

# UNDERSTANDING LEARNING DYNAMICS OF ZERO-ORDER OPTIMIZATION

Zhe Li<sup>1</sup>, Bicheng Ying<sup>2</sup>, Zidong Liu<sup>3</sup>, Haibo Yang<sup>1</sup>

<sup>1</sup>Rochester Institute of Technology, Rochester, NY 14623, USA

<sup>2</sup>Google, Los Angeles, CA 90034, USA

<sup>3</sup>ComboCurve Inc., Houston, TX 77005, USA

z14063@rit.edu, ybc@google.com, z.liu@combocurve.com,

hbycis@rit.edu

## ABSTRACT

We derive the one-step learning dynamics of ZO SGD, where the empirical Neural Tangent Kernel (eNTK) naturally emerges as the key term governing the learning behavior. Inspection of the eNTK produced by ZO-SGD reveals that each element corresponds to the inner product of neural tangent vectors projected onto a random low-dimensional subspace. Thus, by invoking the Johnson-Lindenstrauss Lemma, our analysis shows that the fidelity of the ZO eNTK is governed primarily by the number of perturbations. Crucially, the approximation error depends on the model output size rather than the massive parameter dimension. This dimension-free property provides a theoretical justification for the scalability of ZO methods to LLMs finetuning tasks. We believe that this kernel-based framework offers a novel perspective for understanding ZO methods within the context of learning dynamics.

## 1 INTRODUCTION

By estimating gradients solely through function evaluations, Zeroth-order (ZO) optimization (Spall, 2002; Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017) methods offer a memory-efficient (Malladi et al., 2023a; Chen et al., 2025) and communication-efficient (Qin et al., 2024; Li et al., 2025a;b) alternative to first-order (FO) algorithms. These properties have made ZO optimization increasingly popular for deploying large-scale models on resource-constrained edge devices and for black-box adversarial attacks (Chen et al., 2017; Liu et al., 2020). Despite these practical advantages, ZO methods have historically faced a significant theoretical disadvantage compared to their gradient-based counterparts (Spall, 2002; Conn et al., 2009). Classical optimization theory establishes that ZO algorithms suffer from a dimension-dependent slowdown, with convergence rates typically scaling with the model dimension  $d$  (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017; Shamir, 2017). In the worst-case scenarios, the variance of the gradient estimator scales linearly with  $d$ , suggesting that ZO optimization should be prohibitively slow for high-dimensional models (Duchi et al., 2015).

Yet, recent empirical breakthroughs contradict this pessimistic theoretical outlook. A growing body of work reports the successful application of ZO methods to fine-tuning LLMs with billions of parameters (Malladi et al., 2023a; Chen et al., 2024; Zhao et al., 2025). In these high-dimensional tasks, ZO algorithms frequently achieve performance competitive with FO methods, exhibiting convergence behaviors that defy the worst-case scaling laws predicted by classical analysis (Yu et al., 2025). This discrepancy suggests that the standard optimization perspective, which compresses learning dynamics into scalar loss values, fails to capture the structural nuances of how ZO updates drive knowledge acquisition in deep networks (Aghajanyan et al., 2021; Jin & Tan, 2026).

In this work, we resolve this paradox by investigating ZO optimization through the lens of the empirical Neural Tangent Kernel (eNTK) (Jacot et al., 2018). Rather than focusing solely on parameter-space convergence, we analyze the learning dynamics in function space. We show that the effective kernel induced by ZO optimization, which we term the ZO-eNTK, can be understood as a geometric projection of the standard FO eNTK onto a random subspace spanned by the perturbation vectors. This geometric perspective reveals a fundamental connection between ZO learning dynamics and the Johnson-Lindenstrauss (JL) Lemma (Johnson et al., 1984). By leveraging the JL Lemma, we derive rigorous approximation bounds between the ZO and FO optimization trajectories. Our analysis yields the following critical insights:

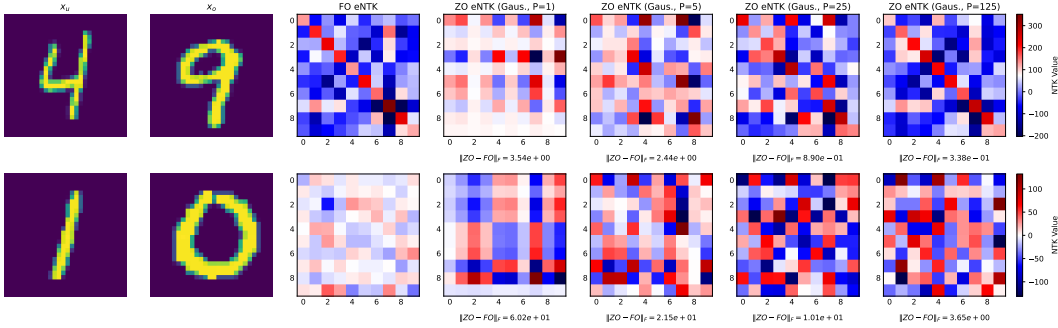


Figure 1: ZO eNTK v.s. FO eNTK. The pair with high similarity: 4 and 9. The pair with low similarity: 1 and 0. The relative Frobenius norm error  $\|ZO - FO\|_F = \frac{\|\mathcal{K}(\mathbf{x}_u, \mathbf{x}_o) - \mathcal{K}(\mathbf{x}_u, \mathbf{x}_o; U_t, P)\|_F}{\|\mathcal{K}(\mathbf{x}_u, \mathbf{x}_o)\|_F}$ .

- The fidelity of the ZO eNTK is mainly governed by the number of perturbations.
- Dimension Independence: Most importantly, we show that the efficacy of the ZO method depends not on the massive model dimension  $d$ , but rather on the output vocabulary size  $V$ . This dimension-free property provides a theoretical justification for the scalability of ZO methods to LLMs, explaining why they remain efficient even when  $d$  is vast.

Beyond these specific insights, we believe another contribution of this work lies in establishing a kernel-based framework for ZO optimization, offering a novel perspective that complements the existing optimization-centric view, enriching our understanding of derivative-free learning dynamics.

## 2 PRELIMINARIES OF LEARNING DYNAMICS AND ZO OPTIMIZATION

The standard empirical risk minimization problem in machine learning is typically formulated as  $\min_{\theta} \ell(\theta) = \sum_{(x,y) \in \mathcal{D}} \mathcal{L}(f_{\theta}(x), y)$ , where  $\theta$  is the model parameters,  $\ell$  is the objective function,  $\mathcal{D}$  is the dataset containing input-label pairs  $(x, y)$ , and  $\mathcal{L}$  is the loss function. From a purely optimization-centric perspective, learning algorithms are often analyzed solely through the objective function  $\ell$ , bringing considerable analytical simplicity. For example, the ZO (stochastic) gradient descent rule can be concisely written as

$$\theta_{t+1} = \theta_t - \eta \left[ \frac{\ell(\theta_t + \mu u_t) - \ell(\theta_t - \mu u_t)}{2\mu} u_t \right] \tag{1}$$

$$= \theta_t - \eta \langle \nabla \ell(\theta_t), u_t \rangle u_t + \mathcal{O}(\mu\eta), \tag{2}$$

where  $t$  is the iteration index,  $\eta$  is the learning rate,  $\mu > 0$  is the smoothing parameter, and  $u \in \mathbb{R}^d$  is the random perturbation vector sampled from a certain distribution (e.g., Gaussian). The estimator  $\frac{\ell(\theta_t + \mu u_t) - \ell(\theta_t - \mu u_t)}{2\mu} u_t$  corresponds to the Simultaneous Perturbation Stochastic Approximation (SPSA) gradient with central differences (Spall, 2002; Nesterov & Spokoiny, 2017). Eq. (2), derived via Taylor expansion, provides a more tractable form for analysis. Crucially, this formulation implies that for a sufficiently small smoothing radius  $\mu$ , the SPSA estimator serves as a proxy for the directional derivative along the random direction  $u_t$ .

While analytically convenient, an exclusive focus on optimization obscures structural information intrinsic to the learning process of modern ML models (Zhang et al., 2017; Belkin et al., 2019). More specifically, as shown in Eq. (1), standard optimization metrics focus exclusively on the objective  $\ell(\cdot)$ , compressing the complex interplay between the model  $f_{\theta}$  and the data  $(x, y)$  into a single scalar loss value. This reduction significantly obscures the structural nuances of the learning trajectory. Thus, it fails to capture how model updates influence predictions at the level of individual data points.

To bridge the gap between optimization and data, we adopt the framework of *learning dynamics*, which characterizes how a model’s confidence in the observed data evolves after a training step on the training sample (Ren & Sutherland, 2025). Consider a concrete supervised learning model case:

$$z = h_{\theta}(x), \quad f_{\theta}(x) = \text{Softmax}(z) \in \mathbb{R}^V,$$

where  $h(\cdot)$  denotes the neural network backbone, and  $V$  represents the vocabulary size or the number of classes. To quantify the model’s confidence in output  $y$  given input  $x$  at iteration  $t$ , we

define  $\pi_{\theta_t}(y|x)$ , or concisely  $\pi_t(y|x)$ . In a multi-class classification setting, this is expressed as:  $\pi_t(y|x) = \mathbb{1}_y^\top f_\theta(x) = \text{Softmax}(h_\theta(x))[y]$ , where  $\mathbb{1}_y$  is the one-hot encoding vector with the  $y$ -th entry equal to 1. For the NLP generation problem,  $x$  and  $y$  represent a sequence of wordings. We can reformulate the model belief as  $\pi_t(y|x) = \prod_{l=1}^L \mathbb{1}_{y_l}^\top f_\theta(x, y_{<l})$ , where  $y_{<l}$  is the model’s previous output sequence smaller than  $l$ . For clarity, this paper focuses on the multi-class classification case.

Now we are ready to present the learning dynamics for an observed data point  $\mathbf{x}_o$ , which is defined as the change in log-probability:  $\Delta \log \pi_t(y|\mathbf{x}_o) := \log \pi_{\theta_{t+1}}(y|\mathbf{x}_o) - \log \pi_{\theta_t}(y|\mathbf{x}_o)$ . Suppose we would like to study how the model’s prediction on  $\mathbf{x}_o$  changes after one-step update of ZO-SGD using  $\mathbf{x}_u$ . Applying a FO Taylor expansion yields:

$$\Delta \log \pi_t(y|\mathbf{x}_o) = \log \pi_t(y|\mathbf{x}_o) + \langle \nabla_\theta \log \pi_t(y|\mathbf{x}_o), \theta_{t+1} - \theta_t \rangle + \mathcal{O}(\|\theta_{t+1} - \theta_t\|^2) - \log \pi_t(y|\mathbf{x}_o).$$

For a sufficiently small learning rate, introducing the ZO-SGD update gives<sup>1</sup>

$$\begin{aligned} \langle \nabla_\theta \log \pi_t(y|\mathbf{x}_o), \theta_{t+1} - \theta_t \rangle &\approx -\eta [\nabla_\theta \log \pi_t(y|\mathbf{x}_o)] u_t u_t^\top \nabla_\theta \mathcal{L}(f_\theta(\mathbf{x}_u), \mathbf{y}_u) \\ &= -\eta \underbrace{\nabla_z \log \pi_t(y|\mathbf{x}_o)}_{V \times V} \underbrace{[\nabla_\theta z(\mathbf{x}_o)]^\top}_{V \times d} \underbrace{u_t u_t^\top}_{d \times d} \underbrace{\nabla_\theta z(\mathbf{x}_u)}_{d \times V} \underbrace{\nabla_z \mathcal{L}(z, \mathbf{y}_u)}_{V \times 1} \\ &= -\eta \mathcal{A}_t(\mathbf{x}_o) \mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u; u_t) \mathcal{G}_t(\mathbf{x}_u, \mathbf{y}_u), \end{aligned} \quad (3)$$

where  $\mathcal{A}_t(\mathbf{x}_o) = I - \mathbb{1}_{\pi_{\theta_t}(\mathbf{x}_o)}^\top$  relies only on the model’s prediction on the observing data  $\mathbf{x}_o$ , and  $\mathcal{G}_t$  is the last layer’s gradient of the previous updating data point. The central quantity in Eq. (3) is  $\mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u; u_t)$ , which can be interpreted as an eNTK of the logits  $z$ , projected onto the random perturbation  $u_t$ . Comparing Eq. (3) with the FO learning dynamics kernel  $\mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u) := \nabla_\theta z(\mathbf{x}_o)^\top \nabla_\theta z(\mathbf{x}_u)$  studied in (Jacot et al., 2018; Ren & Sutherland, 2025), we observe that the key distinction between ZO and FO SGD lies in the presence of the random perturbation  $u_t$ .

This perspective naturally motivates a kernel-based formulation of ZO optimization via  $\mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u; u_t)$ , enabling a systematic analysis of how different design choices affect learning behavior. For clarity, we begin by analyzing the single-perturbation case. However, since practical applications typically employ multi-perturbation strategies, we will subsequently extend our analysis to that setting. Before delving into the theoretical rigor of the general multi-perturbation case, we first examine empirical results to gain intuitive insights from the convergence trajectories.

**Example of Learning Dynamics: Zeroth-Order v.s. First-Order SGD.** To empirically evaluate the approximation fidelity of the ZO eNTK, we conduct experiments using the LeNet model ( $d = 29, 624$ ) on the MNIST (LeCun, 1998) dataset. Figure 1 shows the progressive denoising of the ZO estimate because, as  $P$  increases from 1 to 125, the kernel begins to recover the distinct block-diagonal structure of the FO eNTK, indicating improved geometric alignment. The convergence is also determined by the similarity of data samples.

### 3 UNDERSTANDING ZO OPTIMIZATION FROM THE KERNEL LENS

To improve the fidelity of gradient approximation, ZO methods usually employ a multi-perturbation strategy. Formally, the update rule for multi-perturbation ZO-SGD is:

$$\theta_{t+1} = \theta_t - \frac{\eta}{P} \sum_{p=1}^P \left[ \frac{\ell(\theta_t + \mu u_{t,p}) - \ell(\theta_t - \mu u_{t,p})}{2\mu} u_{t,p} \right],$$

where  $P$  denotes the number of random perturbations sampled at each iteration. By extending the derivation procedure used for the single-perturbation eNTK, we can readily formulate the multi-perturbation projected eNTK as:

$$\mathcal{K}^t(\mathbf{x}_o, \mathbf{x}_u; U_{t,P}) = \underbrace{\nabla_\theta z(\mathbf{x}_o)^\top}_{V \times d} \underbrace{U_{t,P} U_{t,P}^\top}_{d \times d} \underbrace{\nabla_\theta z(\mathbf{x}_u)}_{d \times V}, \quad (4)$$

where  $U_{t,P} \in \mathbb{R}^{d \times P}$  represents the random projection matrix induced by the multiple perturbations. Specifically,  $U_{t,P}$  is constructed by horizontally stacking the normalized perturbation vectors  $\{u_{t,p}/\sqrt{P}\}_{p=1}^P$ . We visually demonstrate the calculation of projected eNTK  $\mathcal{K}(\cdot; U_{t,P})$  in Figure 2.

<sup>1</sup>In some literature,  $\nabla_\theta z(\mathbf{x}_u)$  is denoted as a  $V \times d$  matrix. Yet, in this paper, we choose the  $d \times V$  one due to the later inner product interpolation instead of the outer product one.

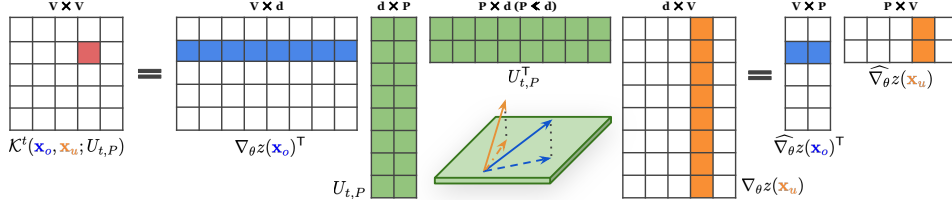


Figure 2: Calculating and Interpolating of Projected Empirical Neural Tangent Kernel.

**Difference between FO and ZO eNTK Update:** Direct analysis of the angle between the ZO gradient estimator  $\frac{\ell(\theta_t + \mu u_t) - \ell(\theta_t - \mu u_t)}{2\mu} u_t$  and the exact FO gradient  $\nabla \ell(\theta_t)$  reveals that they are nearly orthogonal in high-dimensional spaces when  $u_t$  is drawn from a Gaussian or similar isotropic distribution (Vershynin, 2018). However, practical machine learning applications prioritize model output confidence over raw parameter alignment. By characterizing the discrepancy in learning dynamics between FO and ZO methods, we derive the following relationship:

$$\Delta_{\text{fo}} \log \pi_{t+1}(y|\mathbf{x}_o) - \Delta_{\text{zo}} \log \pi_{t+1}(y|\mathbf{x}_o) \approx -\eta \mathcal{A}_t(\mathbf{x}_o) \underbrace{[\mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u) - \mathcal{K}_t(\mathbf{x}_o, \mathbf{x}_u; U_{t,P})]}_{\Delta \mathcal{K}: \text{Kernel Approximation Error}} \mathcal{G}_t(\mathbf{x}_u, \mathbf{y}_u).$$

The above quantifies the update discrepancy by mapping it directly to the deviation between the standard eNTK,  $\mathcal{K}_t(\cdot)$ , and the ZO-induced projected kernel,  $\mathcal{K}_t(\cdot; U_{t,P})$ . This formulation implies that the fidelity of the ZO update depends on how well the random projection matrix  $U_{t,P}$  preserves the geometry of the original kernel space. Unlike the raw log-probability difference, the term  $\Delta \mathcal{K}$  provides a more structured and tractable metric. Thus, we utilize this eNTK difference to analyze how the choice of sampling distribution and the number of perturbations influence ZO optimization.

We observe that the  $(i, j)$ -th entry of the kernel discrepancy matrix  $\Delta \mathcal{K}$  takes the following form:

$$\Delta \mathcal{K}[i, j] = \nabla_{\theta z_i}(\mathbf{x}_o)^\top \nabla_{\theta z_j}(\mathbf{x}_u) - \nabla_{\theta z_i}(\mathbf{x}_o)^\top U_{t,P} U_{t,P}^\top \nabla_{\theta z_j}(\mathbf{x}_u), \quad (5)$$

where  $\nabla_{\theta z_i}(\mathbf{x}_o) \in \mathbb{R}^d$  is the gradient with respect to the  $i$ -th output component (logit). This vector corresponds to the  $i$ -th column of the Jacobian matrix  $\nabla_{\theta z}(\mathbf{x}_o)$ . We define the projected gradient as  $\widehat{\nabla}_{\theta z_j}(\mathbf{x}_u) := U_{t,P}^\top \nabla_{\theta z_j}(\mathbf{x}_u) \in \mathbb{R}^P$  that represents a linear projection of the original gradient from  $\mathbb{R}^d$  into a lower-dimensional subspace  $\mathbb{R}^P$ . Thus, the kernel discrepancy matrix can be reformulated to explicitly highlight the difference in inner products between the original and projected spaces:

$$\Delta \mathcal{K}[i, j] = \langle \nabla_{\theta z_i}(\mathbf{x}_o), \nabla_{\theta z_j}(\mathbf{x}_u) \rangle - \langle \widehat{\nabla}_{\theta z_i}(\mathbf{x}_o), \widehat{\nabla}_{\theta z_j}(\mathbf{x}_u) \rangle. \quad (6)$$

The discrepancy derived above is intimately related to the Johnson-Lindenstrauss (JL) Lemma Johnson et al. (1984); Dasgupta & Gupta (2003). Although classically formulated in terms of distance preservation, the JL Lemma naturally extends to inner products via the polarization identity.

**Lemma 1 (Johnson-Lindenstrauss Lemma - Inner Product Version).** *For any set  $W$  of  $n$  points in a high-dimensional space  $\mathbb{R}^d$  and any given arbitrarily small error  $\epsilon \in (0, 1)$ , there exists a linear mapping  $f: \mathbb{R}^d \rightarrow \mathbb{R}^P$ , where the much lower dimension  $P = \mathcal{O}(\ln n / \epsilon^2)$ , such that the inner product of two points in set  $V$  remains similar after applying  $f$ :  $\forall \omega_i, \omega_j \in W$ ,*

$$|\langle f(\omega_i), f(\omega_j) \rangle - \langle \omega_i, \omega_j \rangle| \leq \frac{\epsilon}{2} (\|\omega_i\|^2 + \|\omega_j\|^2).$$

*When  $\|\omega_i\| = \|\omega_j\| = 1$ , a more concise expression is  $|\langle f(\omega_i), f(\omega_j) \rangle - \langle \omega_i, \omega_j \rangle| \leq \epsilon$ .*

A fundamental and surprising consequence of the JL Lemma is that **the sufficient projection dimension  $P$  is independent of the original feature dimension  $d$ , depending solely on the number of points  $n$  and the error tolerance  $\epsilon$** . In the context of our ZO and FO comparison, this implies that once the number of perturbations  $P$  exceeds a threshold determined by the vocabulary size  $V$ , the kernel discrepancy remains tightly bounded, regardless of how large the model dimension  $d$  becomes. Consequently, a large enough  $P$  ensures that the projected kernel  $\mathcal{K}_t(\cdot; U_{t,P})$  converges to the true eNTK, thereby aligning the ZO learning trajectory with the ideal FO dynamics and theoretically validating that multi-perturbation strategies minimize the variance of the model's confidence updates.

## ACKNOWLEDGEMENT

This work is supported in part by RIT CHAI Faculty Seed Grant, NIH award R16GM159671 and NSF grant CNS-2112471. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

## REFERENCES

- Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: long papers)*, pp. 7319–7328, 2021.
- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Raghu Bollapragada, Cem Karamanli, and Stefan M Wild. Derivative-free optimization via adaptive sampling strategies. *arXiv preprint arXiv:2404.11893*, 2024.
- Aochuan Chen, Yimeng Zhang, Jinghan Jia, James Diefenderfer, Konstantinos Parasyris, Jiancheng Liu, Yihua Zhang, Zheng Zhang, Bhavya Kaikhura, and Sijia Liu. Deepzero: Scaling up zeroth-order optimization for deep model training. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=qBWhjsNPEY>.
- Jiahe Chen and Ziyu Ma. Vamo: Efficient large-scale nonconvex optimization via adaptive zeroth order variance reduction. *arXiv preprint arXiv:2505.13954*, 2025.
- Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.
- Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=9BiVepgmWW>.
- Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.
- Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.
- John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.
- Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N Jones, and Yong Zhou. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022.
- Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pp. 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.
- Katelyn Gao and Ozan Sener. Generalizing gaussian smoothing for random search. In *International Conference on Machine Learning*, pp. 7077–7101. PMLR, 2022.

- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJgndT4KwB>.
- Wenyang Hu, Yao Shu, Zongmin Yu, Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, See-Kiong Ng, and Bryan Kian Hsiang Low. Localized zeroth-order prompt optimization. *Advances in Neural Information Processing Systems*, 37:86309–86345, 2024.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.
- Feihu Jin and Ying Tan. Hi-zfo: Hierarchical zeroth-and first-order llm fine-tuning via importance-guided tensor selection. *arXiv preprint arXiv:2601.05501*, 2026.
- William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.
- Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=ryup8-WCW>.
- Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Achieving dimension-free communication in federated learning via zeroth-order optimization. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=omrLHFzC37>.
- Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Reconciling hessian-informed acceleration and scalar-only communication for efficient federated zeroth-order fine-tuning. *arXiv preprint arXiv:2506.02370*, 2025b.
- Dandan Liang, Jianing Zhang, Evan Chen, Zhe Li, Rui Li, and Haibo Yang. Towards straggler-resilient split federated learning: An unbalanced update approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=MOAmJKj2Uc>.
- Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.
- Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.
- Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL <https://openreview.net/forum?id=Vota6rFhBQ>.
- Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023b.
- Zhendong Mi, Qitao Tan, Xiaodong Yu, Zining Zhu, Geng Yuan, and Shaoyi Huang. Kerzoo: Kernel function informed zeroth-order optimization for accurate and accelerated llm fine-tuning. *arXiv preprint arXiv:2505.18886*, 2025.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

- Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.
- Yuezhong Peng, Yuxin Liu, Fei Wen, and Xie Chen. Muzo: Leveraging multiple queries and momentum for zeroth-order fine-tuning of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8577–8595, 2025.
- Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=cit0hg4sEz>.
- Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tPNH0oZF19>.
- Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.
- James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 2002.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.
- Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.
- Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data*, pp. 581–590. IEEE, 2020.
- Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of llms in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4475–4485, 2025.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiayang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/zhang24ad.html>.
- Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=bEqI61iBue>.

## A MORE RELATED WORKS

**Neural Tangent Kernel (NTK):** The NTK framework, originally introduced to characterize the training dynamics of over-parameterized neural networks, establishes that in the infinite-width limit, gradient descent optimization is equivalent to kernel regression with a static kernel (Jacot et al., 2018). While this lazy training regime provides strong theoretical guarantees (Arora et al., 2019), subsequent research has focused on the finite-width setting, where the kernel evolves during training, capturing the intricate feature learning process of the network (Hanin & Nica, 2020; Yang & Hu, 2021). Understanding this kernel evolution is essential for accurately describing the learning dynamics of practical deep learning models. However, the application of NTK theory has been predominantly confined to FO optimization methods. In the realm of ZO optimization, the literature has extensively analyzed perturbation distributions, such as Gaussian or uniform smoothing, strictly through the lens of estimator efficiency. Recent works have developed unified frameworks to compare these distributions based on bias, variance, and dimension dependence (Liu et al., 2020; Bollapragada et al., 2024), and have derived optimal distributions to minimize the Mean Squared Error (MSE) of gradient estimators within the parameter space (Gao & Sener, 2022). Recent attempts, such as ZOPO (Hu et al., 2024) and KerZOO (Mi et al., 2025), have begun to incorporate kernel perspectives into derivative-free optimization to improve query efficiency. Crucially, however, the intersection of ZO optimization and NTK dynamics remains unexplored.

**Zeroth-Order (ZO) Optimization:** ZO optimization has been extensively studied in the literature, primarily from a theoretical optimization perspective. Early works focused on convergence guarantees for smooth or stochastic objectives (Flaxman et al., 2005; Nesterov & Spokoiny, 2017) and practical extension in other areas (Fang et al., 2022; Liang et al., 2025) while more recent studies explored the impact of different perturbation distributions on gradient estimation accuracy and optimization efficiency. These efforts have provided valuable insights into variance reduction, estimator bias, and function-value convergence (Liu et al., 2018; Li et al., 2018; Gao & Sener, 2022; Bollapragada et al., 2024). Despite these advances, existing studies largely concentrate on optimization performance and provide limited understanding of the underlying training dynamics induced by ZO methods. In particular, how random perturbations interact with parameter updates and affect the evolution of model representations during training remains largely unexplored. Addressing this gap is crucial for a deeper theoretical and empirical understanding of ZO optimization in modern machine learning and deep learning applications. To reconcile the discrepancy between the empirical success of ZO fine-tuning and its pessimistic worst-case theoretical analysis, Malladi et al. (2023b) proposed the low-effective rank assumption. This hypothesis builds on prior observations that the Hessian spectrum of a well-trained model is dominated by a few significant eigenvalues, while the remaining eigenvalues cluster near zero (Papayan, 2020; Yao et al., 2020) - a condition distinct from being strictly low-rank. Yet, for LLMs, explicitly computing the effective rank is computationally prohibitive. Thus, it remains difficult to empirically verify whether this value is truly independent of the model dimension.

## B TRADE-OFFS, LIMITATIONS, AND FUTURE WORK

As an explanatory study, this work primarily investigates the theoretical and empirical roles of the perturbation budget  $P$  in ZO optimization. While our kernel-based analysis confirms that increasing  $P$  enhances trajectory fidelity, this improvement comes with significant engineering trade-offs that warrant further discussion.

- **Computational Trade-Offs and Parallelization.** Implementing multiple perturbations introduces a linear increase in function query costs, which can be prohibitive for large-scale models. A natural question arises: *can we parallelize these multiple perturbations akin to mini-batch processing to amortize the cost?* While conceptually similar to data parallelism, parallelizing ZO perturbations is non-trivial due to the memory constraints of GPU accelerators. Unlike standard gradient accumulation, evaluating  $P$  perturbed forward passes in parallel effectively multiplies the batch size by  $P$ . For LLMs that already operate near the limit of device memory, this naive parallelization creates a severe bottleneck, risking out-of-memory errors. Future work is needed to investigate efficient activation-free parallelization strategies or hybrid schemes that balance query fidelity with memory overhead.

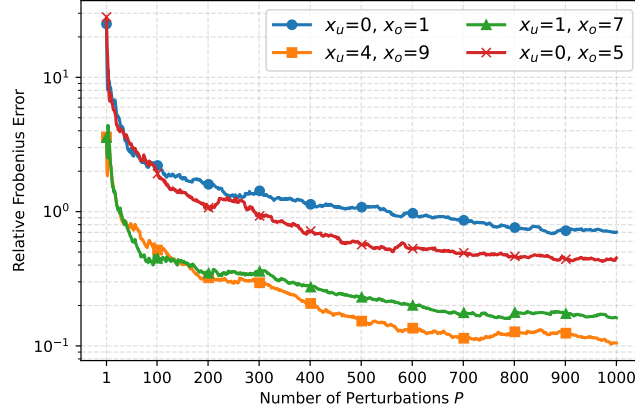


Figure 3: Convergence of Frobenius Norm Error between the ZO and FO eNTK. Pairs with high similarity: 4 and 9, 1 and 7. Pairs with low similarity: 0 and 1, 0 and 5.

- **Sequential Dependencies.** Our current eNTK framework analyzes the learning dynamics of standard prediction tasks (e.g., classification or single-step generation). Yet, it does not explicitly account for sequential reasoning tasks, such as Chain-of-Thought prompting, where the model’s output at step  $t$  conditions its input at step  $t + 1$ . In such autoregressive scenarios, the error accumulation from ZO approximations may propagate non-linearly through the reasoning chain. Extending the projected kernel perspective to capture these temporal dependencies remains a critical open problem.
- **Dynamics of Pre-training vs. Fine-tuning.** Our analysis implicitly assumes a relative stability of the eNTK, a condition that typically holds during the fine-tuning of pre-trained models where parameters remain close to their initialization, often referred to as the lazy training regime. This stability allows the random projection induced by ZO to approximate the geometry effectively. Yet, this assumption breaks down during pre-training from scratch. In the initial phases of pre-training, model features evolve rapidly, causing the underlying kernel to shift dramatically. Under such chaotic dynamics, the random subspace projection may fail to track the rapidly rotating true gradient direction. This kernel instability offers a theoretical explanation for the lack of success reports in using ZO methods to train LLMs from scratch, suggesting that ZO is uniquely suited for the stable, low-rank adaptation characteristic of fine-tuning.

## C PROOF OF LEMMA 1

First of all, recall the polarization identity that connects the inner product and distances:

$$\langle u, v \rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2). \quad (7)$$

For any linear function  $f$ , we also have

$$\langle f(u), f(v) \rangle = \frac{1}{4}(\|f(u + v)\|^2 - \|f(u - v)\|^2). \quad (8)$$

Hence, we just need to show the norms of all vectors in the set  $W' = \{u + v | u, v \in W\} \cup \{u - v | u, v \in W\}$  are all  $\epsilon$ -preserved. This set contains at most  $2n^2 + n$  elements. Leveraging the original JL lemma, we know there exists a linear mapping function  $f : \mathbb{R}^d \rightarrow \mathbb{R}^P$ , where  $P = \mathcal{O}(\ln(|W'|/\epsilon^2)) = \mathcal{O}(\ln(n)/\epsilon^2)$ , such that for any  $w \in W'$  we have

$$(1 - \epsilon)\|w\|^2 \leq \|f(w)\|^2 \leq (1 + \epsilon)\|w\|^2. \quad (9)$$

Substituting the above result into the polarization identity

$$\langle f(u), f(v) \rangle \leq \frac{1}{4} [(1 + \epsilon)\|u + v\|^2 - (1 - \epsilon)\|u - v\|^2]$$

$$\begin{aligned}
&= \frac{1}{4} \left[ (\|u+v\|^2 - \|u-v\|^2) + \frac{\epsilon}{2} (\|u+v\|^2 + \|u-v\|^2) \right] \\
&\leq \langle u, v \rangle + \frac{\epsilon}{2} (\|u\|^2 + \|v\|^2).
\end{aligned} \tag{10}$$

Similarly, we can establish the lower bound:

$$\langle f(u), f(v) \rangle \geq \langle u, v \rangle - \frac{\epsilon}{2} (\|u\|^2 + \|v\|^2). \tag{11}$$

Combining the above two results, we complete the proof of this Lemma.  $\square$

## D THEORETICAL FOUNDATION (CONTINUED)

### D.1 THE IMPACT OF THE NUMBER OF PERTURBATIONS

Besides the sampling distribution, the number of perturbations is another key hyper-parameter in the ZO method. Existing studies (Ghadimi & Lan, 2013; Liu et al., 2018; 2020; Zhang et al., 2024; Li et al., 2025a; Peng et al., 2025; Li et al., 2025b; Chen & Ma, 2025) have established that increasing the number of perturbations in ZO optimization can improve gradient approximation and accelerate convergence from an optimization-theoretic perspective. These works typically characterize the benefit of using multiple perturbations through variance reduction or improved convergence rates, and thus provide theoretical justification for employing larger perturbation budgets during training. However, a fundamental question remains largely unexplored: *how many perturbations are sufficient to achieve learning dynamics comparable to first-order optimization?* In other words, beyond asymptotic convergence guarantees, it is unclear at what point the ZO updates become a sufficiently accurate approximation of their FO counterparts in practice. In this subsection, we address this question through a kernel-based analysis. By examining how the number of perturbations affects the eNTK induced by ZO updates, we provide a principled criterion for determining when the resulting learning dynamics closely match those of FO optimization.

#### D.1.1 OPTIMIZATION POINT OF VIEW

From the optimization point of view, we just need to examine the ZO-SGD update equation 1 and the properties of the loss function  $\ell$ .

Suppose the loss function  $\ell(\theta)$  is  $L$ -smooth and the gradient associated with stochastic function value oracle is unbiased with variance bounded by  $\sigma^2$ . When the learning rate satisfies  $\eta \leq \frac{P}{Ld}$ , the ZO-SGD algorithm satisfies the following one-step descent bound:

$$\mathbb{E} \|\nabla \ell(\theta_t)\|^2 \leq \frac{\mathbb{E} \ell(\theta_t) - \mathbb{E} \ell(\theta_{t+1})}{\eta} + \frac{\eta d L}{2P} \sigma^2 + O(\eta \mu)$$

For a sufficiently large number of iterations  $T$ , setting the learning rate  $\eta = \mathcal{O}\left(\sqrt{\frac{P}{dLT}}\right)$  yields the convergence rate:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla \ell(\theta_t)\|^2 = \mathcal{O}\left(\sqrt{\frac{dL}{PT}}\right). \tag{12}$$

We highlight two key implications of this convergence rate below. First, the convergence rate improves with the number of perturbations, scaling as  $\mathcal{O}(1/\sqrt{P})$ , which aligns with what we predicted from JL lemma. Second, the convergence rate scales with the ambient model dimension  $d$ . Since  $d$  is typically vast in modern deep learning models, this dependency underpins the conventional wisdom that zeroth-order methods are significantly slower and less sample-efficient than FO methods in high-dimensional settings.

Recently, Malladi et al. (2023b) proposed the low-effective rank assumption to improve the convergence rate. This hypothesis builds on prior observations that the Hessian spectrum of a well-trained model is dominated by a few significant eigenvalues (Papayan, 2020; Yao et al., 2020). However, for LLMs, explicitly computing the effective rank is computationally prohibitive. Consequently, it

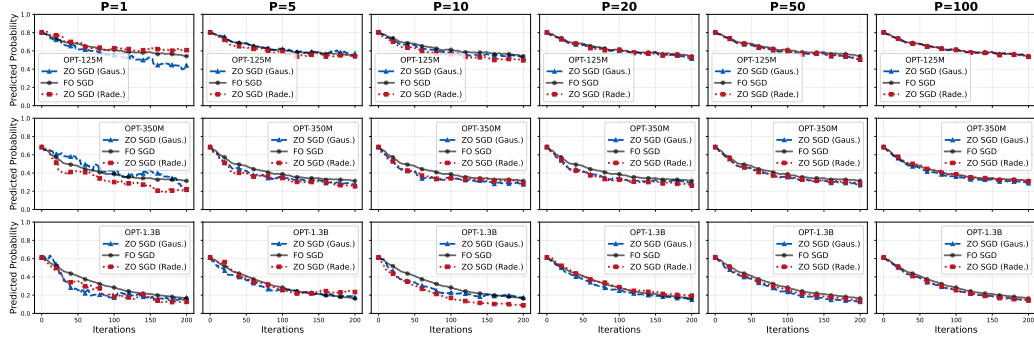


Figure 4: ZO Trajectory Comparison between OPT-125M to OPT-1.3B model on SST-2 Task over Different Perturbations.

remains difficult to empirically calculate or verify if the low effective rank is truly independent of the model dimension.

### D.1.2 ENTKE POINT OF VIEW

We proceed to establish a theoretical upper bound for the discrepancy in model belief updates between the FO and ZO methods. Starting from the operator norm inequality, we have:

$$\|\Delta_{\text{fo}} \log \pi_{t+1}(y|\mathbf{x}_o) - \Delta_{\text{zo}} \log \pi_{t+1}(y|\mathbf{x}_o)\|_2 \leq \eta \|A\|_2 \|\Delta\mathcal{K}\|_F \|\mathcal{G}_t(\mathbf{x}_o, \mathbf{y}_u)\|_2. \quad (13)$$

The critical term is the Frobenius norm of the kernel discrepancy matrix,  $\|\Delta\mathcal{K}\|_F$ . We bound this term by leveraging the inner-product preservation property of the Johnson-Lindenstrauss Lemma 1. Substituting it into the definition of the squared Frobenius norm, we obtain:

$$\begin{aligned} \|\Delta\mathcal{K}\|_F^2 &\leq \frac{\epsilon^2}{4} \sum_{i=1}^V \sum_{j=1}^V (\|\nabla_{\theta} z_i(\mathbf{x}_o)\|_2^2 + \|\nabla_{\theta} z_j(\mathbf{x}_u)\|_2^2)^2 \\ &\stackrel{(a)}{\leq} \frac{\epsilon^2}{2} \sum_{i=1}^V \sum_{j=1}^V (\|\nabla_{\theta} z_i(\mathbf{x}_o)\|_2^4 + \|\nabla_{\theta} z_j(\mathbf{x}_u)\|_2^4) \\ &\stackrel{(b)}{\leq} \frac{\epsilon^2 V}{2} (\|\nabla_{\theta} z(\mathbf{x}_o)\|_F^4 + \|\nabla_{\theta} z(\mathbf{x}_u)\|_F^4) \\ &\leq \frac{\epsilon^2 V}{2} (\|\nabla_{\theta} z(\mathbf{x}_o)\|_F^2 + \|\nabla_{\theta} z(\mathbf{x}_u)\|_F^2)^2. \end{aligned} \quad (14)$$

In step (a), we apply Jensen's inequality  $(a+b)^2 \leq 2a^2 + 2b^2$ . In step (b), we utilize the property that the sum of squares is bounded by the square of the sum for non-negative terms (i.e.,  $\sum x_i^2 \leq (\sum x_i)^2$ ), allowing us to bound the sum of fourth powers by the square of the Frobenius norm.

Denote the Frobenius norms of the Jacobian matrix as  $\Xi_t = \max(\|\nabla_{\theta} z(\mathbf{x}_o)\|_F, \|\nabla_{\theta} z(\mathbf{x}_u)\|_F)$ ; Then we have the following succinct formula  $\|\Delta\mathcal{K}\|_F \leq \epsilon \cdot \Xi_t \cdot \sqrt{V}$ . Substituting this back into Eq. equation 13 and recalling that  $\epsilon \sim \sqrt{(\log V)/P}$ , we arrive at the final convergence rate:

$$\|\text{Diff}\|_2 \lesssim \underbrace{\sqrt{\frac{V \log V}{P}}}_{\text{JL Scaling}} \cdot \underbrace{\eta \Xi_t \|\mathcal{G}_t\|_2 \|\mathcal{A}_t\|_2}_{\text{Model Dynamics}}. \quad (15)$$

This result indicates that maintaining trajectory fidelity in tasks with large vocabulary sizes  $V$  requires the number of perturbations  $P$  to scale appropriately with  $V$ . Typically, the learning rate  $\eta$  is chosen to be sufficiently small such that the term  $\eta \Xi_t \|\mathcal{G}_t\|_2$  remains bounded, thereby preventing algorithmic divergence. Furthermore, it is straightforward to show that the product  $\|\mathcal{A}_t\|_2 \leq 1$  is inherently bounded.

Thus, we arrive at a striking conclusion: the ZO method can effectively approximate the FO trajectory provided the number of perturbations is sufficiently large. Crucially, this requirement depends

on the output dimension  $V$  rather than the model dimension  $d$ . This finding offers an alternative explanation for the modest convergence degradation of ZO methods observed empirically: ZO methods can achieve a dimension-independent convergence rate even without the strict low-effective-rank assumption on the Hessian.

## D.2 SUMMARY AND VALIDATION

We conclude this section by synthesizing the insights derived from both the optimization and kernel perspectives and validating it in LLMs tasks. While distinct in their mathematical formulations, these two frameworks offer complementary rather than conflicting views on ZO dynamics.

Both perspectives agree on the fundamental role of the number of perturbations  $P$ . Whether minimizing the gradient variance (Optimization view) or bounding the kernel approximation error via the Johnson-Lindenstrauss Lemma (eNTK view), the convergence rate consistently improves with  $\mathcal{O}(1/\sqrt{P})$ . This confirms that increasing the number of perturbations is a universal strategy for enhancing the fidelity of ZO updates.

However, the crucial distinction lies in how each framework treats dimensionality. The classical optimization perspective typically yields bounds dependent on the massive model dimension  $d$ , suggesting that ZO methods are prohibitively slow for large-scale models. In contrast, our eNTK analysis reveals a more optimistic reality: the learning trajectory’s fidelity depends primarily on the output vocabulary size  $V$ . This dimension-free property of the projected kernel offers a rigorous theoretical justification for the empirical success of ZO methods in fine-tuning LLM, where  $d$  is vast but  $V$  is comparatively moderate.

**Validation on LLMs:** In Figure 4, we empirically validate this theoretical scaling law by examining the learning trajectories of OPT Zhang et al. (2022) models ranging from 125M, 350M to 1.3B parameters on the SST-2 Socher et al. (2013); Wang et al. (2018) task. We observe a consistent pattern: as  $P$  increases from 1 to 100, the ZO trajectories (for both Gaussian and Rademacher distributions) progressively align with the FO baseline. Crucially, this alignment behavior exhibits remarkable invariance to model scale. Despite the model dimension  $d$  increasing by an order of magnitude (from 125M to 1.3B), the threshold of  $P$  required to recover the FO dynamics remains virtually constant. For instance, at  $P = 50$ , ZO-SGD effectively mimics the FO trajectory across all model sizes. This empirical evidence strongly corroborates our kernel-based derivation: the fidelity of ZO learning dynamics is governed by the output dimension  $V$  (which remains constant across these models), rather than being diluted by the massive expansion of  $d$ .

## E EXPERIMENT SETUP AND MORE EMPIRICAL RESULTS

In this section, we provide more experiments setup details and empirical results to support our findings and conclusions in our main paper. To empirically validate our theoretical findings regarding the learning dynamics of ZO optimization, we conduct a series of experiments analyzing the approximation fidelity of the ZO eNTK. If the ZO eNTK can approach FO eNTK (i.g., low error, or high visual similarity), we will regard it as a good ZO approximation. We compute the exact FO eNTK as the ground truth baseline. We then compare this against the ZO eNTK estimated using stochastic perturbations. Our experiment setup are provided as follows:

### E.1 GROUP 1: LENET MODEL, MNIST DATASET.

Our experiments start with some examples on the relatively low-dimensional LeNet ( $d = 29, 624$ ) model trained on the MNIST dataset. Unless otherwise stated, we employ standard Gaussian perturbations  $u \sim \mathcal{N}(0, I)$ . We systematically vary the number of perturbations  $P$  (e.g.,  $P \in \{1, 5, 25, 125\}$ ) to observe the convergence trajectory of the ZO kernel towards the FO limit. To intuitively demonstrate the geometric alignment, we visualize the  $V \times V$  kernel matrix (where  $V = 10$  is the number of classes) as heatmaps, comparing the structural similarity between FO and ZO updates across different input pairs.

To rigorously quantify the approximation fidelity observed in the heatmaps, we track the convergence trajectory of the ZO eNTK as a function of the perturbation budget  $P$ . We utilize the relative Frobenius error to measure the distance between the ZO eNTK approximation and the ground-truth

FO eNTK. Formally, for a given pair  $(\mathbf{x}_u, \mathbf{x}_o)$ , this is defined as:

$$\text{Relative Frobenius Error} = \frac{\|\mathcal{K}_{ZO}(\mathbf{x}_u, \mathbf{x}_o; U_{t,P}) - \mathcal{K}_{FO}(\mathbf{x}_u, \mathbf{x}_o)\|_F}{\|\mathcal{K}_{FO}(\mathbf{x}_u, \mathbf{x}_o)\|_F}.$$

Unlike the discrete snapshots used for visualization, we evaluate the error across a continuous range of perturbation counts  $P$ , extending from 1 to 1,000. This allows us to verify the asymptotic behavior of the error decay. We plot individual convergence curves for each observing sample  $\mathbf{x}_o$  (classes 0 – 9) against the fixed update sample  $\mathbf{x}_u$ . This stratification allows us to decouple the impact of sample similarity on convergence speed, distinguishing between "easy" (high similarity) and "hard" (low similarity) alignment tasks.

## E.2 GROUP 2: OPT MODELS, SST2 DATASET

To demonstrate the generality of our theoretical framework beyond low-dimensional networks, we extend our analysis to high-dimensional LLMs with varying parameter scales. We employ the OPT (Zhang et al., 2022) model family, specifically evaluating OPT-125M, 350M, and 1.3B. This progression allows us to empirically verify the dimension-free property of ZO learning dynamics, as the parameter count  $d$  increases by an order of magnitude while the output vocabulary dimension  $V$  remains constant. We conduct experiments on the SST-2 sentiment classification task (Socher et al., 2013; Wang et al., 2018). Critically, instead of replacing the final layer with a task-specific classification head, we formulate the task using a prompt-based approach (i.e., next-token prediction). This design choice ensures that the model’s output dimension remains the full vocabulary size  $V$  (e.g.,  $V \approx 50,272$  for OPT), rather than being reduced to the number of classes. This setup strictly aligns with our theoretical assumption where the output dimension  $V$  is invariant to the model scaling. We compare the training trajectory of FO SGD against ZO SGD. For the ZO optimizer, we evaluate two perturbation distributions: Gaussian and Rademacher, to validate our theoretical claims regarding distribution efficiency. We vary the number of perturbations  $P \in \{1, 5, 10, 20, 50, 100\}$  to observe how the ZO trajectory asymptotically approaches the FO baseline.

## E.3 EXTRA LLM EXPERIMENT RESULTS.

From Figure 5, we observe that as  $P$  increases, the difference between the probability distributions of ZO and FO optimization consistently decreases toward zero. This indicates that with a sufficient number of perturbations, the ZO training trajectory becomes increasingly indistinguishable from standard FO optimization such as SGD. Furthermore, the trend discussed above is remarkably consistent across varying model sizes. Although the likelihood region of OPT-1.3B model has a wider spread at lower perturbation numbers, the difference and spread decrease and become the same as smaller models as the number of perturbations increases to 100.

**Sample 1:** well-nigh unendurable ... though the picture strains to become cinematic poetry , it remains depressingly prosaic and dull .

**Sentiment:** **Negative**

**Sample 2:** a big , gorgeous , sprawling swashbuckler that delivers its diversions in grand , uncomplicated fashion .

**Sentiment:** **Positive**

**Sample 3:** it ’s not that kung pow is n’t funny some of the time – it just is n’t any funnier than bad martial arts movies are all by themselves , without all oedekerk ’s impish augmentation.

**Sentiment:** **Negative**

**Sample 4:** the jobs it employs are short , carefully placed and dead-center .

**Sentiment:** **Positive**

**Sample 5:** so unremittingly awful that labeling it a dog probably constitutes cruelty to canines .

**Sentiment:** **Negative**

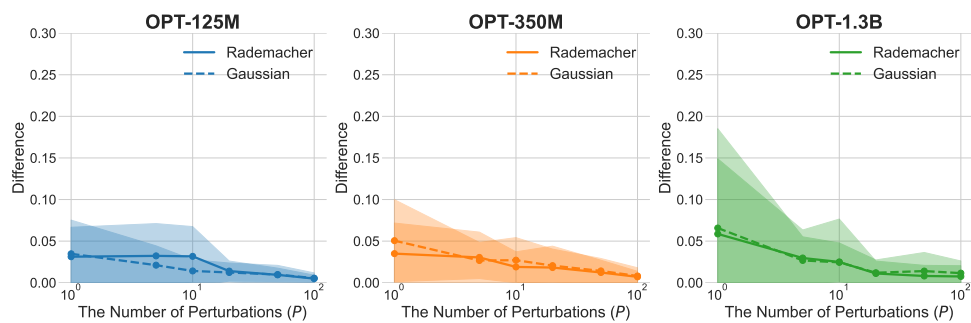


Figure 5:  $\ell_2$  Norm Difference between ZO and FO Model Belief

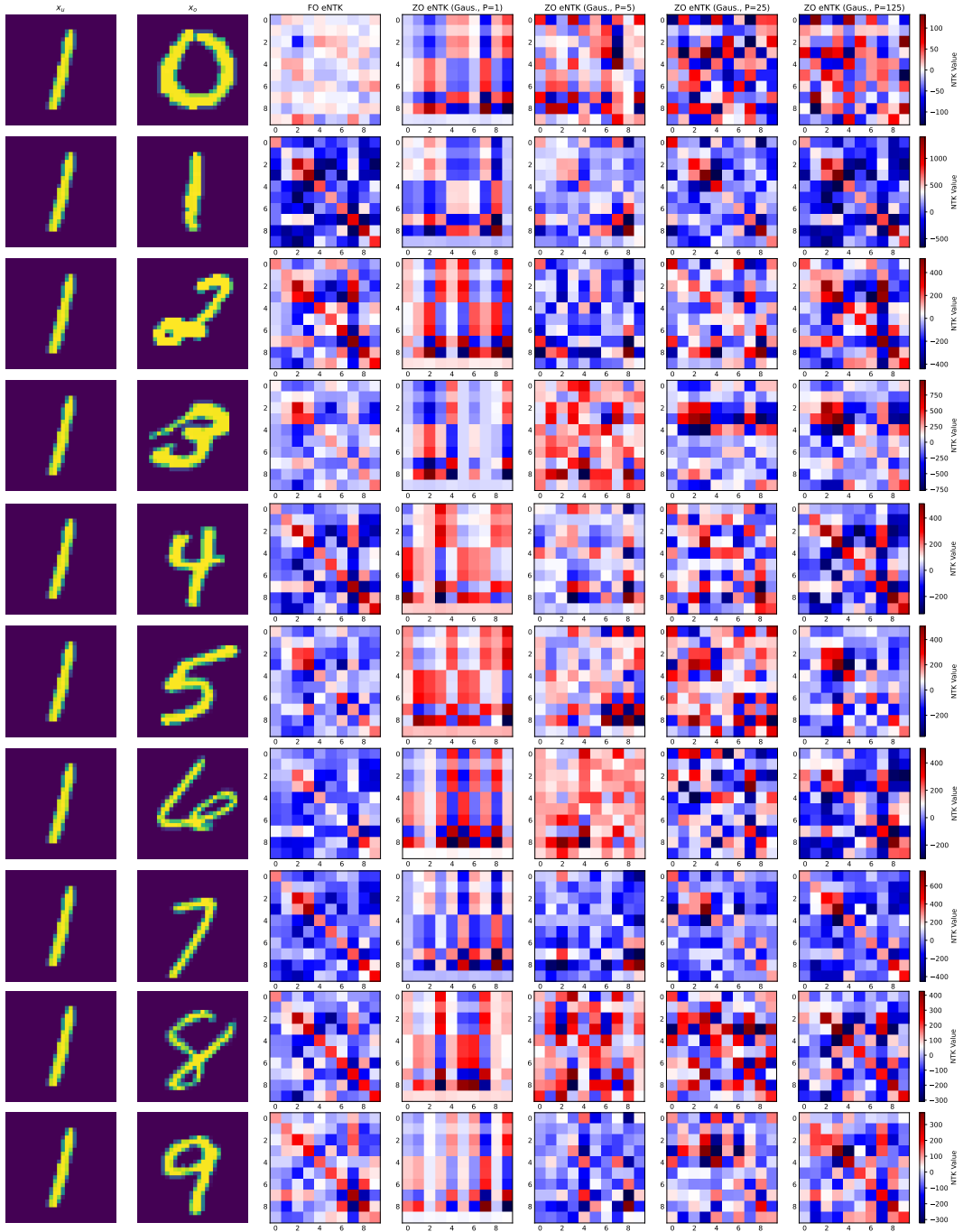


Figure 6: ZO eNTK v.s. FO eNTK under different test samples  $x_o$  and a fixed  $x_u = 1$ . (LeNet, MNIST)

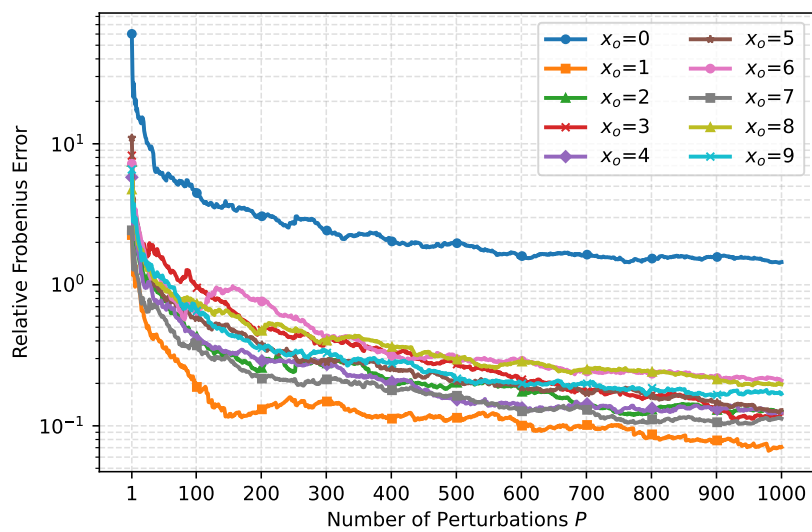


Figure 7:  $x_u = 1$  (LeNet, MNIST)

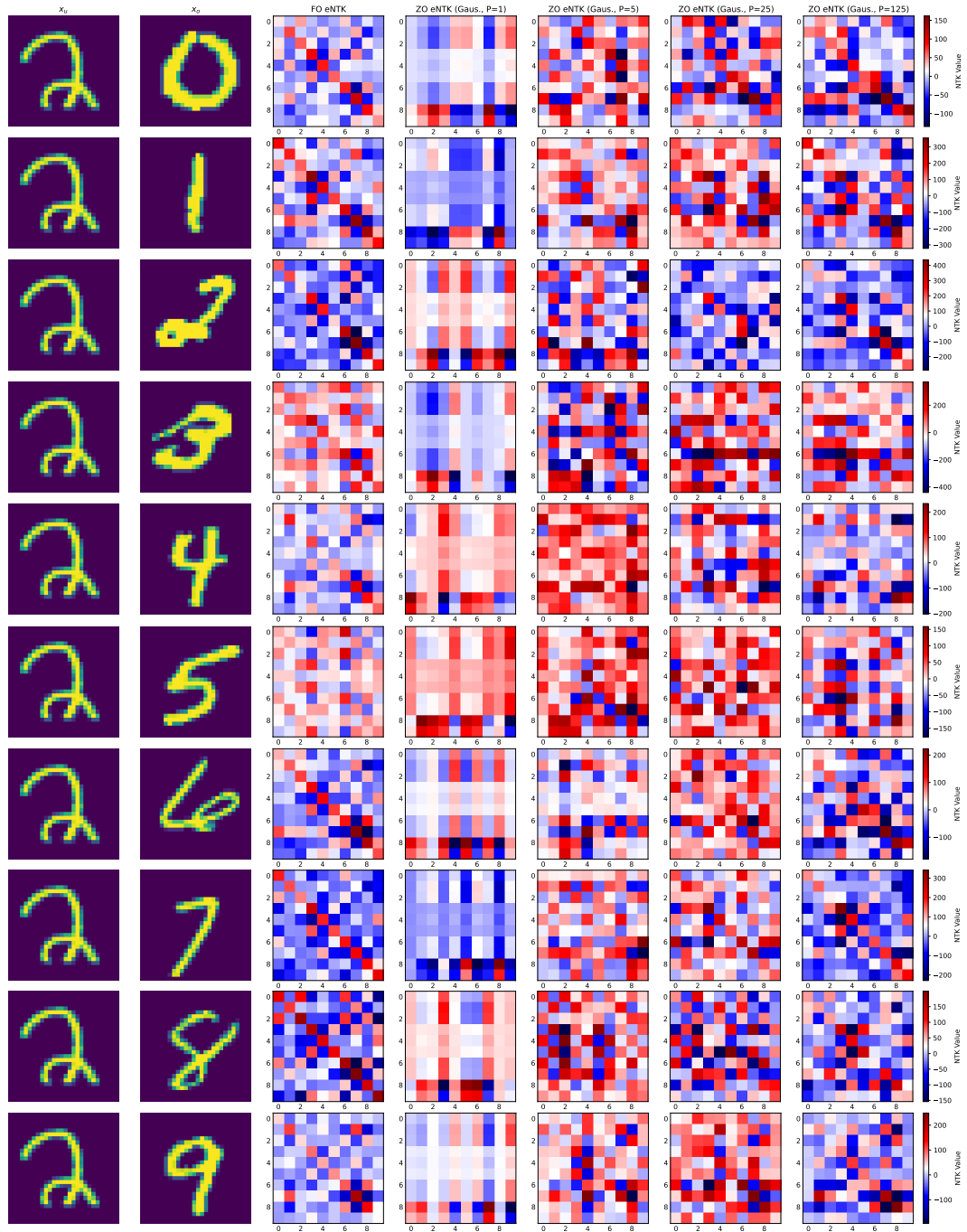


Figure 8: ZO eNTK v.s. FO eNTK under different test samples  $x_o$  and a fixed  $x_u = 2$ . (LeNet, MNIST)

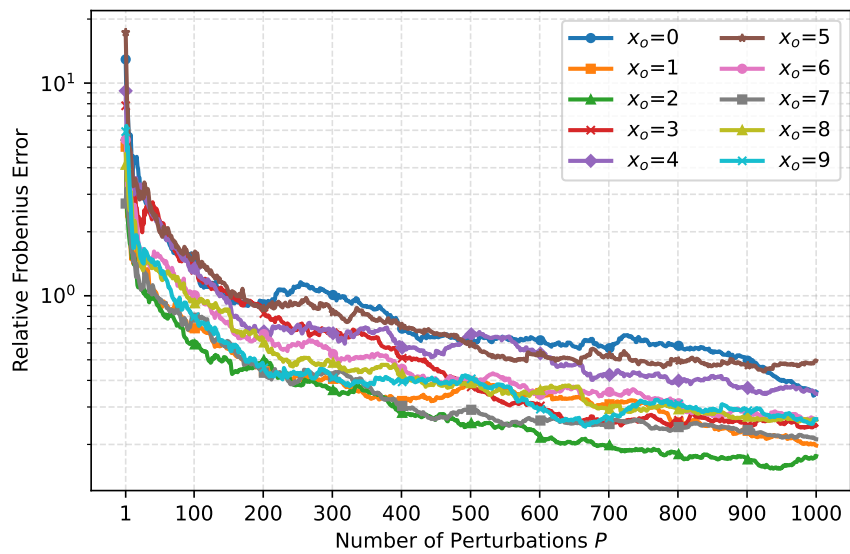


Figure 9:  $x_u = 2$  (LeNet, MNIST)

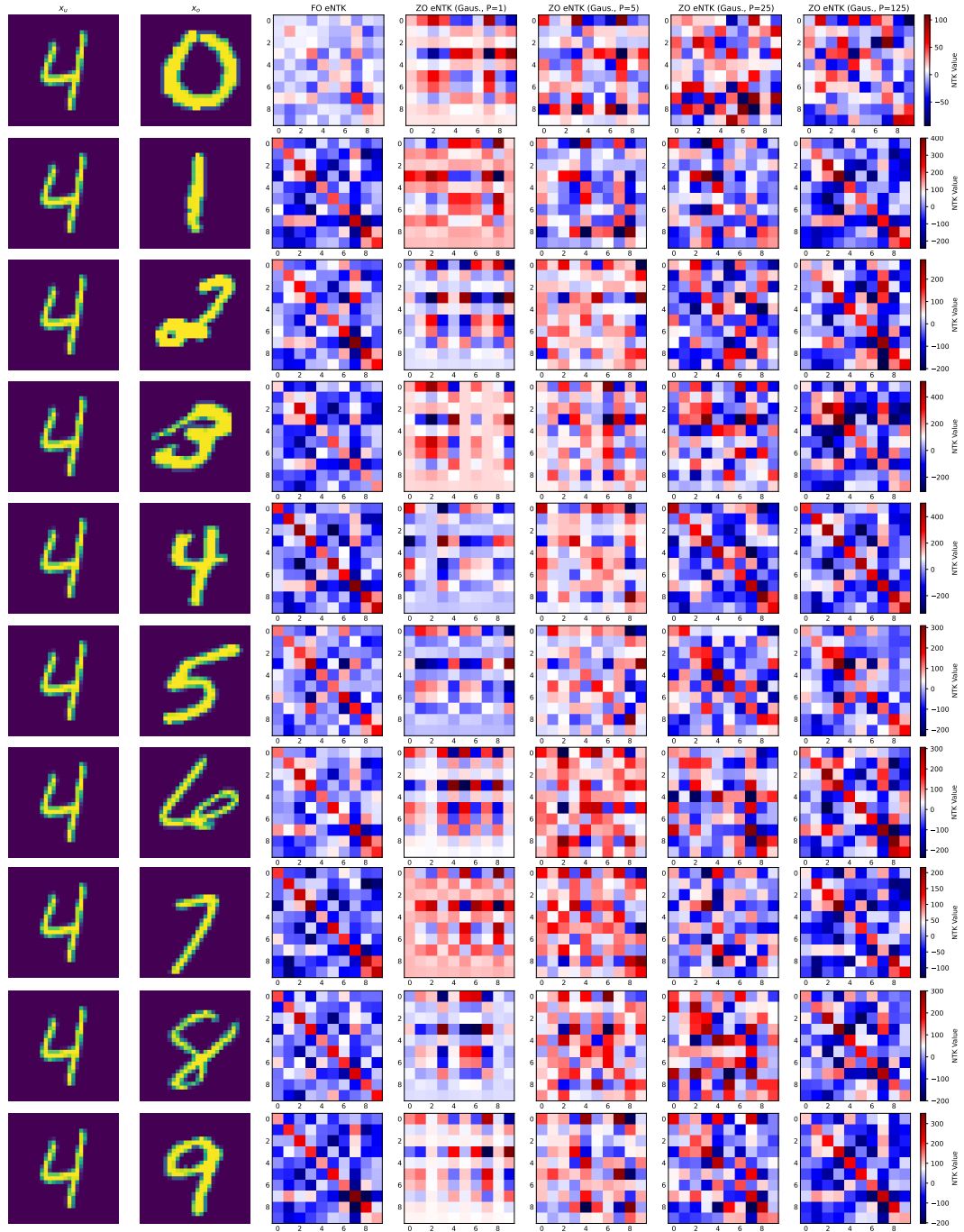


Figure 10: ZO eNTK v.s. FO eNTK under different test samples  $x_o$  and a fixed  $x_u = 4$ . (LeNet, MNIST)

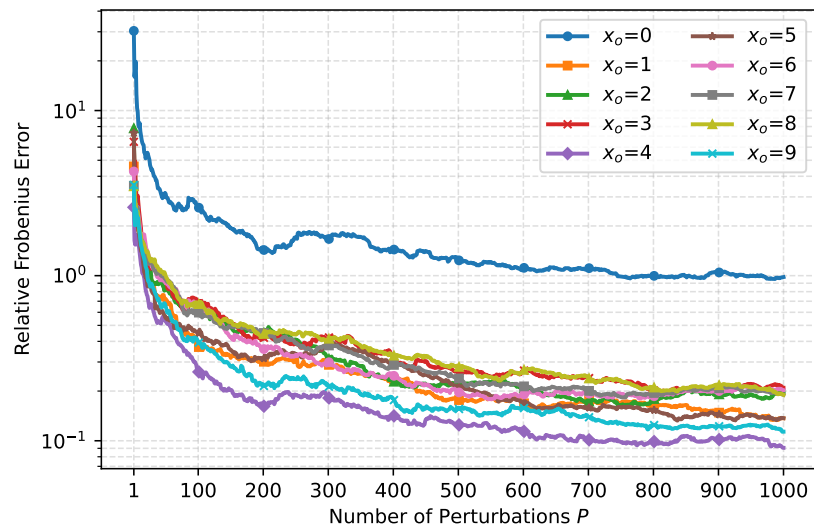


Figure 11:  $x_u = 4$  (LeNet, MNIST)