^{n_\theta} \|\nabla_{\theta_j} l(\theta; x_i) - \nabla_{\theta_j} l(\theta'; x_i)\| \\
& \leq \sum_{j=1}^{n_\theta} \sum_{\ell \in \mathcal{J}_j} \sum_{j=0}^{g_4(\ell)} C_{l,\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{l,\ell,j,q}} \|\theta - \theta'\|
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla_\theta \mathcal{L}(\theta) - \nabla_\theta \mathcal{L}(\theta')\| & \leq \frac{1}{N} \sum_{i=1}^N \|\nabla_\theta l(\theta; x_i) - \nabla_\theta l(\theta'; x_i)\| \\
& \leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{n_\theta} \sum_{\ell \in \mathcal{J}_j} \sum_{j=0}^{g_4(\ell)} C_{l,\ell,j,i} \prod_{q=1}^{n_\theta} \|\theta_{\max,q}\|^{p_{l,\ell,j,q}} \|\theta - \theta'\|
\end{aligned}$$

□

E.3 ADDITIONAL LEMMAS IN MEASURE THEORY AND DIFFERENTIAL TOPOLOGY

In this section, we first introduce some basic concepts and theorems in measure theory and differential topology, and then state the supplementary lemmas of the proof.

E.3.1 BASIC DEFINITIONS AND THEOREMS

We first introduce the parametric transversality theorem (Theorem 6), together with the necessary concepts including residual set and transversality. We then state other useful theorems that will be applied in the lemmas of the following section

We first introduce the notions of nowhere dense set and meager set in order to define the residual set.

Definition 7 (Nowhere dense set). *A set is called nowhere dense if it is not dense in any nonempty open subset, or equivalently, the interior of its closure is empty.*

Definition 8 (Meager set). *A set is called meager if it is a countable union of nowhere dense sets.*

Definition 9 (Residual set). *A set is called residual if its complement is a countable union of nowhere dense sets, or equivalently, its complement is meager.*

Proposition 3. *The meager set and residual set have the following properties:*

1. *All countable unions of meager sets are meager.*
2. *Consider a continuous function $f : (a, \infty) \rightarrow \mathbb{R}$ for some $a > 0$. Then if $X \subseteq (a, \infty)$ is residual in (a, ∞) , then $f(X)$ is residual in $(0, 1/a)$.*

We then introduce transversality of a map.

Definition 10. (Transversality) *A map $f : M \rightarrow N$ is transversal to a submanifold $A \subseteq N$ if for any $x \in f^{-1}(A)$, we have $\text{Im}(Df_x) + T_{f(x)}A = T_{f(x)}N$, i.e., the tangent space of N at $f(x)$ is spanned by the image of the differential of f at x and the tangent space of A at $f(x)$.*

In this paper, we only consider the case where the submanifold $A = \{0\}$. Then, transversality is equivalent to non-degeneracy of the map.

Theorem 6 (Parametric transversality theorem (Hirsch, 2012)). *Let V, M, N be C^r manifold without boundary and $A \subset N$ a C^r submanifold. Let $F : V \rightarrow C^r(M, N)$ satisfy the following conditions:*

- $F^{\text{ev}} : V \times M \rightarrow N$, $(v, x) \mapsto F_v(x)$, is C^r ;
- F^{ev} is transverse to A ;
- $r > \max\{0, \dim N + \dim A - \dim M\}$.

Then the set $\{v \in V : F_v \text{ is transverse to } A\}$ is residual and therefore dense.

In addition to the parametric transversality theorem, we also employ the following useful results.

Theorem 7 (Mityagin (2015)). *Let $f(x)$ be a real analytic function on (a connected open domain U of) $\mathbb{R}^d$. If f is not identically zero, then its zero set*

$$f^{-1}(0) = \{x \in U : f(x) = 0\}$$

has a zero measure, i.e., $\text{Leb}(f^{-1}(0)) = 0$.

Theorem 8 (Theorem G.3 in Wang et al. (2022b)). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and $f \in C^1$. If the set of critical points of f is a null-set, i.e.,*

$$\mathcal{L}(\{x \in \mathbb{R}^d : \nabla f(x) \text{ is not invertible}\}) = 0,$$

then $\mathcal{L}(f^{-1}(B)) = 0$ for any null-set B .

Theorem 9 (Lemma 3 in Rader (1973)). *Consider a function $f : U \subseteq \mathbb{R}^d \rightarrow \mathbb{R}^d$ that is pointwise Lipschitz, i.e., for any $x \in U$, there exists a neighbourhood U_x of x and a constant L_x , s.t.*

$$\|f(x) - f(y)\| \leq L_x \|x - y\|, \quad \forall y \in U_x.$$

Then its image of a measure-zero set is still measure zero.

Theorem 10 (Preimage Theorem). *If y is a regular value of a smooth map $f : X \rightarrow Y$, then the preimage $f^{-1}(y)$ is a submanifold of X , with $\dim f^{-1}(y) = \dim X - \dim Y$.*

E.3.2 SUPPLEMENTARY LEMMAS OF THE PROOF

We then show the supplementary lemmas for the convergence of neural network (1) by applying the above theorems.

Lemma 10. *Fix some θ . Under Assumption 2,*

1. *Let $u_\ell = u_\ell(x) : \mathbb{R}^d \rightarrow \mathbb{R}^d$. The preimage of u_ℓ for any measure-zero set in $\mathbb{R}^d$ is a measure-zero set in $\mathbb{R}^d$, for all $\ell = 0, \dots, L-1$.*

2. Let $U_\ell(x_1, \dots, x_n) = \begin{pmatrix} u_\ell(x_1) \\ \vdots \\ u_\ell(x_N) \end{pmatrix} : \mathbb{R}^{Nd} \rightarrow \mathbb{R}^{Nd}$. The preimage of U_ℓ for any measure-zero set in $\mathbb{R}^{Nd}$ is a measure-zero set in $\mathbb{R}^{Nd}$, for all $\ell = 0, \dots, L-1$.

Proof. Suppose the critical point set of u_ℓ is measure zero, and thus by Theorem 8, the preimage of u_ℓ for any measure zero set is a null set. Consider

$$u_{\ell+1} = u_{\ell+1}(x),$$

where its Jacobian is

$$\nabla u_{\ell+1}(x) = (I + \nabla_u \bar{\varphi}_\ell(u_\ell)) \cdots (I + \nabla_u \bar{\varphi}_0(u_0)).$$

We would like to prove that the critical point set of $u_{\ell+1}$ is measure zero, i.e.,

$$\{x \in \mathbb{R}^d : \nabla u_{\ell+1}(x) \text{ is not full rank}\}.$$

Note that

$$\nabla u_{\ell+1}(x) = (I + \nabla_u \bar{\varphi}_\ell(u_\ell)) \nabla u_\ell(x),$$

and thus

$$\begin{aligned} & \{x \in \mathbb{R}^d : \nabla u_{\ell+1}(x) \text{ is not full rank}\} \\ &= \{x \in \mathbb{R}^d : I + \nabla_u \bar{\varphi}_\ell(u_\ell(x)) \text{ is not full rank}\} \cup \{x \in \mathbb{R}^d : \nabla u_\ell(x) \text{ is not full rank}\}. \end{aligned}$$

By our assumption, the set

$$\text{Leb}_d(\{x \in \mathbb{R}^d : \nabla u_\ell(x) \text{ is not full rank}\}) = 0.$$

Also, by Lemma 11,

$$\text{Leb}_d(\{u \in \mathbb{R}^d : I + \nabla_u \bar{\varphi}_\ell(u) \text{ is not full rank}\}) = 0,$$

and then by Theorem 8, the preimage of the above measure zero set is still measure zero, i.e.,

$$\text{Leb}_d(\{x \in \mathbb{R}^d : I + \nabla_u \bar{\varphi}_\ell(u_\ell(x)) \text{ is not full rank}\}) = 0.$$

Thus the critical point set of $u_{\ell+1}$ is measure zero, i.e.

$$\text{Leb}_d(\{x \in \mathbb{R}^d : \nabla u_{\ell+1}(x) \text{ is not full rank}\}) \leq 0 + 0 = 0.$$

By Theorem 8, the first conclusion holds for $u_{\ell+1}$.

For the second statement, the Jacobian of U is

$$\begin{aligned} \nabla U_{\ell+1}(x_1, \dots, x_N) &= \begin{pmatrix} \nabla u_{\ell+1}(x_1) & & \\ & \ddots & \\ & & \nabla u_{\ell+1}(x_N) \end{pmatrix} \\ &= \begin{pmatrix} I + \nabla_u \bar{\varphi}_\ell(u_\ell(x_1)) & & \\ & \ddots & \\ & & I + \nabla_u \bar{\varphi}_\ell(u_\ell(x_N)) \end{pmatrix} \nabla U_\ell(x_1, \dots, x_N) \\ &= \begin{pmatrix} I_{Nd} + \begin{pmatrix} \nabla_u \bar{\varphi}_\ell(u_\ell(x_1)) & & \\ & \ddots & \\ & & \nabla_u \bar{\varphi}_\ell(u_\ell(x_N)) \end{pmatrix} \end{pmatrix} \nabla U_\ell(x_1, \dots, x_N). \end{aligned}$$

Then the second statement follows a similar derivation as the first one. $\square$

Lemma 11. Let $f(x, y) : \mathbb{R}^m \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}^m$ be an analytic function. Consider $g(x, y, \alpha) = \alpha x + f(x, y)$ where $\alpha \geq c > 0$ for some constant $c < 1$. Then the following statements hold for (1) fixed y and any x except for a measure-zero set, and (2) fixed x and any y except for a measure-zero set:

1. There exists a residual set in the neighborhood of 0, such that, when ϵ_0 is in this set, the Jacobian w.r.t. x of either $x + f(x, y)$ or $x + (1 + \epsilon_0)f(x, y)$ is full rank.

2. There exists a residual set of $(0, \frac{1}{c})$, such that, when η is in this residual set, the Jacobian of $x + \eta f(x, y)$ w.r.t. x is full rank.

Proof. First, consider

$$F(\alpha, x, y) = \det(\alpha I + \nabla_x f(x, y)) = \prod_{i=1}^m (\alpha + \lambda_i(x, y)),$$

where $\lambda_i(x, y)$ is the eigenvalue of $\nabla_x f(x, y)$. By Proposition 1, $F(\alpha, x, y)$ is a analytic function of α, x, y .

Fix x, y . Then $F(\alpha, x, y)$ is a polynomial of α and the critical points of $F(\alpha, x, y)$ are isolated. Thus $F'_\alpha(\alpha, x, y) \neq 0$ for all alpha except for a measure 0 set of isolated points.

Next we no longer fix x, y , let $A = \{0\}$, and consider any open interval I of α that does not contain the critical points, which is a manifold. Then from the above discussion, $\nabla F(\alpha, x, y) \neq \mathbf{0}$, namely F is surjective and thus transverse to A . Therefore, by Theorem 6, the set $\mathcal{S}_\alpha = \{\alpha \in I : F_\alpha \text{ is transverse to } A\}$ is residual and dense, which also implies that $F_\alpha(x, y) = \det(\alpha I + \nabla_x f(x, y))$ is transverse to A for a residual set of α in (c, ∞) . Then, for any fixed α in this residual set, $\nabla F_\alpha(x, y)$ is surjective for any $(x, y) \in F_\alpha^{-1}(0)$, and hence 0 is a regular point. Then by Theorem 10, $F_\alpha^{-1}(0)$ is a submanifold of codimension 1, and therefore

$$\text{Leb}_{mn_2}(F_\alpha^{-1}(0)) = 0.$$

Then we consider fixing x or y , i.e., $F_{\alpha, x}(y)$ and $F_{\alpha, y}(x)$. The above reasoning also holds true for the two functions, and therefore

$$\text{Leb}_m(F_{\alpha, y}^{-1}(0)) = 0, \text{ for some fixed } y, \text{ and } \text{Leb}_{n_2}(F_{\alpha, x}^{-1}(0)) = 0, \text{ for some fixed } x.$$

This concludes the second statement by choosing $\eta = \frac{1}{\alpha}$, which is also residual in $(0, \frac{1}{c})$ by Proposition 3.

For the first statement, if -1 is not an eigenvalue of $\nabla_x f(x, y)$, then $I + \nabla_x f(x, y)$ has full rank. Otherwise, since α is dense in any open set that does not contain 1, we can always find a small enough ϵ_1 , such that $1 - \epsilon_1$ is inside this set $\mathcal{S}_\alpha$. Then $I + \frac{1}{1-\epsilon_1} \nabla_x f(x, y)$ is has full rank. The first statement follows from choosing $\epsilon_0 = \frac{1}{1-\epsilon_1} - 1$.

□

The following lemmas are not directly used in the proof but reflect properties of this neural network, so we include them here.

Lemma 12. For any $(x_1, \dots, x_N)$ except for a measure-zero set in $\mathbb{R}^{Nd}$, the Jacobian along the line segment between any two points $u_{\ell, i}, u_{\ell, j}$ has full rank at every layer $\ell = 0, \dots, L-1$. Furthermore, for any point u_ℓ on the line segment between x_i and x_j , the Jacobian at any layer $\ell = 0, \dots, L-1$ also has full rank.

Proof. First, by Lemma 10, the critical point set of all u_ℓ , denoted by $\mathcal{B}$, is a finite union of measure-zero sets in $\mathbb{R}^d$, and thus is measure zero, i.e.,

$$\begin{aligned} \text{Leb}_d(\mathcal{B}) &= \text{Leb}_d\left(\bigcup_{\ell=0}^{L-1} \{x \in \mathbb{R}^d : \nabla_u u_\ell(x) \text{ is not full rank}\}\right) \\ &\leq \sum_{\ell=0}^{L-1} \text{Leb}_d(\{x \in \mathbb{R}^d : \nabla_u u_\ell(x) \text{ is not full rank}\}) = 0. \end{aligned}$$

Next, we would like to show that the following set is measure zero in $\mathbb{R}^{Nd}$

$$\mathcal{D} = \{(x_1, \dots, x_N) : \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]; i, j = 1, \dots, N\} \cap u_\ell(\mathcal{B}) \neq \emptyset \text{ for some } \ell = 0, \dots, L-1\}$$

Let

$$\bar{U}(\mathcal{B}) = \bigcup_{\ell=0}^{L-1} u_\ell(\mathcal{B}), \text{ and } \bar{\mathcal{B}} = \bigcup_{\ell=0}^{L-1} \{x \in u_\ell^{-1}(\bar{U}(\mathcal{B}))\}.$$

Since $u_\ell(x)$ is analytic, and thus differentiable, then $u_\ell(x)$ is pointwise Lipschitz. By Theorem 9, we have

$$\text{Leb}_d(\bar{U}(\mathcal{B})) \leq \sum_{\ell=0}^{L-1} \text{Leb}_d(u_\ell(\mathcal{B})) = 0.$$

Then by Theorem 8,

$$\text{Leb}_d(\bar{\mathcal{B}}) \leq \sum_{\ell=0}^{L-1} \text{Leb}_d(u_\ell^{-1}(\bar{U}(\mathcal{B}))) = 0.$$

We consider

$$\begin{aligned} \mathcal{D} &\subseteq \{(x_1, \dots, x_n) : \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]; i, j = 1, \dots, N; \ell = 0, \dots, L-1\} \cap \bar{U}(\mathcal{B}) \neq \emptyset\} \\ &\subseteq \bigcup_{\ell=0}^{L-1} \bigcup_{i,j} \bigcup_{z \in \bar{U}(\mathcal{B})} \{(x_1, \dots, x_N) : z \in \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]\}\} \end{aligned}$$

We denote $\mathcal{D}_{i,j,\ell,z} = \{(x_1, \dots, x_N) : z \in \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]\}\}$ and consider the measure of $\bigcup_{z \in \bar{U}} \mathcal{D}_{i,j,\ell,z}$, i.e.,

$$\text{Leb}_{Nd}(\bigcup_{z \in \bar{U}} \mathcal{D}_{i,j,\ell,z}) \leq \int \cdots \int \mathbb{1}_{z \in \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]\}} dx_1 \cdots dx_N dz$$

By Fubini-Tonelli's Theorem,

$$\text{Leb}_{Nd}(\bigcup_{z \in \bar{U}} \mathcal{D}_{i,j,\ell,z}) \leq \int \cdots \int \mathbb{1}_{z \in \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]\}} dx_i dz dx_j \cdots dx_N$$

For any fixed x_j and z , the set $\{u_\ell(x_i) : z \in \{tu_\ell(x_i) + (1-t)u_\ell(x_j) : t \in [0, 1]\}\}$ is a ray in $\mathbb{R}^d$ and thus is measure zero. Therefore, its preimage is measure-zero in $\mathbb{R}^d$. Also, from the above derivation, $\text{Leb}_d(\bar{U}(\mathcal{B})) = 0$. Then we have

$$\text{Leb}_{Nd}(\bigcup_{z \in \bar{U}} \mathcal{D}_{i,j,\ell,z}) \leq 0.$$

We then consider $\bigcup_{\ell=0}^{L-1} \bigcup_{i,j} \bigcup_{z \in \bar{U}(\mathcal{B})} \mathcal{D}_{i,j,\ell,z}$. Since there are $\binom{N}{2}$ pair of (i, j) and L such ℓ , which are both finite, we have

$$\text{Leb}_{Nd}(\mathcal{D}) \leq \text{Leb}_{Nd}(\bigcup_{\ell=0}^{L-1} \bigcup_{i,j} \bigcup_{z \in \bar{U}(\mathcal{B})} \mathcal{D}_{i,j,\ell,z}) \leq \sum_{\ell=0}^{L-1} \sum_{(i,j)} 0 = 0$$

Next, we would like to show that the following set is measure zero in $\mathbb{R}^{Nd}$

$$\mathcal{D} = \{(x_1, \dots, x_n) : \{u_\ell(tx_i + (1-t)x_j) : t \in [0, 1]; i, j = 1, \dots, N\} \cap u_\ell(\mathcal{B}) \neq \emptyset \text{ for some } \ell = 0, \dots, L-1\}.$$

The proof follows the same idea as the previous statement. The only difference is that for any fixed x_j and z , the set $\{x_i : z \in \{u_\ell(tx_i + (1-t)x_j) : t \in [0, 1]\}\}$ is a ray in $\mathbb{R}^d$ and thus is measure zero.

□

Lemma 13. Fix θ and consider the network under some small adjustment of the scale of φ_ℓ for $\ell = 0, \dots, L-1$. For any $(x_1, \dots, x_N)$ except for a measure-zero set in $\mathbb{R}^{Nd}$, each layer satisfies the following lower Lipschitz inequality

$$\|u_{\ell,i} - u_{\ell,j}\| \geq \mu_{u,\ell,\theta} \|x_j - x_i\|, \quad \forall \ell = 0, \dots, L-1, \quad \forall i, j = 1, \dots, N.$$

where $\mu_{u,\ell,\theta} > 0$ is some constant depending on θ, ℓ .

Proof. First, by Lemma 12, the Jacobians of $u_\ell(x)$ the line segments between each point $u_{\ell,i}, u_{\ell,j}$ are full rank for all $\ell = 0, \dots, L-1$, except for a measure-zero set of $X = (x_1, \dots, x_N)$ in $\mathbb{R}^{Nd}$, denoted by $\mathcal{B}_{Nd}$.

Then for fixed θ and $X = (x_1, \dots, x_N) \notin \mathcal{B}_{Nd}$, there exists some constant $\mu_{u,\ell,\theta} > 0$, s.t.

$$\mu_{u,\ell,\theta} = \min_{\{x: u_\ell(x) = t u_\ell(x_i) + (1-t) u_\ell(x_j), t \in [0,1]\}} \lambda_{\min}(\nabla_x u_\ell(x))$$

where $\lambda_{\min}$ is the minimum singular value.

Then

$$\begin{aligned} \|u_\ell(x_i) - u_\ell(x_j)\| &= \left\| \int_0^1 \nabla u_\ell(x_i + t(x_j - x_i))(x_j - x_i) dt \right\| \\ &= \left\| \int_0^1 \nabla u_\ell(x_i + t(x_j - x_i)) dt \cdot (x_j - x_i) \right\| \\ &\geq \mu_{u,\ell,\theta} \|x_j - x_i\|. \end{aligned}$$

□

F APPROXIMATION: PROOF OF THEOREM 3

Below is the formal version of Theorem 3, and we will prove it in this section.

Theorem 11. Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$ under Assumption 1. Let $\mathcal{I}_i = \text{conv}\{x_{i_1}, \dots, x_{i_{d+1}}\}$ with $x_j \notin \mathcal{I}_i, \forall j \neq i_1, \dots, i_{d+1}$, and define $h_{\max_{d+1}} = \max_i \max_{x_q, x_j \in \mathcal{I}_i} h_{qj}$. Suppose there is a family of neural networks that can approximate polynomial of degree $n_d = n_d(N, d)$, i.e., for any $\epsilon_2 > 0$ and $\psi \in P_{n_d}$, there exists some f s.t. $\|f - \psi\|_p \leq \epsilon_2$. Then, for some $C_1 > 0$ and $m > \max\{r, n_d\}$, there exists some $\phi \in W^{m,p}(\Omega)$ satisfying $\|g - \phi\|_{r,p} \leq C_2 h_{\max_{d+1}}^m h_{\min}^{-r} \|\phi\|_{m,p, \mathcal{I}^\circ}$, s.t., with probability 1 over the joint distribution of $x_1, \dots, x_N$, we have

$$\|f - g\|_{p, \mathcal{I}^\circ} \leq C_1 \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p}.$$

Proof of Theorem 11. Consider the ground truth function $g \in W^{r,p}(\Omega)$. Since $W^{m,p}(\Omega)$ is dense in $W^{r,p}(\Omega)$ for $m \geq r$, we have that for any $\epsilon_1 > 0$, there exists $\phi \in W^{m,p}(\Omega)$ such that

$$\|g - \phi\|_{r,p} \leq \epsilon_1.$$

Also, by Lemma 14 and Theorem 13, there exists a polynomial ψ of order at most $n_d(N, d)$, s.t.,

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

For a family of neural network functions that can approximate polynomial of order n_d , i.e., for any $\epsilon_2 > 0$ and $\psi \in P_{n_d}$, there exists some f in this class s.t.

$$\|f - \psi\|_p \leq \epsilon_2$$

Let $\epsilon_1 = \epsilon_2 = \frac{1}{2} C \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p}$. Then combining all the results above, we have

$$\begin{aligned} \|f - g\|_{p, \mathcal{I}^\circ} &\leq \|f - \psi\|_p + \|\psi - \phi\|_p + \|\phi - g\|_p \\ &\leq \|f - \psi\|_p + \|\psi - \phi\|_{r,p} + \|\phi - g\|_{r,p} \\ &\leq 2C \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p}. \end{aligned}$$

□

F.1 DEFINITIONS AND LEMMAS RELATED TO COMPUTATIONAL GEOMETRY

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

Definition 11 (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 12 (Delaunay triangulation). Given a set $\mathcal{P}$ of points in the n -dimensional Euclidean space, a Delaunay triangulation is a triangulation $DT(\mathcal{P})$ such that if for every n -simplex (whose vertices belong to $\mathcal{P}$) in the triangulation, there exists an open n -ball that passes through all the vertices of the simplex and that does not contain any other point of $\mathcal{P}$ in its interior.

Definition 13 (n -simplex). A n -simplex in n -dimensional space is the convex hull of $n + 1$ affinely independent points in $\mathbb{R}^n$.

Proposition 4. Let $\{v_0, v_1, \dots, v_d\} \subset \mathbb{R}^n$ be the vertices of an n -simplex. Let $u_i = v_i - v_0$, for $i = 1, \dots, n$.

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

Definition 14 (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 12 (Theorem 3.1, and Chapter 9, in De Berg (2000)). Assume the set $\{x_i\}_{i=1}^N$ is in general position. Then there exists a unique Delaunay triangulation, which consists of n -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 14. Suppose $x_1, \dots, x_N \in \mathbb{R}^n$ are sampled iid from a distribution P and $P \ll \text{Leb}$. Then with probability one over the joint distribution, the set $\{x_i\}_{i=1}^N$ is in general position, i.e., any $n + 1$ points are affinely independent.

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

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

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

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

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.

Also, P is absolutely continuous with respect to Lebesgue measure Leb . Thus, $P(\mathcal{N}) = 0$, i.e.,

$$\mathbb{P}((x_0, \dots, x_n) | F(x_0, \dots, x_n) = 0) = 0.$$

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. $\square$

F.2 DATA-DEPENDENT POLYNOMIAL APPROXIMATION IN SOBOLEV SPACE

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

Theorem 13 (adaptive Bramble-Hilbert). Let $m > n/p$, $1 \leq p < \infty$. 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., $\mathcal{I} = \text{conv}\{x_1, \dots, x_N\} \subseteq \Omega$. For any $\phi \in W^{m,p}(\mathcal{I}^\circ)$, there exists a polynomial interpolation ψ of order $p(N, n)$, s.t.,

$$\|\phi - \psi\|_{r,p} \leq C \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p}$$

where $h_{\min} = \min_i h_{i,\min}$ and $h_{\max_{n+1}} = \max_{\mathcal{I}_i} h_{i,\max}$ is the maximum h_{ij} over all the n -simplex.

The proof is an adaptation of Watkins (1979).

Proof. For any n -simplex $\mathcal{I}_i$, define $\Delta v_i = (v_{i1} - v_{i0}, v_{i2} - v_{i0}, \dots, v_{in} - v_{i0}) \in \mathbb{R}^{n \times n}$, and $h_{ij} = \|v_{ij} - v_{i0}\|$. Also, $\mathcal{I} = \bigcup_i \mathcal{I}_i$.

For $x \in \mathcal{I}_i$, define an affine map $x = v_0 + \Delta v_i \xi$, where $\xi \in \mathcal{D} = \{\xi \in \mathbb{R}^n | \xi \in [0, 1]^n, \sum_{i=1}^n \xi_i \leq 1\}$. Then this map is a one-to-one mapping of $\mathcal{D}$ onto $\mathcal{I}_i$. Let the vertices of $\mathcal{D}$ be $\bar{v}_{ij}$ for $j = 0, \dots, n$. Then for any function ϕ on $\mathcal{I}$, define $\bar{\phi}(\xi) = \phi(x) = \phi(v_0 + \Delta v_i \xi)$. Then

$$\|D^\alpha \phi(x)\|_{0,p,\mathcal{I}_i^\circ} = \prod_{j=1}^n h_{ij}^{-\alpha_j} |J_i|^{1/p} \|D^\alpha \bar{\phi}(\xi)\|_{0,p,\mathcal{D}^\circ}$$

where $|J_i|$ is the determinant of the Jacobian of the affine map. Then, we have

$$\begin{aligned}\|\bar{\phi}(\xi)\|_{m,p,\mathcal{D}^\circ} &= \left(\sum_{|\alpha| \leq m} \|D^\alpha \bar{\phi}\|_{0,p,\mathcal{D}^\circ}^p \right)^{1/p} \\ &= \left(\sum_{|\alpha| \leq m} \left(\prod_{j=1}^n h_{ij}^{\alpha_j} |J_i|^{-1/p} \right)^p \|D^\alpha \phi(x)\|_{0,p,\mathcal{I}_i^\circ}^p \right)^{1/p} \\ &\leq \max\{h_{i,\max}^m, 1\} |J_i|^{-n/p} \|\phi(x)\|_{m,p,\mathcal{I}_i^\circ}\end{aligned}\quad (24)$$

and similarly,

$$\|\phi(x)\|_{m,p,\mathcal{I}_i^\circ} \leq \max\{h_{i,\min}^{-m}, 1\} |J_i|^{n/p} \|\bar{\phi}(\xi)\|_{m,p,\mathcal{D}^\circ} \quad (25)$$

where $h_{i,\min} = \min_j h_{ij}$, and $h_{i,\max} = \max_j h_{ij}$.

For any function $\phi \in C(\mathcal{I})$, we can define a unique polynomial interpolant $\psi = B\phi$, where $\psi \in C(\mathcal{I})$ is a polynomial with degree $\leq p(N, n)$ for some $p(N, n) \geq m-1$ such that $p_{m-1} = Bp_{m-1}$ for any polynomial p_{m-1} of order at most $m-1$, and the operator B is linear (see for example Saniee (2008)). The two functions ϕ, ψ are equal at the vertices of the n -simplex, i.e., $\phi(v_{ij}) = \psi(v_{ij})$, for all $j = 0, \dots, n$.

Let $\bar{\phi}(\xi) = \phi(x)$ and $\bar{\psi}(\xi) = \psi(x)$. We can define the linear operator $\bar{B}$, s.t.,

$$(\phi - B\phi)(x) = (\bar{\phi} - \bar{B}\bar{\phi})(\xi).$$

Then, we can find a set of barycentric coordinate functions $\bar{\phi}_0(\xi), \dots, \bar{\phi}_n(\xi)$, which is the basis function of a natural coordinate system for n -simplex, s.t.

$$\bar{\phi}_s(\bar{v}_{ij}) = \delta_{sj} \text{ and } \bar{\psi}(\xi) = \sum_{j=0}^n \bar{\phi}_j(\xi) \bar{\phi}(\bar{v}_{ij}).$$

Then

$$\begin{aligned}\|\bar{B}\bar{\phi}\|_{r,p,\mathcal{D}^\circ} &= \|\bar{\psi}\|_{r,p} \leq \max_j \|\bar{\phi}(\bar{v}_{ij})\| \sum_{s=0}^n \|\bar{\phi}_s\|_{r,p,\mathcal{D}^\circ} \\ &\leq \|\bar{\phi}\|_\infty \sum_{s=0}^n \|\bar{\phi}_s\|_{r,p,\mathcal{D}^\circ} \\ &\leq C' \|\bar{\phi}\|_{m,p,\mathcal{D}^\circ} \sum_{s=0}^n \|\bar{\phi}_s\|_{r,p,\mathcal{D}^\circ} = C'' \|\bar{\phi}\|_{m,p,\mathcal{D}^\circ}\end{aligned}$$

where the last inequality follows from the Sobolev embedding theorem, and $C', C'' > 0$. This shows that $\bar{B}$ is a bounded operator from $W^{m,p}(\mathcal{D}^\circ)$ to $W^{r,p}(\mathcal{D}^\circ)$, and so is $I - \bar{B}$, i.e.,

$$\|(I - \bar{B})\bar{\phi}\|_{r,p,\mathcal{D}^\circ} \leq C''' \|\bar{\phi}\|_{m,p,\mathcal{D}^\circ} \quad (26)$$

for some universal constant $C''' > 0$.

Then

$$\begin{aligned}\|\phi - \psi\|_{r,p,\mathcal{I}_i^\circ} &= \|\phi - B\phi\|_{r,p,\mathcal{I}_i^\circ} \leq \max\{h_{i,\min}^{-r}, 1\} |J_i|^{n/p} \|\bar{\phi} - \bar{B}\bar{\phi}\|_{r,p,\mathcal{D}^\circ} \\ &= \max\{h_{i,\min}^{-r}, 1\} |J_i|^{n/p} \|(I - \bar{B})\bar{\phi}\|_{r,p,\mathcal{D}^\circ} \\ &\leq C''' \max\{h_{i,\min}^{-r}, 1\} |J_i|^{n/p} \|\bar{\phi}\|_{m,p,\mathcal{D}^\circ} \\ &\leq C''' \max\{h_{i,\max}^m, 1\} \max\{h_{i,\min}^{-r}, 1\} \|\phi\|_{m,p,\mathcal{I}_i^\circ}\end{aligned}$$

where the first inequality follows from (25), the second inequality follows from (26), the last inequality follows from (24).

Then we have for any $\phi \in W^{m,p}(\mathcal{I}^\circ)$,

$$\begin{aligned}\|\phi - \psi\|_{r,p} &\leq C \max\{h_{\max_{n+1}}^m, 1\} \max\{h_{\min}^{-r}, 1\} \|\phi\|_{m,p} \\ &= C \max\{h_{\max_{n+1}}, 1\}^m \min\{h_{\min}, 1\}^{-r} \|\phi\|_{m,p}\end{aligned}$$

where $h_{\min} = \min_i h_{i,\min}$ and $h_{\max_{n+1}} = \max_{\mathcal{I}_i} h_{i,\max}$ is the maximum h_{ij} over all the n -simplex, and $C > 0$ is some universal constant. $\square$

G ADDITIONAL LEMMAS

Lemma 15. *Let Ω be a bounded open set. Then the Sobolev space $W^{m,p}(\Omega)$ is dense in $L^p(\Omega)$.*

Proof. By definition, $W^{m,p}(\Omega) \subseteq L^p(\Omega)$. Also, we have $C^\infty(\Omega)$ is dense in $L^p(\Omega)$, and in $W^{m,p}(\Omega)$. Then $W^{m,p}(\Omega)$ is dense in $L^p(\Omega)$. $\square$

Lemma 16. *Under Assumption 2, the neural network $f(x) \in W^{r,p}(\Omega)$ for any $1 \leq r, p < \infty$.*

Proof. By Assumption 2 and Proposition 1, the neural network function $f(x)$ is analytic on some open neighbourhood Ω_1 of Ω . Then $f(x) \in C^\infty(\Omega_1)$, and therefore

$$\sup_{x \in \Omega} \|D^\alpha f(x)\|_{\ell_p} < \infty$$

for any non-negative $\alpha = (\alpha_1, \dots, \alpha_d)$, and any $1 \leq p < \infty$. By $|\Omega| < \infty$, we have the $L^p(\Omega)$ norm of any derivative

$$\|D^\alpha f(x)\|_p^p \leq |\Omega| \sup_{x \in \Omega} \|D^\alpha f(x)\|_{\ell_p}^p < \infty.$$

Therefore, $f(x) \in W^{r,p}(\Omega)$ for any $1 \leq r, p < \infty$. $\square$

H EXTENDED EXPERIMENTS

In our experiments, we first introduce the data uniformity strategy—iteratively selecting data points that maximize the distances among all previously chosen points, and present visualizations comparing the selected uniform data, randomly selected data of the same size, and the full dataset. The purpose of comparing the selected uniform data with randomly selected data of the same size is to isolate the effect of the minimum distance $h_{\min}$ while keeping the dataset size constant. We then conduct supervised fine-tuning with two complementary objectives: ℓ^2 loss optimized via SGD for theoretical validation, and cross-entropy loss optimized via Adam for practical evaluation. This setting enables us to systematically assess both performance and training efficiency across varying source datasets, optimization strategies, dataset scales, and model sizes, thereby demonstrating the robustness of our method. Specifically, we deploy our approach into state-of-the-art data distillation and data selection baselines, including TeaMs-RL (Gu et al., 2024), WizardLM (Xu et al., 2024), Zcore (Griffin et al., 2024), and LESS (Xia et al., 2024). Comparative experiments show that our method not only improves training efficiency but also achieves superior or comparable performance relative to these strong baselines. The experiments validate our theoretical results by showing that training with the selected uniform data, which implies a larger $h_{\min}$, accelerates convergence (see Corollary 3) while maintaining comparable performance to using the full dataset (see Theorem 3 and the discussions below).

H.1 DATA SELECTION AND VISUALIZATION

Data Selection via Distance Maximization. To construct compact yet representative instruction-tuning datasets, we apply a greedy uniformity selection strategy based on data distance. Specifically, we represent each instruction–input–output triple as an average Word2Vec embedding trained on the full corpus. Starting from a randomly selected seed, we iteratively add the data point that maximizes the minimum distance to all previously selected points, ensuring maximal coverage of the data space. We apply this strategy to four baselines: TeaMs-RL (Gu et al., 2024), WizardLM (Xu et al., 2024), Zcore (Griffin et al., 2024), and LESS (Xia et al., 2024). The TeaMs-RL and WizardLM datasets are both distilled from ChatGPT-3.5/4, using different strategies to generate diverse instructional data. From the 9k TeaMs-RL dataset, we select a 4.5k maximized distance subset and compare it to a 4.5k random sample, which is composed of half randomly sampled data points and half selected to minimize pairwise distances from the 9k full dataset³; from the 20k WizardLM dataset, we select a 10k maximized distance subset comparing it against a 10k random sample, and the

³This choice of random data is made because the dataset size is relatively small, and we aim to highlight the effect of different minimum distances between data points $h_{\min}$.

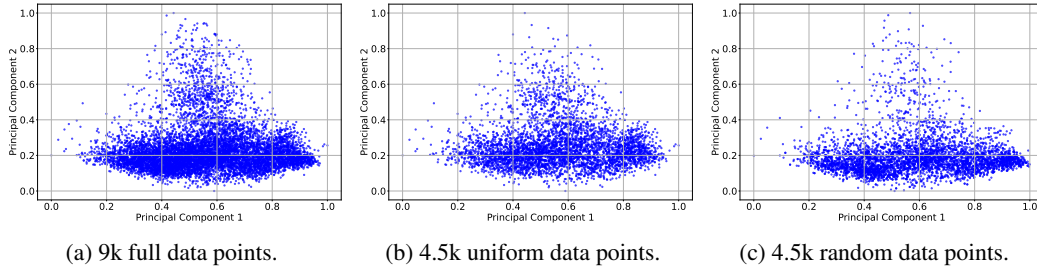


Figure 7: Visualization of data point distributions for datasets selected using different methods from TeaMs-RL (Gu et al., 2024).

random subset consists of the first 10k samples from the original 20k dataset, which is already randomly shuffled. This approach allows us to fairly evaluate the benefits of uniformity-oriented sampling across different dataset scales. Beyond instruction-tuning datasets, we also apply the same strategy to the California Housing dataset (Pace & Barry, 1997) and the LESS dataset (Xia et al., 2024), thereby enabling direct benchmarking against baselines such as Zcore (Griffin et al., 2024) and LESS (Xia et al., 2024).

Data Visualization via PCA Projection. To qualitatively assess the distributional coverage of different data selection methods, we project high-dimensional sentence embeddings into two dimensions using PCA (Maćkiewicz & Ratajczak, 1993). For consistency, PCA is fit on the full dataset (either TeaMs-RL 9k or WizardLM 20k) and applied to both the full set and its corresponding uniform and random subsets (the uniform data is selected based on maximizing pairwise distance). The resulting 2D coordinates are normalized using MinMax scaling to $[0,1]$ and plotted for visualization. As shown in Figures 3 and 7, the uniform subsets from TeaMs-RL (Gu et al., 2024) and WizardLM (Xu et al., 2024) (4.5k for TeaMs-RL, 10k for WizardLM) exhibit broader and more dispersed coverage across the embedding space compared to the random subsets and the full dataset, which cluster more tightly and redundantly. These patterns highlight the ability of uniformity-based sampling to retain semantic coverage with reduced sample sizes, supporting its utility for efficient training. A similar strategy is employed in Zcore (Griffin et al., 2024) and LESS (Xia et al., 2024).

H.2 ℓ^2 -SGD TRAINING EXPERIMENTS

We present training dynamics under ℓ^2 loss with SGD to validate our theoretical insights. While ℓ^2 -(S)GD provides a useful controlled setting for analysis, it is generally less efficient in practice compared to cross-entropy loss optimized with Adam, which is a widely used approach for training LLMs (Gunel et al., 2021; Mao et al., 2023).

Training Comparison with Uniform Subsets. Figure 8 shows the training loss curves of LLaMA-1-7B (Touvron et al., 2023)⁴ using ℓ^2 loss with SGD (Bottou, 2010; Cortes et al., 2012) on the TeaMs-RL dataset under three data configurations: the full dataset (9k), a 4.5k subset selected by maximizing pairwise distance, and a 4.5k randomly sampled subset. The figure includes both unsmoothed (left) and smoothed (right) loss curves, where unsmoothed curves reflect raw step-wise fluctuations in loss, while smoothed curves (e.g., via moving average) highlight overall training trends more clearly. In both plots, we zoom in the early and middle stage of the training for clearer comparison (see full training in Appendix I), and observe that the model trained on the uniform subset consistently converges faster than the one trained on the random subset across training steps. Moreover, despite using only half the number of samples, the uniform subset achieves convergence behavior comparable to or better than the full dataset. This validates our theory (see Section 4) and suggests that uniformity-aware data selection can substantially enhance training efficiency and convergence, even under limited data budgets. More experiments are provided in Appendix I.

⁴<https://huggingface.co/huggyllama/llama-7b>

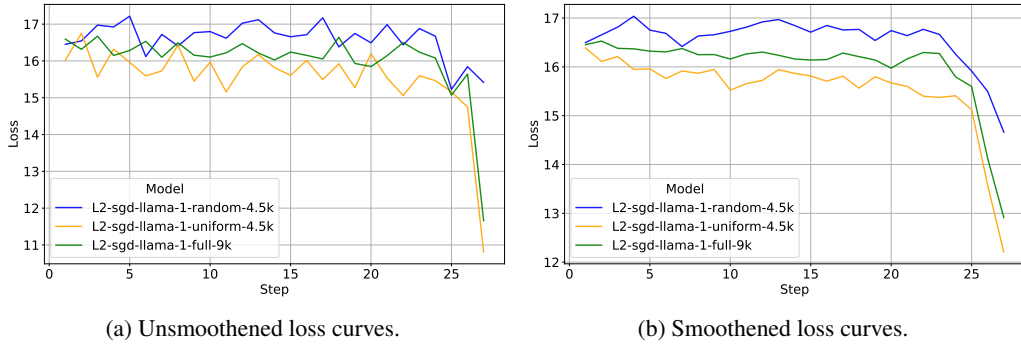


Figure 8: ℓ^2 SGD Training Experiments: Training loss comparison of TeaMs-RL data point distributions for different dataset sizes: 4.5k uniform (maximized distance) dataset, 4.5k random dataset, 9k full dataset.

H.3 CROSS ENTROPY LOSS ADAM TRAINING EXPERIMENTS

Beyond the ℓ^2 -SGD, we conduct the experiments using cross-entropy loss with Adam, the standard optimization setup for LLMs (Gunel et al., 2021; Mao et al., 2023), we observe consistent improvements. As shown in Figure 6 presents training dynamics of LLaMA-1-7B using cross-entropy loss with AdamW on the LESS dataset (Xia et al., 2024), comparing uniform subsets (ours), random subsets, and the full dataset. Panel (a) shows that the 7k uniform subset reaches the loss threshold more than $2\times$ faster than the 14k full dataset and substantially faster than the 7k random subset, highlighting improved training efficiency. Panel (b) reports MMLU accuracy, where the uniform subset achieves performance comparable to or exceeding both random and full datasets despite using fewer samples. Panel (c) illustrates training loss trajectories across seeds, demonstrating that uniform subsets yield smoother and more stable convergence relative to baselines. Together, these results indicate that uniformity-aware data selection not only reduces training time but also improves robustness and generalization, validating its effectiveness for efficient large-scale model training.

Cross-Entropy Loss and Adam Optimization with maximized-distance Subsets. Figures 9 (a) and 9 (b) compare training loss curves under cross-entropy loss (Mao et al., 2023) using the Adam optimizer (Kingma & Ba, 2014) for three TeaMs-RL datasets: the full 9k dataset, a 4.5k maximized-distance subset selected via maximized distance, and a 4.5k random subset. Figure 9 (a) presents raw training loss, while Figure 9 (b) shows smoothed loss for clarity. In both cases, the model trained on the maximized-distance based subset consistently converges faster and achieves significantly lower loss than the one trained on the random subset, despite both using the same number of samples. Notably, the maximized-distance 4.5k subset also outperforms the full 9k dataset in both convergence speed and final loss, highlighting the efficiency gains from data uniformity and well-distributed data. These results reaffirm that maximized-distance data selection not only improves optimization but also reduces training redundancy under standard cross-entropy objectives with modern optimizers. Notably, models trained on uniform data exhibit lower initial loss compared to those trained on random or full datasets. This may be attributed to the higher diversity and broader coverage of the uniform subset, which is more aligned with the base model’s existing knowledge, likely due to its pretraining on large-scale corpora exceeding 1T tokens (Touvron et al., 2023).

Scalability of Uniform Data Benefits with Larger Datasets. Figures 10 (a) and 10 (b) show the training dynamics of models trained on WizardLM data using cross-entropy loss with the Adam optimizer across three dataset settings: the full 20k dataset, a 10k maximized-distance subset selected by maximizing distance, and a 10k random subset. In both plots, the maximized-distance subset consistently outperforms the random counterpart throughout the training process, achieving lower and more stable loss values. Notably, the maximized-distance 10k subset even surpasses the full 20k dataset regarding convergence, suggesting that data uniformity can compensate for and even outperform brute-force data scaling. This pattern, observed across both TeaMs-RL and WizardLM data, highlights the robustness and scalability of uniform-based sampling strategies, offering

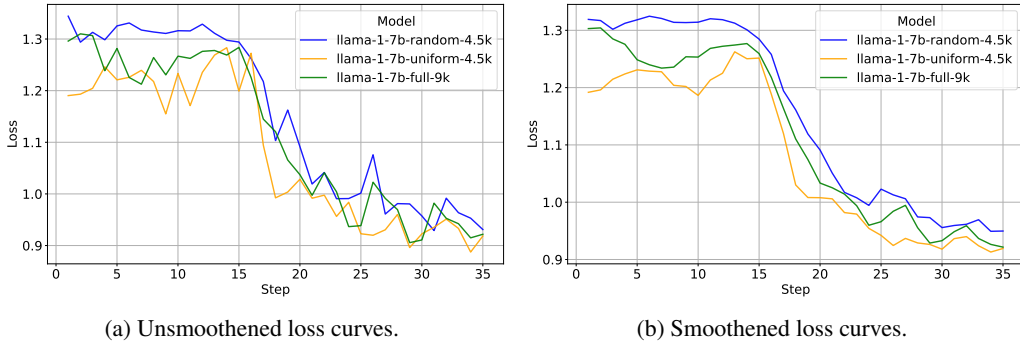


Figure 9: Training loss comparison of TeaMs-RL data point distributions for different dataset sizes using cross entropy loss and Adam: 4.5k uniform (maximized distance) dataset, 4.5k random dataset, 9k full dataset.

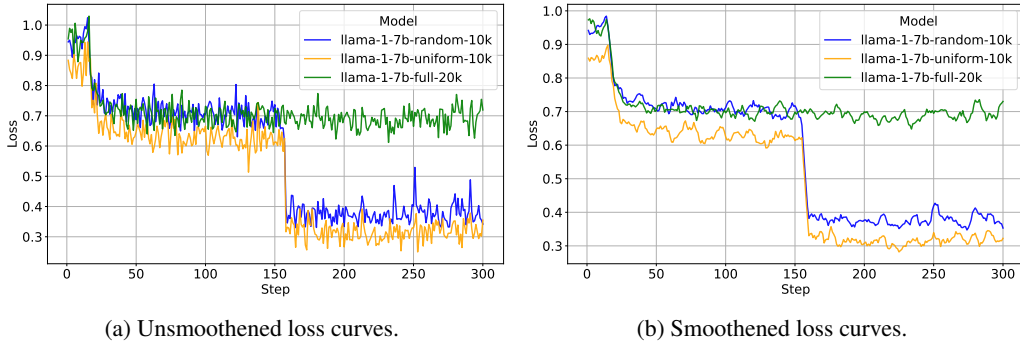


Figure 10: Training loss comparison of WizardLM data point distributions for different dataset sizes using cross entropy and adam: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

a compelling approach for reducing computational costs without sacrificing learning quality. The complete training experiments across all steps are provided in Appendix I.4.

Training on large-scale models. Figures 11 through 12 present training loss comparisons for the LLaMA-1 13B model (Touvron et al., 2023)⁵ using full datasets, random subsets, and uniformity-selected subsets across TeaMs-RL (4.5k/9k) and WizardLM (10k/20k) data. In all settings, the models trained on uniformity-maximized subsets exhibit consistently lower or comparable loss compared to those trained on full datasets, and significantly outperform the randomly sampled subsets. This trend holds both with and without smoothing applied to the loss curves. Particularly in Figures 12 (a) and 12 (b), the 10k uniform subset achieves better final loss than even the full 20k dataset, reinforcing the efficiency of data uniformity in training large-scale models. These results demonstrate that uniformity-aware sampling remains effective even at larger model scales, offering substantial gains in training efficiency without compromising performance. The complete training experiments across all steps are provided in Appendix I.4.

H.4 PERFORMANCE AND TRAINING TIME COMPARISON

Evaluation Performance with Diverse Subsets. Figures 13 and 14 present evaluation results of LLaMA-1-13B on the ARC Challenge (Clark et al., 2018) and TruthfulQA MC benchmarks (Lin et al., 2021), using datasets from TeaMs-RL and WizardLM, respectively. In both settings, models trained on data selected via maximum pairwise distance (orange bars) consistently outperform those trained on random subsets (green) and the original baselines (blue). Notably, the 4.5k and 10k uniform subsets achieve performance comparable to or slightly better than the full datasets

⁵<https://huggingface.co/huggyllama/llama-13b>

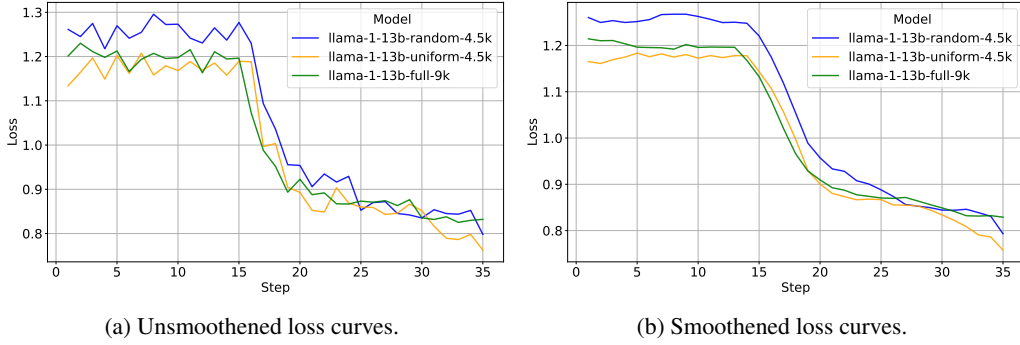


Figure 11: Training loss comparison of TeaMs-RL data point distributions for different dataset sizes using cross entropy loss and Adam on llama-1-13b models: 4.5k uniform (maximized distance) dataset, 4.5k random dataset, 9k full dataset.

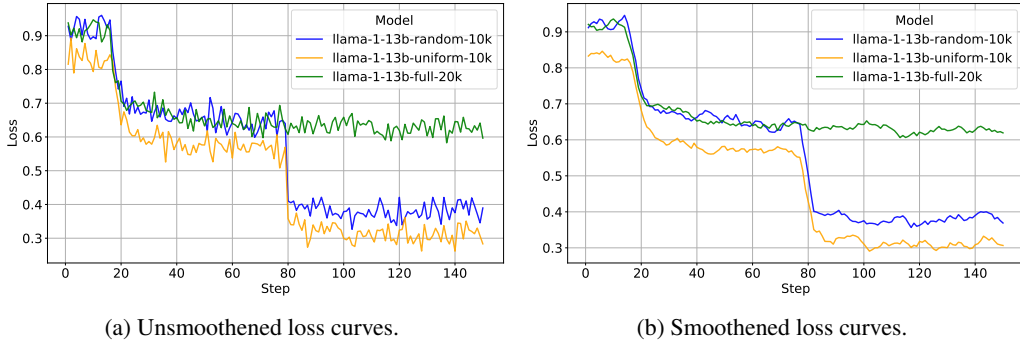


Figure 12: Training loss comparison of WizardLM data point distributions for different dataset sizes using cross entropy and adam on llama-1-13b models: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

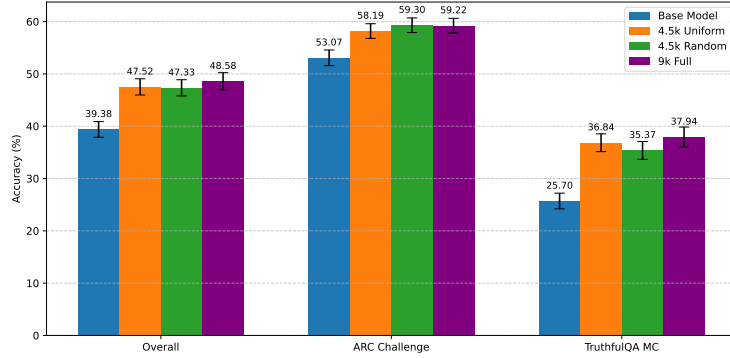


Figure 13: ARC and MC performance on Llama-1-13b using TeaMs-RL diverse datasets.

(red) while using significantly fewer samples. This is particularly evident on the ARC Challenge task (e.g., Figure 15), where the uniform subsets closely match or exceed the accuracy of the full datasets. These results further validate the effectiveness of uniformity-based data selection in improving model generalization and sample efficiency for LLM training. Note that models trained on different datasets may reach different loss levels at the same training steps; therefore, we select checkpoints with approximately equal loss values to ensure a fair performance comparison, e.g., 0.8733 for the TeaMs-RL dataset and 0.7058 for the WizardLM dataset.

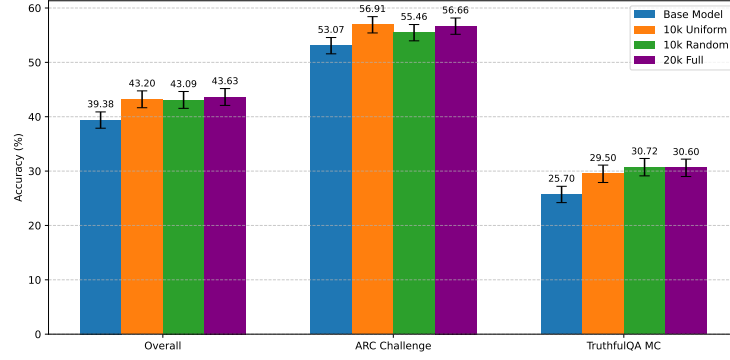


Figure 14: ARC and MC performance on Llama-1-13b using WizardLM diverse datasets.

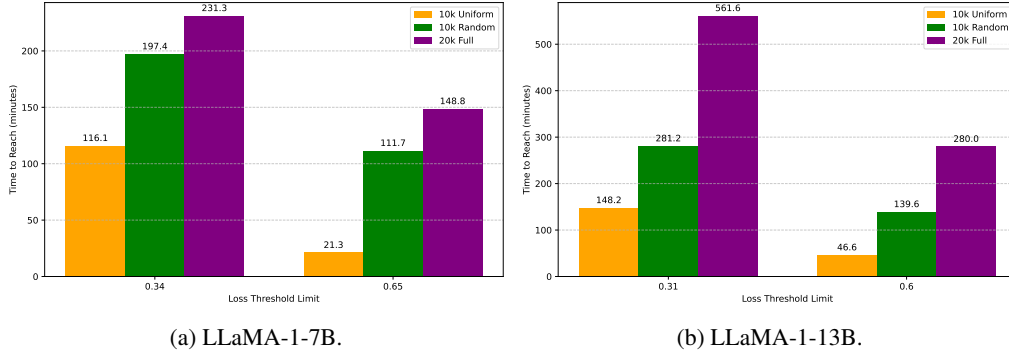


Figure 15: **Training time to reach loss thresholds using a 3-step moving average, measured on a single A100 GPU.** For both (a) LLaMA-1-7B (0.34, 0.65) and (b) LLaMA-1-13B (0.31, 0.6), we compare the convergence speed of Uniform, Random, and Full data selection strategies. Uniform sampling consistently achieves faster convergence, indicating more efficient training.

Time to Threshold: Faster Convergence with Uniform Data Selection. As shown in Figures 15 (a) and (b), we compare the convergence efficiency of three data selection strategies—Uniform, Random, and Full, under different loss thresholds using LLaMA-1-7B and LLaMA-1-13B. All times are measured based on wall-clock time using a single A100 GPU. For LLaMA-1-7B, Uniform reaches the 0.34 and 0.65 loss thresholds in 116.1 and 21.3 minutes, respectively, outperforming Random (197.4, 111.7) and Full (231.3, 148.8). Similarly, LLaMA-1-13B trained on the 10k Uniform subset reaches the 0.31 and 0.6 thresholds in 148.2 and 46.6 minutes, significantly faster than Random (281.2, 139.6) and 20k Full (561.6, 280.0). These results highlight that Uniform selection not only reduces the dataset size but also accelerates training, indicating a more efficient optimization path and improved data quality.

I SUPPLEMENTARY EXPERIMENTS

I.1 ADDITIONAL EXPERIMENTS USING ℓ^2 -SGD

I.2 EFFECTIVENESS OF MAXIMIZED-DISTANCE BASED SAMPLING ON SMALLER SCALE DATASETS.

Figure 16 illustrates the ℓ^2 SGD training loss curves for LLaMA-1-7B across different data selection strategies on the TeaMs-RL dataset. We compare a 4.5k **uniform subset**, a 4.5k **random subset**, and the **full 9k dataset**. Subfigure (a) shows the raw loss trajectories, while (b) applies smoothing to better visualize convergence patterns. The 4.5k uniform subset achieves convergence comparable to or faster than the full dataset and consistently outperforms the random subset, particularly in the early stages of training. However, since this dataset size is relatively small, the advantage of data

uniformity is not very pronounced in the final stage of training. When the dataset size increases, uniform data selection becomes more effective in the later phase (see Figure 18).

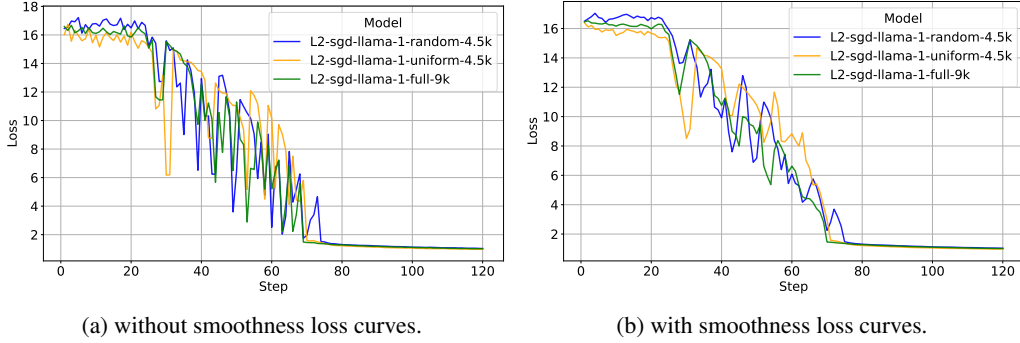


Figure 16: ℓ^2 SGD Training Experiments—All Curves: Training loss comparison of TeaMs-RL data point distributions for different dataset sizes: 4.5k uniform (maximized distance) dataset, 4.5k random dataset, 9k full dataset.

I.3 EFFECTIVENESS OF MAXIMIZED-DISTANCE BASED SAMPLING ON LARGER SCALE DATASETS.

Figures 17 and 18 show the ℓ^2 -SGD training loss curves of llama-1-7b models trained on the WizardLM dataset under three different data selection settings: the full 20k dataset, a 10k maximized-distance subset selected via maximization distance, and a 10k random subset. In both the raw (left) and smoothed (right) plots, the maximized-distance 10k subset outperforms the random 10k subset across training steps, achieving faster convergence and more stable convergence. Notably, despite being half the size, the maximized-distance subset often matches or even surpasses the performance of the full dataset in early and end training regarding the convergence. This result indicates the stability and efficiency of data uniformity sampling strategies, particularly in large-scale training tasks. It highlights the finding that uniform data offers stronger learning signals than merely increasing sample count without regard to data uniformity.

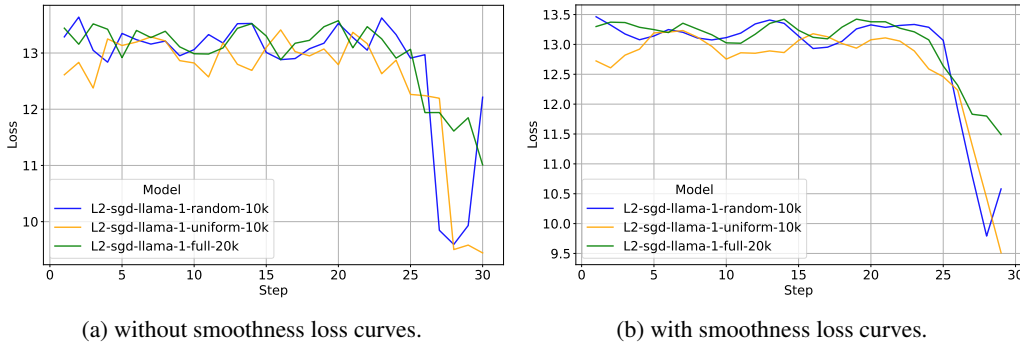


Figure 17: ℓ^2 SGD Training Experiments: Training loss comparison of WizardLM data point distributions for different dataset sizes: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

I.4 FULL STEP TRAINING: 7B AND 13B MODELS

As shown in Figure 19 and Figure 20, models generally converge with sufficient training time, but our method achieves faster convergence.

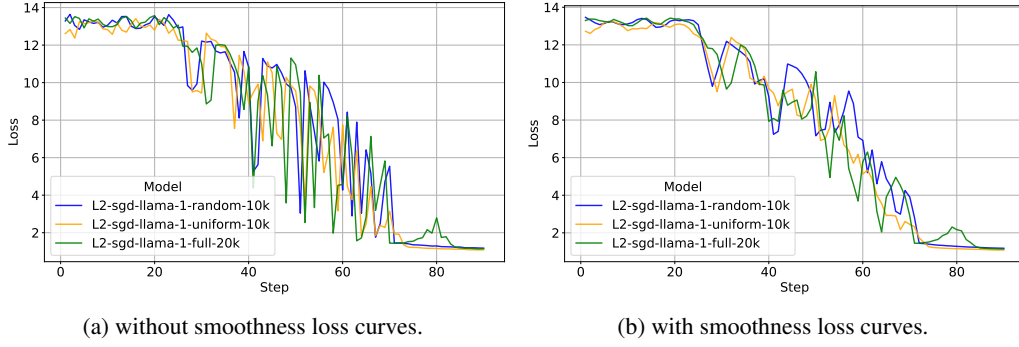


Figure 18: ℓ^2 SGD Training Experiments—All Curves: Training loss comparison of WizardLM data point distributions for different dataset sizes: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

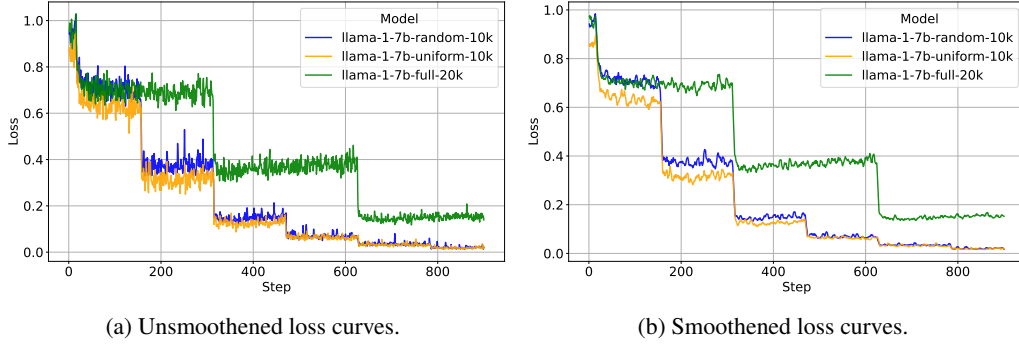


Figure 19: Full Steps: Training loss comparison of WizardLM data point distributions for different dataset sizes using cross entropy and adam: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

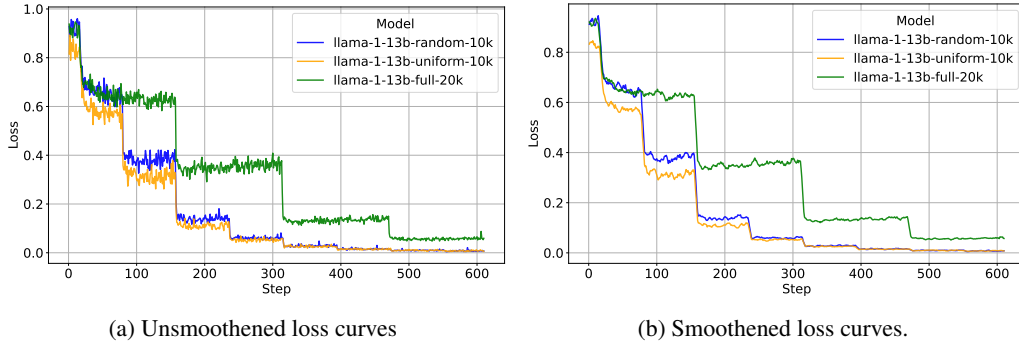


Figure 20: Full Steps: Training loss comparison of WizardLM data point distributions for different dataset sizes using cross entropy and adam on llama-1-13b models: 10k uniform (maximized distance) dataset, 10k random dataset, 20k full dataset.

I.5 ABLATION EXPERIMENT

As shown in Figure 21, it shows that data selection has a clear impact on training efficiency. Across both raw and smoothed curves, the minimize-10k datasets consistently reach lower loss than other datasets, and uniform-10k datasets achieve faster than other datasets, such as random-10k and 20k-full datasets, demonstrating that more principled selection can achieve stronger optimization. This indicates that additional low-quality or redundant data can hinder training efficiency. Overall, these

results highlight that careful data selection can yield both faster convergence and better final model quality, even with substantially less data.

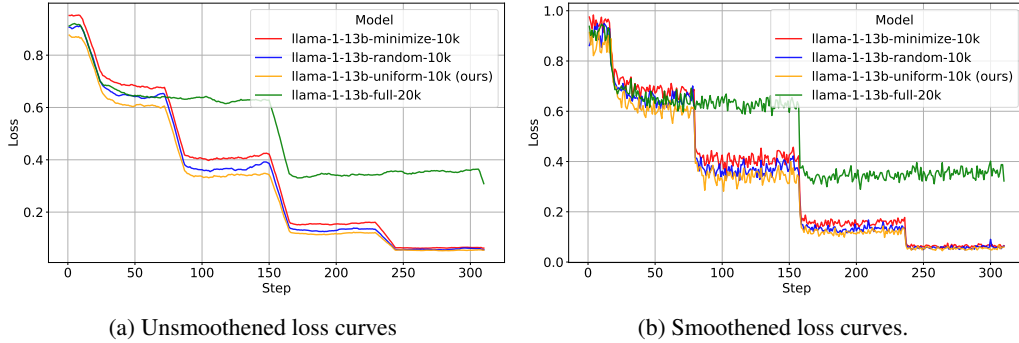


Figure 21: Training-efficiency ablation comparing data-selection strategies for Llama-1-13B. Left: raw (unsmoothed) loss curves. Right: smoothed loss curves. We compare minimize-10k, random-10k, uniform-10k (Ours), and full-20k training subsets.

As shown in Figure 22, this ablation examines whether the choice of embedding model affects the effectiveness of our uniform data-selection strategy. The results show that word2vec-based and sentence-BERT-based (s-bert) uniform sampling yield nearly identical training dynamics: both variants reduce loss at a similar rate, maintain stability throughout optimization, and converge to almost the same final value. The close overlap of the curves indicates that the uniformity principle itself, not the specific embedding space, is the primary factor driving performance gains. This suggests that our method is robust, lightweight, and flexible: simple embedding models like word2vec are already sufficient to realize the benefits of uniform selection, while stronger encoders like s-bert do not materially change the outcome.

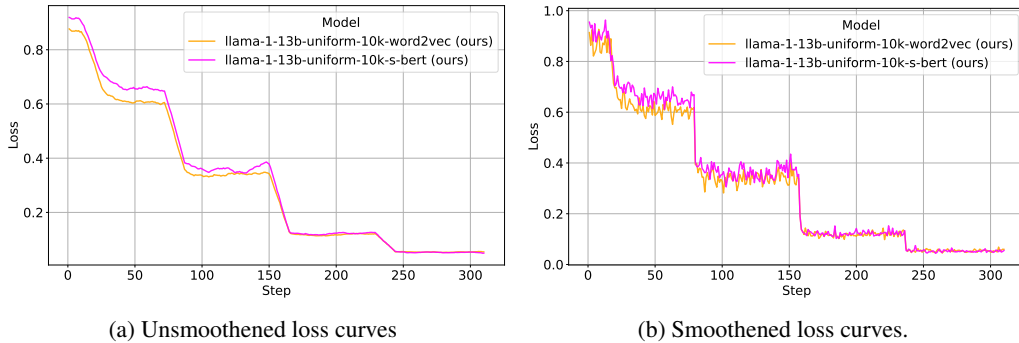


Figure 22: Ablation on embedding methods used in our uniform data-selection framework. Left: raw loss curves. Right: smoothed loss curves. We compare uniform-10k-word2vec (ours) and uniform-10k-s-bert (ours) for Llama-1-13B.

As shown in Figure 23, to ensure that the full-dataset and random-subset baselines train without exhibiting the gradient instabilities occasionally observed in default SGD settings (As shown in Figure 4), we fine-tuned their optimization hyperparameters, for example, by substantially reducing the learning rate (e.g., 0.01). While this stabilizes training, it also slows convergence and leads to noticeably less efficient learning. The results clearly show that our 8k Uniform (Ours) subset yields the strongest overall performance among all 8k-sized selections. Uniform sampling achieves faster training-loss reduction than both Random and Z-core, and it even surpasses the 16k Full baseline during the early and mid stages of optimization. On the validation set, the advantage of our method is more substantial: Uniform converges to the lowest MAE, demonstrating that it selects a more in-

formative and better-balanced subset that generalizes more effectively. In contrast, Random suffers from slow convergence, and Z-core plateaus early, while Full is consistently weaker than Uniform despite using twice the data. These results highlight that data quality and coverage are more important than dataset size, and that our uniform selection algorithm consistently provides the most efficient and generalizable training signal.

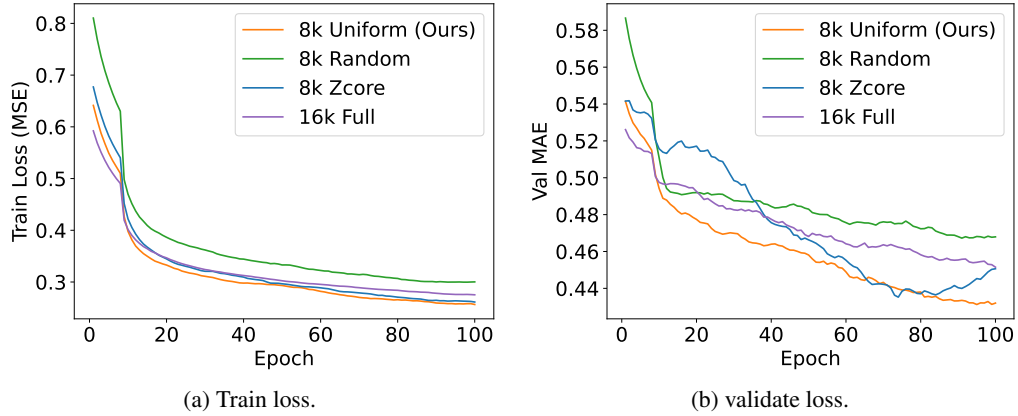


Figure 23: Training-efficiency ablation on the California Housing dataset comparing our uniform selection method with random, z-core, and full-data baselines (Smoothened loss). Left: training MSE loss. Right: validation MAE loss.

J EXPERIMENT SETTINGS

We summarize the key configurations⁶ used in our experiments to ensure reproducibility and clarity. Table 1 details the training hyperparameters. Based on these and building on Llama-X⁷, we train LLaMA-1-7B (one A100 GPU) and 13B (two A100 GPUs) models on datasets of varying sizes: WizardLM⁸ (10k uniform datasets, 10k random datasets, and 20k full datasets) and TeaMs-RL⁹ (4.5k uniform datasets, 4.5k random datasets, and 9k full datasets). The uniform subsets from TeaMs-RL and 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. For TeaMs-RL, the random subset is composed of half randomly sampled data points and half selected to minimize pairwise distances from the 9k full dataset. In the case of WizardLM, the random subset consists of the first 10k samples from the original 20k dataset.

Table 2 shows the DeepSpeed configuration using cross-entropy loss with the Adam optimizer. Table 3 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 4 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.

Table 1: 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

⁶<https://anonymous.4open.science/r/data-uniformity-1A5C>

⁷<https://github.com/AetherCortex/Llama-X>

⁸https://huggingface.co/datasets/WizardLMTeam/WizardLM_evol_instruct_v2_196k

⁹<https://github.com/SafeRL-Lab/TeaMs-RL>

Table 2: 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 3: 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 4: 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