Model Evolution Under Zeroth-Order Optimization: A Neural Tangent Kernel Perspective

Chen Zhang*[1], Yuxin Cheng*[1], Chenchen Ding[1], Shuqi Wang[1], Jingreng Lei[1], Runsheng Yu[2], Yik-Chung Wu[1], Ngai Wong[1]

[1]The University of Hong Kong [2]The Hong Kong University of Science and Technology

Motivation and Problem Setting

Zeroth-order (ZO) optimization updates neural networks using only function evaluations, avoiding backpropagation memory costs.

Challenge: random directional finite differences make training dynamics hard to characterize.

Key idea: introduce Neural Zeroth-order Kernel (NZK) and interpret ZO as kernel gradient descent in function space, parallel to NTK analysis for first-order methods.

Takeaway: with shared random vectors, ZO can converge much faster while retaining memory efficiency.

Method: ZO Dynamics via NZK

For model $f(\mathbf{x};\theta)$ and loss $\mathcal{L}$, ZO update is

$$\theta_{t+1} = \theta_t - \eta\,\mathcal{G}_t,$$

$$\mathcal{G}_t = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathcal{L}(f(\mathbf{x}_i;\theta_t+\epsilon\mathbf{z})) - \mathcal{L}(f(\mathbf{x}_i;\theta_t-\epsilon\mathbf{z}))}{2\epsilon}\,\mathbf{z}.$$

The expected NZK is built from ZO tangent random features

$$\bar{\mathcal{K}}(\mathbf{x}_i,\mathbf{x}_j) = \left\langle \frac{\widehat{\partial f(\mathbf{x}_i;\theta)}}{\partial\theta}, \frac{\widehat{\partial f(\mathbf{x}_j;\theta)}}{\partial\theta} \right\rangle,$$

which governs evolution under squared loss.

Interpretation: ZO is characterized in function space by NZK, analogous to NTK analysis for FO methods.

Theoretical Results

- For linear models, the expected NZK is time-invariant and has a closed-form dependence on perturbation moments.
- For linearized neural networks, expected NZK remains fixed and yields explicit function-space dynamics.

Under squared loss,

$$\left[f^{\mathrm{lin}}_{\theta_t}(\mathbf{x}_i)\right]_N = \left(I - (I-\eta\bar{\mathcal{K}})^t\right)\left[f^*(\mathbf{x}_i)\right]_N + (I-\eta\bar{\mathcal{K}})^t\left[f_{\theta_0}(\mathbf{x}_i)\right]_N.$$

Implication: convergence speed is controlled by the NZK spectrum, enabling algorithmic acceleration through sampling design.

Practical Takeaways and Scope

- **From the paper:** NZK gives a function-space lens for ZO, analogous to NTK for first-order training.
- **From the paper:** for linear models, expected NZK is time-invariant and determined by perturbation moments.
- **From the paper:** linear/linearized squared-loss dynamics admit closed-form evolution trajectories.
- **From the paper:** reusing one shared random vector can substantially accelerate convergence.
- In practice, start with shared vectors for speed, then gradually increase randomness for robustness.
- The same framework suggests a data-dependent tuning rule: stronger sharing early, weaker sharing near convergence.

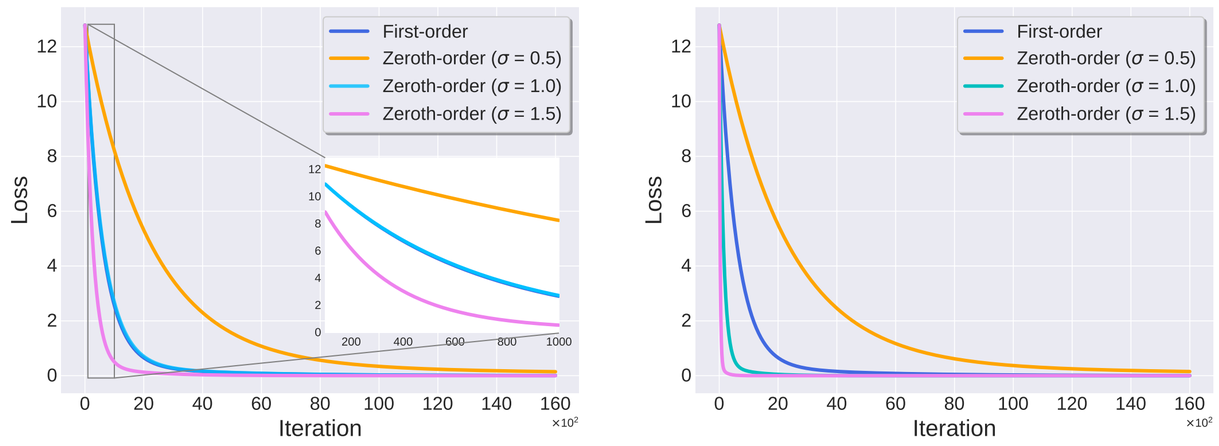
Additional Evidence from the Paper

- Under zero-mean, unit-variance perturbations, expected NZK reduces to the NTK form in function space.
- Experiments include synthetic tasks and real datasets (MNIST, CIFAR-10, Tiny ImageNet).
- Across these settings, using a shared random vector consistently accelerates optimization.

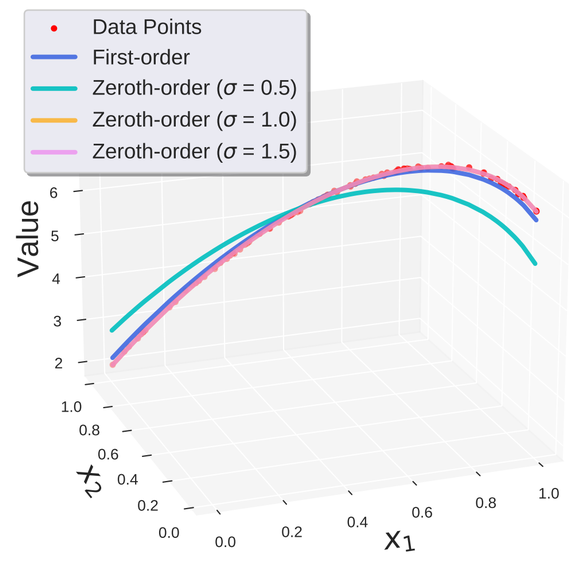
Experimental Setup Snapshot

- Benchmarks cover synthetic linear tasks and real datasets: MNIST, CIFAR-10, and Tiny ImageNet.
- Perturbation patterns compare independent random vectors against shared random vectors.
- Metrics include optimization loss, final function fit, and kernel-structure visualization (NTK vs NZK).
- Observed trends align with the NZK-based theoretical analysis in the paper.

Experiments I: Linear Models

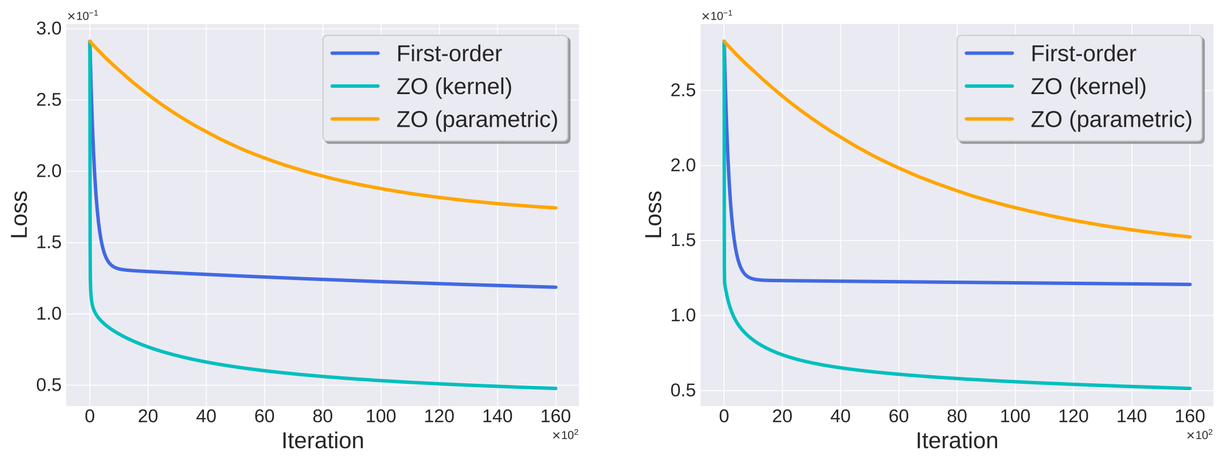


- Independent vectors (left) show that with $\sigma_z = 1$, both FO and ZO exhibit similar convergence rates. Besides, we find that for ZO, the evolution rate accelerates as σ_z increases.
- Identical vectors (right), using a single random vector for zeroth-order optimization and estimating the rate of change of $f(\boldsymbol{x};\theta)$ w.r.t. θ, consistently accelerate optimization.
- The final trained models (Independent vector) under FO and ZOs with varying variances after 16,000 iterations is presented below.

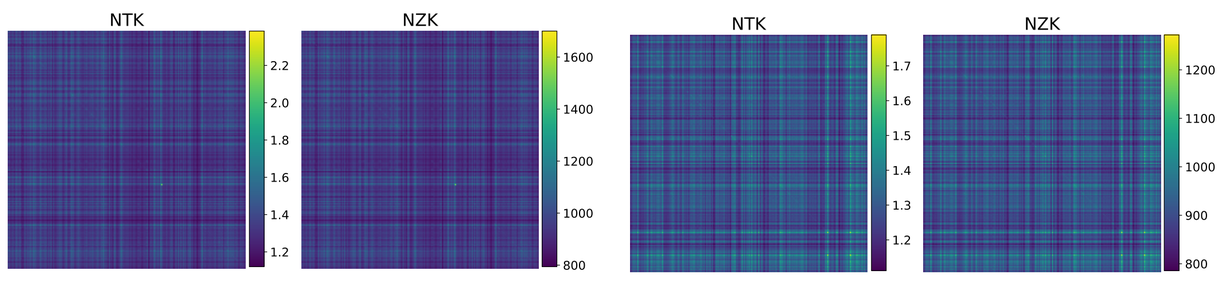


Experiments II: Linearized Neural Networks

Loss on CIFAR-10 and Tiny ImageNet



Kernel structure (NTK vs NZK)



Conclusion and Practical Guidance

- NZK provides a principled function-space view of ZO dynamics.
- Expected NZK invariance enables closed-form trajectories in linear/linearized regimes.
- Sharing random vectors is a simple and effective way to speed convergence.
- This perspective suggests new ZO designs via kernel shaping and sampling control.
- Empirically, matching perturbation structure to data geometry consistently improves optimization speed.

Message: NZK analysis suggests perturbation design as a practical lever to improve ZO optimization.