Towards Identifying Dominant Low-Rank Subspaces in Zeroth-Order Fine-Tuning

Jinjie Fang[1] **Chengxun Jin**[1] **Yi Chang**[1,2,3†] **Bin Gu**[1†]

[1]School of Artificial Intelligence, Jilin University, China
[2]International Center of Future Science, Jilin University, China
[3]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
{fangjj24,jincx1521}@mails.jlu.edu.cn {yichang, gubin}@jlu.edu.cn

Abstract

Adapting Large Language Models (LLMs) to downstream tasks is increasingly bottlenecked by the staggering memory overhead of first-order (FO) backpropagation. Zeroth-order (ZO) optimization has emerged as a compelling, memory-efficient alternative; however, it suffers from the curse of dimensionality, which introduces prohibitively high variance in gradient estimation. While LLM gradients empirically reside in low-rank subspaces, existing state-of-the-art ZO methods typically rely on randomly generated low-rank subspaces, which often fail to align with the underlying dominant gradient manifold, fundamentally limiting optimization efficiency. To bridge this gap, we propose AHZO, an efficient low-rank **ZO** fine-tuning algorithm that dynamically identifies dominant subspaces via the **Average Historical** Gradient (AHG). Motivated by the directional coherence of optimization trajectories, AHZO leverages the average historical ZO gradient over a period as a principled proxy for the true gradient. By applying singular value decomposition to AHG matrices, AHZO distills principal spectral signatures to construct low-rank bases that tightly align with the true dominant subspaces. Theoretically, we prove that AHZO significantly mitigates estimation variance, yielding superior convergence guarantees compared to random-subspace ZO methods. Extensive evaluations across diverse LLM architectures and benchmarks demonstrate that AHZO consistently outperforms existing ZO baselines, achieving performance highly competitive with memory-intensive FO fine-tuning.

1 Introduction

The emergence of large language models (LLMs), such as LLaMA [30], OPT [42], and Qwen [2], has substantially advanced the modeling of complex linguistic patterns and world knowledge from large-scale corpora. Despite their strong generalization ability, these models frequently exhibit a marked decline in efficacy when deployed in specialized domains that require domain-specific knowledge [26, 9, 1, 37, 32]. As a result, fine-tuning has become a critical component for adapting pre-trained LLMs to downstream tasks, effectively bridging the gap between generic pre-training and task-specific performance [40, 16, 19].

However, as model scales continue to expand, fine-tuning with first-order (FO) optimizers such as Adam [18] faces severe hardware limitations, particularly regarding memory overhead. The primary bottleneck stems from the necessity of retaining intermediate activations for backpropagation [20]. Parameter-efficient fine-tuning (PEFT) methods, exemplified by LoRA [14], reduce the count of trainable parameters but fail to fundamentally circumvent the activation bottleneck, as gradient-based

†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

updates still necessitate extensive backpropagation through the model's architecture [33, 36, 15]. Empirical studies suggest that full fine-tuning and PEFT can require up to several times the memory footprint of inference in long-context scenarios due to soaring activation overhead [22]. Such prohibitive overhead renders gradient-based adaptation increasingly infeasible in resource-constrained environments [45, 28, 5].

To mitigate memory overhead, MeZO [22] leverages zeroth-order (ZO) optimization [23] to achieve backpropagation-free fine-tuning. By estimating gradients through forward-pass loss differences, this approach obviates activation storage and matches the memory footprint of inference [43, 3]. Beyond basic ZO optimization, MeZO-SVRG [12] improves scalability through a variance-reduced gradient estimator; HiZOO [45] and [44] utilize diagonal Hessian information to ensure more stable convergence; LMAO [41] employs an alternating optimization scheme to achieve full-rank fine-tuning while maintaining parameter efficiency. Despite these enhancements, ZO optimization remains fundamentally constrained by the curse of dimensionality, as its estimation variance scales linearly with the parameter dimension [13, 6]. Consequently, the optimization process suffers from severe instability and a performance disparity compared to FO fine-tuing [12, 23].

A compelling insight to surmount this dimensionality bottleneck is to leverage the empirical observation that LLM gradients are predominantly concentrated within low-rank subspaces [14, 20]. This property suggests that critical gradient information inherently resides within computationally tractable subspaces. Consequently, in the context of ZO optimization, explicitly constraining perturbations to low-rank subspaces offers a promising pathway to address the intrinsic rank mismatch inherent in standard full-rank ZO methods, thereby effectively overcoming the prohibitive variance explosion [7, 39]. However, existing low-rank ZO methods typically rely on randomly generated subspaces. For instance, LOZO [7] employs stochastic perturbations formulated as the product of random low-rank matrices, while SubZero [39] implements a decomposition-based random subspace scheme. These random and unguided subspaces often struggle to consistently capture the dominant gradient manifold in high-dimensional parameter landscapes. Consequently, this leads to inefficient optimization and suboptimal convergence.

To resolve this decoupling between random subspaces and dominant gradient manifold, we propose AHZO, an efficient low-rank **ZO** fine-tuning algorithm designed to dynamically identify and align with dominant subspaces via **A**verage **H**istorical Gradient (AHG). Motivated by the inherent directional coherence of optimization trajectories, AHZO leverages the average historical estimated gradient as a principled proxy for the intractable true gradient manifold (Figure 1). Specifically, we aggregate these historical displacements into structured AHG matrices within optimization intervals, effectively capturing the span of the update directions that have driven the most substantial objective minimization. By reusing such historically informative directions, AHG avoids costly repeated queries while providing a stable and low-variance proxy for the true gradient. Subsequently, by applying Singular Value Decomposition (SVD) to these AHG matrices, AHZO distills the principal spectral signatures to construct low-rank bases. This enables the construction of low-rank bases that align well with the underlying dominant manifold of the true gradient, thereby circumventing the inefficiencies inherent in unguided random subspaces. Our main contributions are summarized as follows:

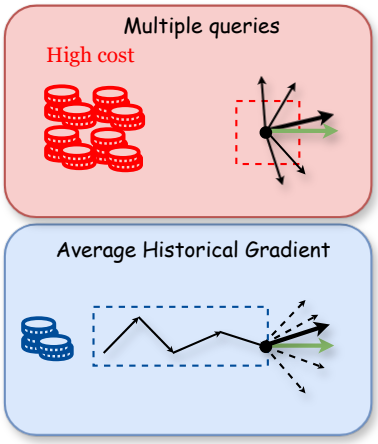

Figure 1: AHG as a proxy

- We propose AHZO, an efficient low-rank ZO optimization framework designed to overcome the limitations of unguided random subspaces. Leveraging AHG to ensure structural consistency and SVD to extract dominant basis vectors, our method dynamically identifies and aligns the gradient estimation space with the true dominant gradient manifold.

- We provide a comprehensive theoretical foundation for the AHZO algorithm, and formally demonstrate that AHZO achieves superior convergence performance and reduced estimation variance compared to low-rank ZO methods that rely on random subspaces.

- We evaluate AHZO across a diverse range of LLMs (including OPT and Qwen) on benchmarks (spanning classification, multiple-choice, and generative). Our results indicate that AHZO consistently outperforms state-of-the-art ZO baselines, substantially narrowing the performance gap between ZO and FO fine-tuning.

2 Related Work

Memory-efficient Fine-tuning. To mitigate the memory challenges of fine-tuning LLMs, recent research has focused on memory-efficient techniques that reduce training costs while preserving competitive performance. Parameter-efficient fine-tuning methods, such as LoRA [14], achieve this by introducing low-rank updates to pre-trained weights, thereby drastically reducing the number of trainable parameters without sacrificing performance [33, 20]. In contrast to FO methods, ZO optimization methods, represented by MeZO [22], have gained increasing attention as a viable alternative to conventional FO methods. By eliminating the need to store activations, ZO methods offer a more memory-efficient alternative, making them particularly appealing for large-scale models and resource-constrained settings [13, 6, 3]. Building upon MeZO, Sparse-MeZO [21] introduces a masking mechanism to sparsify parameter updates, improving efficiency while maintaining competitive performance. MeZO-SVRG [12] proposes a variance-reduced ZO method that enables efficient and scalable fine-tuning. HiZOO [45] improves ZO-based fine-tuning by incorporating diagonal Hessian information, addressing convergence instability arising from the highly non-uniform curvature of the loss landscape. LMAO [41] introduces an alternating optimization framework that integrates LoRA with ZO updates to achieve memory-efficient and full-rank fine-tuning. ConMeZO [5] enhances MeZO by introducing momentum-guided directional sampling, improving convergence speed while preserving the memory efficiency.

Low-rank ZO Fine-tuing. Recent studies have revealed that FO gradients arising during the fine-tuning of LLMs exhibit a pronounced low-rank structure, where the essential information is concentrated in a low-dimensional subspace [14, 8, 33]. This observation has motivated a line of work exploiting low-rank structures to improve the efficiency of LLM fine-tuning. In contrast, standard ZO fine-tuning methods estimate gradients using full-rank perturbations sampled from Gaussian distributions [22, 12]. This results in a mismatch between the intrinsic low-rank structure of FO gradients and the full-rank nature of ZO gradient estimators. Such a discrepancy may lead to suboptimal optimization behavior and has been identified as a potential factor contributing to the performance gap between ZO and FO methods [7]. To address this issue, recent works have proposed low-rank ZO optimization algorithms that restrict gradient estimation to low-dimensional subspaces, thereby better aligning ZO updates with the inherent low-rank structure of LLM gradients and improving optimization efficiency. For instance, LOZO [7] restricts gradient estimation to a low-rank subspace, enabling ZO updates to better align with the intrinsic low-rank structure of gradients during LLM fine-tuning. SubZero [39] further introduces a randomized subspace ZO framework that applies layer-wise low-rank perturbations, enabling efficient fine-tuning of LLMs. RoZO [27] proposes a Riemannian ZO optimization framework that improves performance by constraining updates to LoRA manifold, while relying on pre-trained FO-LoRA models for initialization.

3 Preliminaries

3.1 Memory-efficient ZO-SGD

We formulate the fine-tuning of LLMs as the following optimization problem:

$$\min_{\mathbf{W}} \mathcal{F}(\mathbf{W}) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[\mathcal{F}(\mathbf{W};\xi)\right], \tag{1}$$

where $\mathbf{W}$ denotes the model parameters, ξ is the random data sample and $\mathcal{D}$ is the data distribution, $\mathcal{F}$ is the objective function. For LLM fine-tuning, let $\ell = 1, \ldots, \mathcal{L}$ denote the layer index, and let the parameter matrices be $\mathbf{W} = \{W_\ell\}_{\ell=1}^{\mathcal{L}}$, where $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$. These weight matrices are typically high-dimensional, resulting in substantial memory overhead for FO methods, which must store intermediate activations for backpropagation [38].

To address this challenge, ZO optimization has emerged as an effective alternative that avoids explicit gradient computation. Specifically, MeZO [22] extends ZO stochastic gradient descent (ZO-SGD) to

LLM fine-tuning by employing Simultaneous Perturbation Stochastic Approximation (SPSA) [13, 34] to estimate gradients using only function evaluations:

$$\widehat{\nabla}\mathcal{F}(\mathbf{W};\xi) := \frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{Z};\xi)-\mathcal{F}(\mathbf{W}-\mu\mathbf{Z};\xi)}{2\mu}\cdot\mathbf{Z}, \tag{2}$$

where $\mathbf{Z}\sim\mathcal{N}(0,\mathbf{I}^{m\times n})$ is the full-rank perturbation matrix and μ is the perturbation scale.

We denote η as learning rate, the ZO-SGD proceeds via:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\cdot\widehat{\nabla}\mathcal{F}(\mathbf{W}_t;\xi), \tag{3}$$

A primary bottleneck in conventional ZO-SGD implementations is the substantial memory footprint necessitated by the explicit materialization of the perturbation matrix $\mathbf{Z}\in\mathbb{R}^{m\times n}$. To bypass this constraint, MeZO circumvents the need to store $\mathbf{Z}$ by retaining only the corresponding random seed.

3.2 Low-rank ZO Fine-tuing with Random Subspaces

A fundamental limitation of standard ZO optimization methods lies in the high variance of their gradient estimates, which deteriorates with increasing parameter dimensionality [3, 34, 6]. In LLM fine-tuning, this issue is substantially exacerbated by the enormous parameter scale, making the estimated descent directions noisy and unreliable. A promising solution stems from a key empirical observation: the gradient updates of LLMs are intrinsically concentrated within their inherent low-rank subspaces [14, 20]. To preserve this low-rank structure during gradient estimation, the gradient estimator of LOZO abandons the traditional full-rank perturbation matrix $\mathbf{Z}$, substituting it with the product of two low-rank factor matrices: $\mathbf{U}=\{U_\ell\}_{\ell=1}^{\mathcal{L}}\in\{\mathbb{R}^{m_\ell\times r}\}_{\ell=1}^{\mathcal{L}}$ and $\mathbf{V}=\{V_\ell\}_{\ell=1}^{\mathcal{L}}\in\{\mathbb{R}^{n_\ell\times r}\}_{\ell=1}^{\mathcal{L}}$, where the rank $r\ll\min\{m_\ell,n_\ell\}$. The gradient estimator at ℓ-th layer is defined as:

$$\widehat{\nabla}_\ell\mathcal{F}\left(\mathbf{W},\mathbf{U},\mathbf{V};\xi\right) := \frac{\mathcal{F}\left(\left\{W_\ell+\mu U_\ell V_\ell^\top\right\}_{\ell=1}^{\mathcal{L}};\xi\right)-\mathcal{F}\left(\left\{W_\ell-\mu U_\ell V_\ell^\top\right\}_{\ell=1}^{\mathcal{L}};\xi\right)}{2\mu}\cdot U_\ell V_\ell^\top. \tag{4}$$

To facilitate a concise representation, we define this estimator across all $\mathcal{L}$ layers as follows:

$$\widehat{\nabla}\mathcal{F}\left(\mathbf{W},\mathbf{U},\mathbf{V};\xi\right) := \frac{\mathcal{F}\left(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top;\xi\right)-\mathcal{F}\left(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top;\xi\right)}{2\mu}\cdot\mathbf{U}\mathbf{V}^\top. \tag{5}$$

4 Methods

4.1 The Limitations of LOZO

Optimal Gradient Subspace. We first define the optimal gradient subspace based on the empirical low-rank nature of gradients in LLMs. To extract the principal features from the true gradient, we perform SVD on the FO gradient:

$$\mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top = \mathrm{SVD}\left(\nabla\mathcal{F}(\mathbf{W})\right). \tag{6}$$

Subsequently, we define the optimal gradient subspace $\mathcal{G}^\star$ as the subspace spanned by the top-r left singular vectors:

$$\mathcal{G}^\star \triangleq \mathrm{span}(\mathbf{A}_{1:r}). \tag{7}$$

This subspace represents the "active manifold" where the most significant parameter updates occur. Specifically, by corresponding to the top-r singular values, $\mathcal{G}^\star$ captures the overwhelming majority of the dominant spectral directions while systematically filtering out the highly oscillating, non-informative directions. In the context of ZO estimation, aligning perturbation directions with this dominant subspace effectively decouples the estimation variance from the massive full-parameter dimension, thereby stabilizing the optimization trajectory.

Subspace Misalignment. For random-subspace ZO methods, the basis matrix $\hat{\mathbf{U}}$ is typically randomly sampled. Consequently, the corresponding gradient subspace is defined as follows:

$$\hat{\mathcal{G}} \triangleq \mathrm{span}(\hat{\mathbf{U}}) \tag{8}$$

In high-dimensional settings, a random $\hat{\mathcal{G}}$ is statistically likely to be near-orthogonal to the optimal manifold $\mathcal{G}^\star$ (as stochastic searching for an r-dimensional optimal subspace within an m-dimensional ambient space). This persistent subspace misalignment prevents the estimator from consistently capturing the most effective update directions throughout the optimization trajectory.

Subsequently, we evaluate how the efficiency of gradient signal preservation is determined by geometric alignment. Specifically, this efficiency is proportional to the cosine similarity between the stochastic subspace $\hat{\mathcal{G}}$ and the optimal gradient subspace $\mathcal{G}^\star$, as formalized below:

$$\frac{\|\mathcal{P}_{\hat{\mathcal{G}}}(\nabla\mathcal{F}(\mathbf{W}))\|_F}{\|\mathcal{P}_{\mathcal{G}^\star}(\nabla\mathcal{F}(\mathbf{W}))\|_F} \propto \cos\left(\hat{\mathcal{G}}, \mathcal{G}^\star\right). \tag{9}$$

Inefficient Optimization in Random Subspaces. To characterize the optimization bottleneck, we assume the objective function $\mathcal{F}$ is L-smooth. Under the subspace-constrained ZO setting, the gradient estimator $\widehat{\nabla}\mathcal{F}$ is constructed by perturbing parameters within a subspace $\mathcal{G}$. The expected objective decrease per iteration is governed by the following descent lemma:

$$\begin{aligned}
&\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \\
&\leq -\eta \cdot \mathbb{E}_{\xi_t,\mathbf{V}_t}\left[\left\langle \nabla\mathcal{F}(\mathbf{W}_t), \widehat{\nabla}\mathcal{F}\left(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t\right)\right\rangle\right] + \frac{L\eta^2}{2} \cdot \mathbb{E}_{\xi_t,\mathbf{V}_t}\left[\left\|\widehat{\nabla}\mathcal{F}\left(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t\right)\right\|_F^2\right] \\
&\leq \mathcal{O}\left(\left(L\eta^2 - \eta\right)\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2\right) + \mathcal{O}\left(L^3\eta^2\mu^2 + \eta\mu\right).
\end{aligned} \tag{10}$$

The detailed proof is provided in the Appendix (Lemma 7). The efficiency of this descent is primarily determined by the first term (η is typically set to a small value), which represents the gradient signal projection onto the sampling subspace $\mathcal{G}$; The second term is governed by the stochastic ZO gradient noise, which impairs the descent of the objective function. If $\mathcal{G}$ is sampled uniformly at random ($\hat{\mathcal{G}}$), the expected projection norm $\|\mathcal{P}_{\hat{\mathcal{G}}}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2$ scales as $\mathcal{O}(r/m)$, where r is the subspace rank. Since $r \ll m$, the cosine similarity $\cos(\hat{\mathcal{G}}, \mathcal{G}^\star)$ vanishes almost surely. Consequently, the random subspace captures a negligible fraction of the gradient information, leading to inefficient optimization formulated in Eq.(10), where the descent signal is overwhelmed by noise.

4.2 Align with Dominant Subspaces via Average Historical Gradient

AHG for Principled Proxy. In the ZO optimization regime, the FO gradient is inherently inaccessible. A standard strategy to approximate the FO gradient and reduce the high variance of ZO estimators is to average multiple independently sampled perturbation directions at the current step. However, as illustrated in the top panel of Figure 1, this strategy requires multiple function queries per step, incurring a prohibitively high cost in the LLM setting. To obtain a more efficient proxy without inflating the budget, we leverage the trajectory of the optimization process. Rather than relying on multiple noisy queries at a single point, we utilize the average historical ZO gradient information from the previous period, acting as an AHG. Specifically, for the k-th period $t \in [k\nu, \ldots, (k+1)\nu - 1]$, we define the single-step empirical displacement δ_t, induced by the ZO update at step t. These displacements along the optimization path are averaged to construct the AHG matrix:

$$\Delta = \sum_{t=k\nu}^{(k+1)\nu-1} \delta_t. \tag{11}$$

Compared to single-step multi-query estimation, AHG implicitly averages historical stochastic perturbations over the entire period. This trajectory-based approach significantly reduces variance and aligns with the true gradient without introducing any additional query overhead.

SVD for dominant basis. To periodically update the ZO gradient subspace, we decompose the accumulated AHG matrix to extract the dominant low-rank subspace, which captures the directions most critical for loss reduction, using truncated SVD:

$$\mathbf{A}_z\mathbf{\Sigma}_z\mathbf{B}_z^\top = \mathrm{TruncatedSVD}\left(\Delta, q\right) \tag{12}$$

Subsequently, we utilize the first p singular vectors in $\mathbf{A}_{z,1:p}$, i.e., $\mathbf{A}_z[:, 1:p]$, as $\mathbf{U}_{\text{exploit}}$ to characterize the dominant descent manifold, which is employed for the exploitation phase.

To ensure the algorithm retains the capacity to explore novel directions, we designate the orthogonal complement of $\mathbf{U}_{\text{exploit}}$ as the exploration phase:

$$\mathbf{U}_{\text{explore}}, _ \leftarrow \text{QR}(\Omega - \mathbf{U}_{\text{exploit}} \cdot \mathbf{U}_{\text{exploit}}^{\top} \cdot \Omega). \tag{13}$$

Subsequently, we synthesize the components from both the exploitation and exploration phases to construct the updated subspace matrix:

$$\mathbf{U} = [\mathbf{U}_{\text{exploit}}, \mathbf{U}_{\text{explore}}] \tag{14}$$

Finally, we establish the relationship between our subspace and the true gradient subspace:

$$\cos\left(\hat{\mathcal{G}}, \mathcal{G}^{\star}\right) \propto \cos\left(\mathbf{U}, \mathbf{A}_{1:r}\right) \propto \cos\left(\mathbf{U}_{\text{exploit}}, \mathbf{A}_{1:p}\right) \tag{15}$$

4.3 Algorithmic Design of AHZO

To reduce memory consumption, we store $\Psi \in \mathbb{R}^{r \times n}$ instead of $\mathbf{W} \in \mathbb{R}^{m \times n}$:

$$\Psi = (\eta/r) \cdot \Phi \cdot \mathbf{V}^{\top}. \tag{16}$$

We then calculate the single-step parameter displacement: $\delta = \mathbf{U} \cdot \Psi$. All algorithmic details are presented in Algorithm 1. Figure 2 illustrates the conceptual framework of our algorithm.

Algorithm 1 AHZO

Require: Weight $\mathbf{W}$, loss function $\mathcal{F}$, samples ξ, learning rate η, perturbation size μ, step budget T, number of layers $\mathcal{L}$, interval ν, rank r, truncated rank q, exploited rank p.

Initialize orthogonal basis matrices $\mathbf{U}$
Initialize coefficient matrices $\Psi = \{\Psi_\ell\}_{\ell=1}^{\mathcal{L}}$, where $\Psi_\ell \in \mathbb{R}^{r \times n_\ell}$ and $\Psi_\ell = 0$
for $t = 0$ to $T - 1$ **do**
    Sample $\mathbf{V}$ from a standard normal distribution
    $\mathbf{W} \leftarrow \text{LayerwisePerturbation}(\mathbf{W}, \mu, \mathbf{U}, \mathbf{V})$
    $\mathcal{F}_+ \leftarrow \mathcal{F}(\mathbf{W}; \xi)$
    $\mathbf{W} \leftarrow \text{LayerwisePerturbation}(\mathbf{W}, -2\mu, \mathbf{U}, \mathbf{V})$
    $\mathcal{F}_- \leftarrow \mathcal{F}(\mathbf{W}; \xi)$
    $\mathbf{W} \leftarrow \text{LayerwisePerturbation}(\mathbf{W}, \mu, \mathbf{U}, \mathbf{V})$
    $\Phi \leftarrow (\mathcal{F}_+ - \mathcal{F}_-)/(2\mu)$
    $\mathbf{W} \leftarrow \mathbf{W} - (\eta/r) \cdot \Phi \cdot \mathbf{U}\mathbf{V}^{\top}$; $\Psi \leftarrow \Psi + (\eta/r) \cdot \Phi \cdot \mathbf{V}^{\top}$
    if $(t+1) \bmod \nu = 0$ **then**
        $\mathbf{U} \leftarrow \text{ExploitExploreSubspace}(q, p, r, \mathbf{U}, \Psi)$; $\Psi \leftarrow \mathbf{0}$
    end if
end for

function LayerwisePerturbation($\mathbf{W}, \mu, \mathbf{U}, \mathbf{V}$)
    for $\ell = 1$ to $\mathcal{L}$ **do**
        $W_\ell \leftarrow W_\ell + \mu \cdot U_\ell V_\ell^{\top}$
    end for
end function

function ExploitExploreSubspace($q, p, r, \mathbf{U}, \Psi$)
    $\Delta \leftarrow \mathbf{U} \cdot \Psi$
    $\mathbf{A}_z, _, _ \leftarrow \text{TruncatedSVD}(\Delta, q)$
    Generate $\Omega = \{\Omega_\ell\}_{\ell=1}^{\mathcal{L}} \in \left\{\mathbb{R}^{m \times (r-p)}\right\}_{\ell=1}^{\mathcal{L}}$ from a standard normal distribution
    $\mathbf{U}_{\text{explore}}, _ \leftarrow \text{QR}(\Omega - \mathbf{A}_z(\mathbf{A}_z\Omega))$
    $\mathbf{U} \leftarrow [\mathbf{A}_z[:, 1:p],\ \mathbf{U}_{\text{explore}}]$
end function

5 Convergence Analysis

To provide a rigorous convergence analysis for our low-rank ZO optimization framework, we begin by introducing several standard assumptions for non-convex optimization [29, 4, 17, 11, 35, 24, 34]:

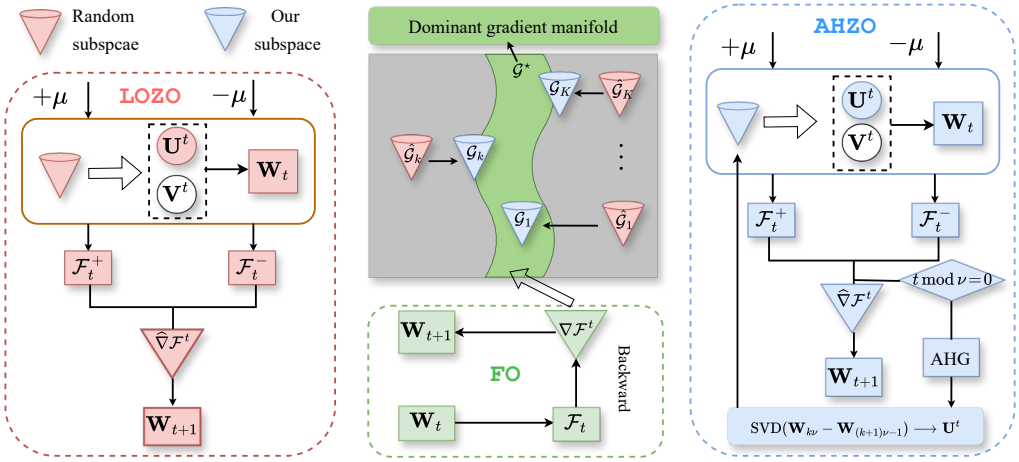

Figure 2: Comparison of the frameworks for LOZO and AHZO.

Assumption 1. ***L-Smoothness:*** *For any ξ, the function $\mathcal{F}(\mathbf{W};\xi)$ is differentiable with respect to $\mathbf{W}$. Furthermore, the gradient $\nabla\mathcal{F}(\mathbf{W};\xi)$ is uniformly L-Lipschitz continuous, i.e., $\forall\, \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{m\times n}$:*

$$\left\|\nabla\mathcal{F}(\mathbf{W}_1;\xi) - \nabla\mathcal{F}(\mathbf{W}_2;\xi)\right\|_F \leq L\left\|\mathbf{W}_1 - \mathbf{W}_2\right\|_F. \tag{17}$$

Assumption 2. ***Bounded Variance:*** *The stochastic gradient $\mathcal{F}(\mathbf{W};\xi)$ is unbiased with bounded variance:*

$$\mathbb{E}_\xi\left[\left\|\nabla\mathcal{F}(\mathbf{W};\xi) - \nabla\mathcal{F}(\mathbf{W})\right\|_F^2\right] \leq \sigma^2. \tag{18}$$

Assumption 3. ***Bounded initial optimality gap:*** *Given an initial point $\mathbf{W}_0$, and $\mathbf{W}^\star$ denotes a global minimizer of $\mathcal{F}$, there exists a constant $H < \infty$ that:*

$$\mathcal{F}(\mathbf{W}_0) - \mathcal{F}(\mathbf{W}^\star) \leq H. \tag{19}$$

Based on the aforementioned assumptions, the following theorem establishes the convergence guarantees for AHZO. We provide the complete mathematical derivation in Appendix A.

Theorem 1. *Under Assumptions 1, 2, and 3, run AHZO for $t = 0, ..., T-1$ steps with period ν, and at each step use the rank-r two-point estimator $\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V};\xi)$, where $\mathbf{U}^\top\mathbf{U} = I_r$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$ have i.i.d. $\mathcal{N}(0,1)$ entries. Let $I = \min\{m,n\}$, $diag(\{s_{t,i}\}_{i=1}^I)$ is the singular matrix by SVD decomposition for $\nabla\mathcal{F}(\mathbf{W}_t)$ and $\mathbf{A}_{t,1:p}$ is the corresponding top-p left singular vectors, we denote $\varepsilon_{t,exploit} := |\cos\Theta(\mathbf{U}_{t,exploit}, \mathbf{A}_{t,1:p})|$ is the similarity measure. Suppose the step size satisfies $0 < \eta < \frac{1}{4L(nr+2)}$, then the following inequality holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{F}(\mathbf{W}_t)\right\|_F^2 \leq \frac{8H}{3\eta T} + \sigma^2 + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \frac{5}{3}G_1(p)\cdot G_2(r-p), \tag{20}$$

where $G_1(p) = \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{i=1}^I s_{t,i}^2 - \varepsilon_{t,exploit}^2\sum_{i=1}^p s_{t,i}^2\right)$ and $G_2(r-p) = (1 - \frac{r-p}{m-q})$.

Remark 1. *Theorem 1 establishes the global convergence behavior of the AHZO algorithm. The terms $G_1(p)$ and $G_2(r-p)$ characterize the effects of exploitation and exploration on the overall convergence, respectively, and together quantify the error induced by subspace misalignment. Specifically, $G_2(r-p) = 1 - \frac{r-p}{m-q}$ corresponds to the exploration component. Since $\frac{r-p}{m-q}$ is typically very small, this factor remains close to one, indicating that exploration helps prevent the algorithm from being trapped at saddle points while introducing negligible degradation to the convergence bound. In contrast, the final convergence error is mainly governed by the exploitation term $G_1(p)$. Since the dominant gradient information is primarily captured by a low-rank structure, the quantity $\sum_{i=1}^p s_i^2$ plays a leading role in $G_1(p)$. Therefore, increasing the alignment similarity $\varepsilon_{exploit}^2$ effectively reduces the residual error caused by subspace misalignment, thereby improving the convergence performance.*

To further quantify the theoretical reliability of AHZO, we derive the following corollary for $\varepsilon_{\text{exploit}}$:

Corollary 1. ***Discussion regarding*** $\varepsilon_{exploit}$*: Under Assumptions 1 and 2, at the* k*-th period, let* $\rho = \max_{t=k\nu,\ldots,(k+1)\nu-1}\{\|\mathbf{W}_j - \mathbf{W}_{j+1}\|_F\}$*, and* $\sigma_z = \sigma_p(\Delta) - \sigma_{p+1}(\Delta)$ *be the singular value gap. The similarity between the estimated subspace and the optimal subspace is bounded as follows:*

$$\varepsilon_{exploit} = |\cos\Theta(\mathbf{U}_{exploit}, \mathbf{A}_{1:p})| \geq \sqrt{1 - \frac{L^2\eta^2(\mu\nu\sqrt{nr} + \nu^2 \cdot \rho)^2}{r^2\sigma_z^2}} \tag{21}$$

Remark 2 (**Comparison with random subspaces**)**.** *For random subspace methods, the expected alignment satisfies* $\mathbb{E}[\varepsilon_{exploit,t}] \leq p/m$*, which becomes negligible for LLMs with extremely large parameter dimension* m*. In contrast, AHZO provides a deterministic lower bound on* $\varepsilon_{exploit,t}$*. Under* $0 < \eta < \frac{1}{4L(nr+2)}$*, we have* $L^2\eta^2 < \frac{1}{16(nr+2)^2}$*, which offsets the dimensional dependence induced by the* $\sqrt{nr}$ *term in Eq. (21). When* μ*,* ν*,* ρ*, and* r *are bounded and* σ_z *is bounded away from zero, the correction term therefore diminishes with dimension. Hence, unlike random subspace selection whose expected alignment scales as* p/m*, AHZO maintains a nontrivial deterministic alignment guarantee in high-dimensional regimes.*

Remark 3 (**Empirical validation**)**.** *Tables 16 and 17 in Appendix B.7 empirically support the regime underlying Corollary 1. Table 16 shows persistent non-vanishing singular-value gaps* σ_z*, with average values of* 1.89–3.87 *on COPA/Qwen2.5-7B and* 9.89–10.50 *on SST-2/OPT-13B. Meanwhile, Table 17 shows limited within-period drift* ρ*, with average layer-wise values of* 0.08–0.19 *and* 0.18–0.35*, respectively. A non-vanishing* σ_z *controls the denominator in Eq. (21), while a small* ρ *limits its drift-dependent numerator. These observations are therefore consistent with the regime in which the correction term remains controlled and* $\varepsilon_{\mathrm{exploit},t}$ *retains a meaningful lower bound.*

6 Experiments

6.1 Experimental Setup

Datasets and Models. To rigorously evaluate optimization performance, we consider tasks from the GLUE [31], SuperGLUE [25], and related benchmarks (e.g., SST-2, RTE, CB, BoolQ, WSC, WiC, MultiRC, COPA, ReCoRD, SQuAD, and DROP), which cover a broad range of linguistic phenomena, including natural language inference, coreference resolution, commonsense reasoning, and machine reading comprehension. We conduct experiments on Qwen2.5-7B and Qwen2.5-32B [2], OPT-1.3B and OPT-13B [42].

Baselines. To comprehensively evaluate the effectiveness of AHZO, we compare it against a diverse set of baselines that represent the current state of the art in LLM fine-tuing. Specifically, we include the memory-efficient ZO baselines: MeZO [22] and MeZO-LoRA [22], HiZOO [45], SubZero [39], and LOZO [7]. We further report results for two widely used non-training baselines: Zero-shot inference (Zero-shot), in-context learning (ICL) [10], and FO baselines: fine-tuning with Adam (FT) [18] and FT-LoRA [14]. The detailed baseline configurations are provided in the Appendix B.1. Our code is available at https://github.com/fangjj2000/Low-rank-AHZO.

6.2 Main Results

We evaluate AHZO across four model scales, including OPT-1.3B, Qwen2.5-7B, OPT-13B, and Qwen2.5-32B. As shown in Tables 1–4, AHZO consistently achieves the best average performance among all ZO fine-tuning methods across the four models, demonstrating strong effectiveness and scalability across diverse tasks and model sizes. Specifically, AHZO improves the best competing ZO baseline by 3.81, 1.6, 1.6, and 2.2 average points on OPT-1.3B, Qwen2.5-7B, OPT-13B, and Qwen2.5-32B, respectively. Notably, on OPT-13B, AHZO achieves an average score of 72.5, matching full fine-tuning while avoiding its substantial memory overhead. The consistent gains from 1.3B to 32B parameters indicate that AHZO remains effective as model scale increases. Figure 3 illustrates the training loss dynamics of AHZO, LOZO, and MeZO on the SST-2, RTE, and COPA datasets using Qwen2.5-7B, the results show AHZO consistently achieves faster loss reduction and converges to lower final training loss values compared with LOZO and MeZO.

Table 1: Performance comparison on OPT-1.3B across different tasks.

Task Type	Classification							Multiple Choice		Generation		Average
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	53.44	53.43	37.50	45.60	44.23	57.52	45.40	70.00	70.00	26.99	11.12	46.84
ICL	85.21	52.71	57.14	65.40	54.81	55.96	54.40	71.00	70.00	59.30	20.55	58.77
FT	93.23	77.62	78.57	73.30	53.85	62.07	72.20	73.00	70.90	80.60	29.30	69.51
FT-LoRA	91.40	56.32	71.43	62.30	47.12	58.93	60.50	78.00	71.90	75.36	23.44	63.34
MeZO	87.84	51.99	69.64	63.10	55.77	57.37	56.30	72.00	70.60	75.66	22.76	62.09
MeZO-LoRA	90.60	54.51	64.29	63.40	52.88	58.31	56.10	74.00	**72.30**	76.97	23.91	62.48
SubZero	87.96	59.21	69.64	62.60	53.85	55.80	58.40	75.00	70.60	72.42	21.77	62.48
HiZOO	91.17	51.99	69.64	62.57	55.77	57.37	56.48	70.00	70.79	73.27	25.16	62.20
LOZO	85.89	57.04	67.86	62.60	52.88	56.58	55.30	72.00	71.10	72.42	23.04	61.52
AHZO	**92.20**	**64.62**	**71.43**	**65.50**	**58.65**	**60.19**	**63.70**	**78.00**	**72.30**	**77.39**	**25.24**	**66.29**

Table 2: Performance comparison on Qwen2.5-7B across different tasks.

Task Type	Classification							Multiple Choice		Generation		Average
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	63.6	84.5	62.5	81.9	68.3	61.0	77.8	86.0	75.2	85.1	60.8	73.3
ICL	95.3	83.4	91.1	85.4	66.3	64.4	86.7	86.0	76.3	74.8	60.7	79.1
FT-LoRA	79.6	89.5	78.6	80.4	73.1	65.8	87.7	86.0	77.0	89.8	61.9	79.0
FT (11× memory)	95.0	89.9	96.4	87.6	76.0	75.5	86.6	90.0	80.0	91.5	69.5	85.3
MeZO	89.1	88.1	89.3	84.1	69.2	68.8	87.4	89.0	76.9	87.4	60.2	80.9
MeZO-LoRA	78.9	87.4	85.7	82.9	71.2	64.9	86.2	89.0	75.3	81.6	60.7	78.5
SubZero	88.1	88.4	89.3	84.1	72.1	69.3	87.7	**90.0**	74.6	86.9	59.6	80.9
HiZOO	92.0	88.8	91.1	82.9	65.4	67.1	87.1	86.0	76.7	87.1	60.7	80.4
LOZO	90.6	86.6	**96.4**	84.3	73.1	70.2	87.9	**90.0**	76.6	87.5	60.7	82.2
AHZO	**93.8**	**89.5**	**96.4**	**84.8**	**73.1**	**73.2**	**88.3**	**93.0**	**77.5**	**89.4**	**63.0**	**83.8**

Table 3: Performance comparison on OPT-13B across different tasks.

Task Type	Classification							Multiple Choice		Generation		Average
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	58.8	59.6	46.4	59.0	38.5	55.0	46.9	80.0	81.0	46.2	14.6	53.3
ICL	87.0	62.1	57.1	66.9	39.4	50.5	53.1	87.0	82.3	75.9	29.5	62.8
FT-LoRA	92.4	62.1	60.7	66.9	40.4	58.2	58.0	84.0	81.1	81.3	29.4	65.0
FT (12× memory)	92.0	70.8	83.9	77.1	63.5	70.1	71.1	79.0	74.1	84.9	31.3	72.5
MeZO	91.3	68.2	66.1	68.1	61.5	59.4	59.4	88.0	81.3	81.8	31.3	68.9
MeZO-LoRA	89.6	67.9	67.8	73.5	63.5	60.2	61.3	84.1	81.5	82.1	31.3	69.4
SubZero	92.1	74.0	73.2	71.2	62.5	60.8	61.0	88.0	81.0	84.5	**32.0**	70.9
HiZOO	92.1	69.3	69.6	63.2	63.5	59.4	61.3	89.0	81.3	80.5	28.2	68.9
LOZO	91.7	70.4	69.6	71.9	63.5	60.8	62.0	89.0	81.3	84.9	30.7	70.5
AHZO	**93.7**	**74.7**	**75.0**	**74.6**	**64.4**	**62.5**	**62.7**	**90.0**	**81.8**	**85.7**	**32.0**	**72.5**

Table 4: Performance comparison on Qwen2.5-32B across tasks.

Task	SST-2	RTE	BoolQ	WSC	WiC	SQuAD	Average
Zero-shot	61.5	87.0	84.3	80.8	61.8	85.5	76.8
MeZO	87.5	90.3	86.7	81.7	65.7	87.9	83.3
MeZO-LoRA	62.0	87.4	84.8	82.7	64.7	84.7	77.7
SubZero	89.1	90.3	87.7	82.7	68.8	89.6	84.7
LOZO	89.4	**91.0**	87.0	84.6	69.3	90.5	85.3
AHZO	**92.3**	**91.0**	**88.7**	**86.5**	**74.5**	**92.1**	**87.5**

Table 5: Ablation study.

Task/Model	Qwen2.5-7B		OPT-13B		Average
	SST-2	RTE	SST-2	RTE	
Ours	**93.8**	**89.5**	**93.7**	**74.7**	**87.9**
Random	92.5	88.1	92.8	73.4	86.7
Fixed	93.0	86.6	90.9	69.7	85.1

6.3 Ablation Study for Subspace Selection

To validate the efficacy of the proposed dynamic subspace alignment mechanism, we conduct controlled ablation studies, keeping the optimization framework and total query budget constant while varying only the subspace selection strategy. Table 5 presents a comparison between our approach and two variants: 1) Random, which employs randomly sampled subspaces at each step; and 2) Fixed, which utilizes a static, pre-defined subspace throughout the entire optimization process. The empirical results demonstrate that our method consistently outperforms both baselines across all models and tasks. These findings underscore the critical importance of subspace updates in accurately capturing the evolving gradient information during LLM fine-tuning.

6.4 Sensitivity Analysis of Hyperparameters

We conduct a sensitivity analysis to study how the subspace configuration affects optimization performance (Figure 4). **For subspace rank (r):** The subspace rank r determines the number of basis

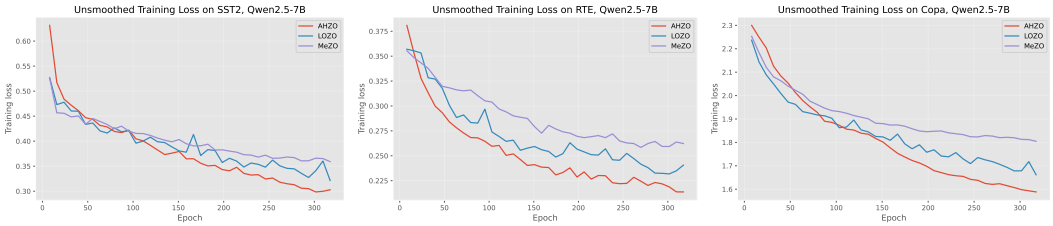

Figure 3: Comparison of Training Loss Dynamics Between AHZO, LOZO, and MeZO

directions maintained during optimization. The results indicates that a larger subspace can preserve richer gradient information and provide more effective update directions. **For exploit rank (p):** The exploit rank p controls the number of dominant singular directions used for greedy exploitation. The results show that different choices of p exhibit similar decreasing trends, with moderate values such as $p = 4$ and $p = 6$ achieving slightly lower training loss than overly small or large choices, suggesting the importance of balancing exploitation and exploration. **For exploit interval (ν):** The exploit interval ν determines the duration of each exploitation phase and the frequency of subspace updates. The results suggest that, within this range, appropriately increasing the period length can enhance the quality of subspace estimation, thereby yielding more effective optimization in large-scale fine-tuning. **For learning rate (η):** The learning rate η is a critical hyperparameter that balances update step sizes with the precision of subspace alignment. Figure 5 in Appendix B.5 indicates that the AHZO algorithm can adapt to a significantly larger learning rate and achieve faster convergence compared to the LOZO algorithm, which typically requires a much smaller range (10^{-7} to 10^{-6}). Furthermore, these findings suggest that within the stable range defined by our theory, a relatively higher learning rate can accelerate the exploitation of the estimated gradient subspace, thereby substantially improving the overall optimization efficiency for large-scale models.

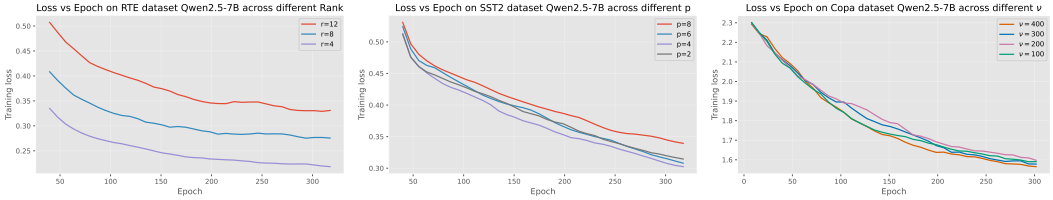

Figure 4: Sensitivity analysis for r, p and ν

7 Conclusion

In this paper, we proposed AHZO, an efficient low-rank ZO optimization framework for LLM fine-tuning. Motivated by the inherent directional coherence of optimization trajectories, AHZO leverages the AHG as a principled proxy for the true gradient manifold. By applying SVD to AHG matrices, AHZO dynamically extracts dominant spectral directions and constructs low-rank bases that better align with true gradient subspaces. Our theoretical analysis shows that AHZO reduces estimation variance and improves convergence compared with existing random-subspace ZO methods. Extensive experiments across different LLM architectures (Qwen and OPT) and benchmark tasks (classification, multiple-choice and generative) further demonstrate that AHZO consistently outperforms state-of-the-art ZO baselines, substantially narrowing the gap with memory-intensive FO fine-tuning while largely preserving the memory efficiency of inference-level optimization.

Acknowledgments

Dr. Yi Chang was supported by the National Key R&D Program of China (Grant No. 2023YFF0905400), the National Natural Science Foundation of China (Grant No. U2341229), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM903), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

[1] DM Anisuzzaman, Jeffrey G Malins, Paul A Friedman, and Zachi I Attia. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184, 2025.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[4] Nail Bashirov, Alexander Gasnikov, and Aleksandr Lobanov. Zeroth-order methods for nonsmooth stochastic problems under heavy-tailed noise. *Optimization Methods and Software*, pages 1–26, 2026.

[5] Lejs Deen Behric, Liang Zhang, Bingcong Li, and Kiran Koshy Thekumparampil. Conmezo: Adaptive descent-direction sampling for gradient-free finetuning of large language models. *arXiv preprint arXiv:2511.02757*, 2025.

[6] 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*, pages 15–26, 2017.

[7] Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. *arXiv preprint arXiv:2410.07698*, 2024.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[9] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128, 2024.

[11] Yasong Feng and Tianyu Wang. Stochastic zeroth-order gradient and hessian estimators: variance reduction and refined bias bounds. *Information and Inference: A Journal of the IMA*, 12(3):1514–1545, 2023.

[12] Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. *arXiv preprint arXiv:2404.08080*, 2024.

[13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[15] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 5254–5276, 2023.

[16] Cheonsu Jeong. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*, 2024.

[17] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pages 3100–3109. PMLR, 2019.

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 365–374, 2024.

[20] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[21] Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse mezo: Less parameters for better performance in zeroth-order llm fine-tuning. *arXiv preprint arXiv:2402.15751*, 2024.

[22] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

[23] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[24] Konstantinos Nikolakakis, Farzin Haddadpour, Dionysis Kalogerias, and Amin Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in neural information processing systems*, 35:31525–31541, 2022.

[25] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[26] Junhong Shen, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolo Fusi. Tag-llm: Repurposing general-purpose llms for specialized domains. *arXiv preprint arXiv:2402.05140*, 2024.

[27] Zichen Song and Weijia Li. Rozo: Geometry-aware zeroth-order fine-tuning on low-rank adapters for black-box large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1773–1783, 2026.

[28] Yan Sun, Tiansheng Huang, Liang Ding, Li Shen, and Dacheng Tao. Tezo: Empowering the low-rankness on the temporal dimension in the zeroth-order optimization for fine-tuning llms. *arXiv preprint arXiv:2501.19057*, 2025.

[29] Qitao Tan, Jun Liu, Zheng Zhan, Caiwei Ding, Yanzhi Wang, Xiaolong Ma, Jaewoo Lee, Jin Lu, and Geng Yuan. Harmony in divergence: Towards fast, accurate, and memory-efficient zeroth-order llm fine-tuning. *arXiv preprint arXiv:2502.03304*, 2025.

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[31] 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*, pages 353–355, 2018.

[32] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.

[33] Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *Advances in Neural Information Processing Systems*, 37:54905–54931, 2024.

[34] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International conference on artificial intelligence and statistics*, pages 1356–1365. PMLR, 2018.

[35] Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

[36] Bingyang Wu, Ruidong Zhu, Zili Zhang, Peng Sun, Xuanzhe Liu, and Xin Jin. {dLoRA}: Dynamically orchestrating requests and adapters for {LoRA}{LLM} serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 911–927, 2024.

[37] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9514–9528, 2021.

[38] Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *Advances in neural information processing systems*, 37:9474–9506, 2024.

[39] 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*, pages 4475–4485, 2025.

[40] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

[41] Jia Zhang, Yu Bai, Hualin Zhang, Tianshuo Chen, Zhaogeng Liu, Yi Chang, Bin Gu, et al. Three forward, one backward: Memory-efficient full-rank fine-tuning of large models via extra forward passes. In *The Fourteenth International Conference on Learning Representations*, 2026.

[42] 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.

[43] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv preprint arXiv:2402.11592*, 2024.

[44] Hanzhen Zhao, Shihong Ding, Cong Fang, and Zhouchen Lin. Pazo: Preconditioned accelerated zeroth-order optimization for fine-tuning llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[45] Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W Tsang. Second-order fine-tuning without pain for llms: A hessian informed zeroth-order optimizer. *arXiv preprint arXiv:2402.15173*, 2024.

A Convergence analysis

A.1 Assumptions

Assumption 4. *For any ξ, the function $\mathcal{F}(\mathbf{W};\xi)$ is differentiable with respect to $\mathbf{W}$. Furthermore, the gradient $\nabla\mathcal{F}(\mathbf{W};\xi)$ is uniformly L-Lipschitz continuous, i.e., $\forall\,\mathbf{W}_1,\,\mathbf{W}_2\in\mathbb{R}^{m\times n}$:*

$$\left\|\nabla\mathcal{F}(\mathbf{W}_1;\xi)-\nabla\mathcal{F}(\mathbf{W}_2;\xi)\right\|_F\le L\left\|\mathbf{W}_1-\mathbf{W}_2\right\|_F. \tag{22}$$

Assumption 5. ***Bounded Variance:*** *The stochastic gradient $\mathcal{F}(\mathbf{W};\xi)$ is unbiased and has bounded variance:*

$$\mathbb{E}_\xi\left[\|\nabla\mathcal{F}(\mathbf{W};\xi)-\nabla\mathcal{F}(\mathbf{W})\|_F^2\right]\le\sigma^2. \tag{23}$$

Assumption 6. ***Bounded initial optimality gap:*** *Given an initial point $\mathbf{W}_0$, and $\mathbf{W}^\star$ denotes a global minimizer of $\mathcal{F}$, there exists a constant $H<\infty$ that:*

$$\mathcal{F}(\mathbf{W}_0)-\mathcal{F}(\mathbf{W}^\star)\le H. \tag{24}$$

A.2 Lemmas

Lemma 1. ***Low-Rank Parameterization Preserves Inner Products.*** *For $\mathbf{U}\in\mathbb{R}^{m\times r}$ with $\mathbf{U}^\top\mathbf{U}=I_r$ and any $\mathbf{V}_1,\mathbf{V}_2\in\mathbb{R}^{n\times r}$,*

$$\langle\mathbf{U}\mathbf{V}_1^\top,\mathbf{U}\mathbf{V}_2^\top\rangle=\langle\mathbf{V}_1,\mathbf{V}_2\rangle,\quad\|\mathbf{U}\mathbf{V}_1^\top\|_F=\|\mathbf{V}_1\|_F. \tag{25}$$

Proof. Using the cyclic property of the trace, we can get:

$$\langle\mathbf{U}\mathbf{V}_1^\top,\mathbf{U}\mathbf{V}_2^\top\rangle=\mathrm{tr}(\mathbf{U}\mathbf{V}_1^\top\mathbf{V}_2\mathbf{U}^\top)=\mathrm{tr}(\mathbf{V}_1^\top\mathbf{V}_2\mathbf{U}^\top\mathbf{U})=\mathrm{tr}(\mathbf{V}_1^\top\mathbf{V}_2)=\langle\mathbf{V}_1,\mathbf{V}_2\rangle_F. \tag{26}$$

Let $\mathbf{V}_1=\mathbf{V}_2$, we can get:

$$\langle\mathbf{U}\mathbf{V}_1^\top,\mathbf{U}\mathbf{V}_1^\top\rangle=\langle\mathbf{V}_1,\mathbf{V}_1\rangle. \tag{27}$$

This step is equivalent to:

$$\|\mathbf{U}\mathbf{V}_1^\top\|_F^2=\|\mathbf{V}_1\|_F^2. \tag{28}$$

Taking the square root of both sides yields:

$$\|\mathbf{U}\mathbf{V}_1^\top\|_F=\|\mathbf{V}_1\|_F. \tag{29}$$

□

Lemma 2. ***Unbiasedness of Two-Point Zeroth-Order Estimator.*** *Define the smoothed function:*

$$\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W}):=\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)], \tag{30}$$

then the estimator

$$\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V}):=\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top \tag{31}$$

satisfies

$$\mathbb{E}_{\mathbf{V}}[\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V})]=\mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W})), \tag{32}$$

where the orthogonal projection onto $\mathbf{U}$ is $\mathcal{P}_{\mathbf{U}}(G)=\mathbf{U}\mathbf{U}^\top G$.

Proof. By the linearity of expectation, we expand the estimator:

$$\mathbb{E}_{\mathbf{V}}[\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V})]=\frac{1}{2\mu}\left(\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)\mathbf{U}\mathbf{V}^\top]-\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)\mathbf{U}\mathbf{V}^\top]\right). \tag{33}$$

Since $\mathbf{V}\sim\mathcal{N}(0,\mathbf{I})$ is symmetric about the origin, let $\mathbf{V}'=-\mathbf{V}$, then we can get:

$$\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)\mathbf{U}\mathbf{V}^\top]=\mathbb{E}_{\mathbf{V}'}[\mathcal{F}(\mathbf{W}+\mu\mathbf{U}(\mathbf{V}')^\top)(-\mathbf{U}(\mathbf{V}')^\top)]=-\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)\mathbf{U}\mathbf{V}^\top]. \tag{34}$$

Substituting this back to Eq.(33), we obtain:

$$\mathbb{E}_{\mathbf{V}}[\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V})] = \frac{1}{\mu}\mathbb{E}_{\mathbf{V}}[\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)\mathbf{U}\mathbf{V}^\top]. \tag{35}$$

Let $\mathcal{H}(\mathbf{V}) = \mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)$, by the chain rule, the gradient with respect to $\mathbf{V}$ is:

$$\nabla_{\mathbf{V}}\mathcal{H}(\mathbf{V}) = \mu\mathbf{U}^\top\nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top). \tag{36}$$

Applying the Stein's Lemma: $\mathbb{E}_x[f(x)x] = \mathbb{E}_x[\nabla_x f(x)]$, we can get:

$$\mathbb{E}_{\mathbf{V}}[\mathcal{H}(\mathbf{V})\mathbf{V}^\top] = \mathbb{E}_{\mathbf{V}}[\nabla_{\mathbf{V}}\mathcal{H}(\mathbf{V})] = \mu\mathbb{E}_{\mathbf{V}}[\mathbf{U}^\top\nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)]. \tag{37}$$

Substituting the result back into the estimator in Eq.(35):

$$\mathbb{E}_{\mathbf{V}}[\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V})] = \frac{1}{\mu}\cdot\mathbf{U}\cdot\big(\mu\mathbb{E}_{\mathbf{V}}[\mathbf{U}^\top\nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)]\big) = \mathbf{U}\mathbf{U}^\top\mathbb{E}_{\mathbf{V}}[\nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)]. \tag{38}$$

By the definition of the smoothed function $\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W})$, we have $\nabla\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W}) = \mathbb{E}_{\mathbf{V}}[\nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)]$. Thus we can get:

$$\mathbb{E}_{\mathbf{V}}[\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V})] = \mathbf{U}\mathbf{U}^\top\nabla\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W}) = \mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}_{\mu,\mathbf{U}}(\mathbf{W})). \tag{39}$$

□

Lemma 3. ***Unbiasedness of True Projected Gradient.*** *The projection of the true gradient satisfies the following inequality:*

$$\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top] = \mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}(\mathbf{W})). \tag{40}$$

Proof. Let $\omega = \mathbf{U}^\top\nabla\mathcal{F}(\mathbf{W}) \in \mathbb{R}^{r\times n}$, using the invariance under transposition, then we can get:

$$\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle = \mathrm{tr}(\nabla\mathcal{F}(\mathbf{W})^\top\mathbf{U}\mathbf{V}^\top) = \mathrm{tr}(\mathbf{V}\mathbf{U}^\top\nabla\mathcal{F}(\mathbf{W})) = \mathrm{tr}(V\omega) = \sum_{a=1}^{n}\sum_{b=1}^{r}\mathbf{V}_{a,b}\,\omega_{b,a}, \tag{41}$$

then, we multiply both sides by $\mathbf{V}_{j,i}$ where $j = 1,...,r$ and $i = 1,...,n$, we can get:

$$\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\,\mathbf{V}_{j,i}] = \sum_{a=1}^{n}\sum_{b=1}^{r}\mathbb{E}_{\mathbf{V}}[\mathbf{V}_{a,b}\mathbf{V}_{j,i}]\,\omega_{b,a}. \tag{42}$$

Since the entries of $\mathbf{V}_{(\cdot,\cdot)}$ are independent standard Gaussian random variables, we can get:

$$\mathbb{E}_{\mathbf{V}}[\mathbf{V}_{a,b}\mathbf{V}_{j,i}] = \delta_{b,j}\,\delta_{a,i}, \tag{43}$$

where $\delta_{(\cdot,\cdot)}$ denotes the Kronecker delta as following:

$$\delta_{k,l} = \begin{cases} 1 & \text{if } k = l\,, \\ 0 & \text{if } k \neq l\,. \end{cases} \tag{44}$$

Therefore:

$$\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\,\mathbf{V}_{j,i}] = \omega_{i,j}. \tag{45}$$

This shows that:

$$\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\,\mathbf{V}^\top] = \omega. \tag{46}$$

We can get:

$$\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top] = \mathbf{U}\,\mathbb{E}_{\mathbf{V}}[\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\,\mathbf{V}^\top] = \mathbf{U}\mathbf{U}^\top\nabla\mathcal{F}(\mathbf{W}) = \mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}(\mathbf{W})). \tag{47}$$

□

Lemma 4. ***Smoothing Bias Bound.*** *Under Assumption 1, the smoothing bias can be bounded by following:*

$$\mathbb{E}_{\mathbf{V}} \left\| \widehat{\nabla}\mathcal{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}) - \langle \nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top \rangle \mathbf{U}\mathbf{V}^\top \right\|_F^2 \leq \frac{L^2\mu^2 nr(nr+2)(nr+4)}{4}. \tag{48}$$

Proof. The L-smoothness property in Assumption 1 implies the following inequalities:

$$-\frac{L}{2}\left\|\mu\mathbf{U}\mathbf{V}^\top\right\|_F^2 \leq \mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) - \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle \leq \frac{L}{2}\left\|\mu\mathbf{U}\mathbf{V}^\top\right\|_F^2, \tag{49}$$

$$-\frac{L}{2}\left\|\mu\mathbf{U}\mathbf{V}^\top\right\|_F^2 \leq \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) + \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle \leq \frac{L}{2}\left\|\mu\mathbf{U}\mathbf{V}^\top\right\|_F^2. \tag{50}$$

Taking the absolute value of both sides yields:

$$\left|\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) - \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \leq \frac{L\mu^2}{2}\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2, \tag{51}$$

$$\left|\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) + \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \leq \frac{L\mu^2}{2}\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2, \tag{52}$$

then, a symmetric difference error bound as following:

$$\begin{aligned}
&\left|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu} - \langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \\
&= \frac{1}{2\mu}\left|\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top) - 2\mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \\
&\leq \frac{1}{2\mu}\left|\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) - \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \\
&\quad + \frac{1}{2\mu}\left|\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}) + \mu\langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\right| \\
&\overset{(1)}{\leq} \frac{L\mu}{2}\|\mathbf{U}\mathbf{V}^\top\|_F^2,
\end{aligned} \tag{53}$$

where (1) use Eq.(51) and Eq.(52).

Multiplying both sides by $\|\mathbf{U}\mathbf{V}^\top\|_F$, we can get:

$$\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top - \langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F \leq \frac{L\mu}{2}\|\mathbf{U}\mathbf{V}^\top\|_F^3. \tag{54}$$

Using Lemma 1 we can get:

$$\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top - \langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F \leq \frac{L\mu}{2}\|\mathbf{V}^\top\|_F^3. \tag{55}$$

Squaring both sides yields:

$$\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top) - \mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top - \langle\nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F^2 \leq \frac{L^2\mu^2}{4}\|\mathbf{V}^\top\|_F^6. \tag{56}$$

Taking the expectation over $\mathbf{V}$, and along with the sixth-order Gaussian moment

$$\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^6\right] = nr(nr+2)(nr+4), \tag{57}$$

according to definition in Eq.(31), we obtain:

$$\mathbb{E}_{\mathbf{V}} \left\| \widehat{\nabla}\mathcal{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}) - \langle \nabla\mathcal{F}(\mathbf{W}), \mathbf{U}\mathbf{V}^\top \rangle \mathbf{U}\mathbf{V}^\top \right\|_F^2 \leq \frac{L^2\mu^2 nr(nr+2)(nr+4)}{4}. \tag{58}$$

□

Lemma 5. ***Second-order Moment Bound for Estimators.*** *Under Assumptions 1 - 3, there exist inequality such that*

$$\mathbb{E}_{\xi,\mathbf{V}}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V};\xi)\right\|_F^2\right] \leq 3(nr+2)(\|\mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2+\sigma^2)+\frac{3L^2\mu^2nr(nr+2)(nr+4)}{4}. \tag{59}$$

Proof. From the symmetric difference decomposition, we can get:

$$\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top;\xi)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top;\xi)}{2\mu} = \frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}+\frac{\zeta_+-\zeta_-}{2\mu}, \tag{60}$$

where $\zeta_\pm := \mathcal{F}(\mathbf{W}\pm\mu\mathbf{U}\mathbf{V}^\top;\xi)-\mathcal{F}(\mathbf{W}\pm\mu\mathbf{U}\mathbf{V}^\top)$.

Taking the expectation of ξ and $\mathbf{V}$, we can get:

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{V}}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V};\xi)\right\|_F^2\right]\\
&=\mathbb{E}_{\xi,\mathbf{V}}\left[\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top;\xi)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top;\xi)}{2\mu}\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&=\mathbb{E}_{\xi,\mathbf{V}}\left[\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top+\frac{\zeta_+-\zeta_-}{2\mu}\cdot\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\overset{(1)}{\leq} 3\underbrace{\mathbb{E}_{\mathbf{V}}\left[\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top-\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]}_{a)}\\
&\quad+3\underbrace{\mathbb{E}_{\mathbf{V}}\left[\left\|\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]}_{b)}+3\underbrace{\mathbb{E}_{\mathbf{V},\xi}\left[\left\|\frac{\zeta_+-\zeta_-}{2\mu}\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]}_{c)},
\end{aligned} \tag{61}$$

where (1) use the inequality: $\|a+b+c\|^2 \leq 3\left(\|a\|^2+\|b\|^2+\|c\|^2\right)$.

For term a), use Lemma 4 we can get:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{V}}\left[\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top-\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\leq\frac{L^2\mu^2nr(nr+2)(nr+4)}{4}.
\end{aligned} \tag{62}$$

For term b), we perform the following transformation:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{V}}\left[\left\|\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&=\mathbb{E}_{\mathbf{V}}\left[\left\|\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\right\|_F^2\cdot\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\overset{(1)}{=}\mathbb{E}_{\mathbf{V}}\left[\left\|\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\mathbf{V}^\top\rangle\right\|_F^2\right]\cdot\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^2\right]\\
&=\mathbb{E}_{\mathbf{V}}\left[\|\langle\nabla\mathcal{F}(\mathbf{W}),\mathbf{U}\rangle\|_F^2\right]\cdot\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^4\right]\\
&\overset{(2)}{=}(nr+2)\left\|\mathbf{U}^\top\nabla\mathcal{F}(\mathbf{W})\right\|_F^2\\
&\overset{(3)}{=}(nr+2)\left\|\mathbf{U}\mathbf{U}^\top\nabla\mathcal{F}(\mathbf{W})\right\|_F^2\\
&=(nr+2)\left\|\mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}(\mathbf{W}))\right\|_F^2,
\end{aligned} \tag{63}$$

where (1) and (3) use Lemma 1; (2) follows from the fourth-order moment identity of standard Gaussian variables: $\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^4\right]=(nr+2)$.

For term c), we perform the following scaling:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{V},\xi}\left[\left\|\frac{\zeta_+-\zeta_-}{2\mu}\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&=\mathbb{E}_{\mathbf{V},\xi}\left[\left\|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top;\xi)-\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top;\xi)+\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&=\mathbb{E}_{\mathbf{V},\xi}\left[\left|\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top;\xi)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top;\xi)}{2\mu}-\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{U}\mathbf{V}^\top)-\mathcal{F}(\mathbf{W}-\mu\mathbf{U}\mathbf{V}^\top)}{2\mu}\right|^2\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\overset{(1)}{=}\mathbb{E}_{\mathbf{V},\xi}\left[\left|\frac{1}{2}\int_{-1}^{1}\left\langle\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top;\xi),\mathbf{U}\mathbf{V}^\top\right\rangle dt-\frac{1}{2}\int_{-1}^{1}\left\langle\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top),\mathbf{U}\mathbf{V}^\top\right\rangle dt\right|^2\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&=\mathbb{E}_{\mathbf{V},\xi}\left[\left|\frac{1}{2}\int_{-1}^{1}\left\langle\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top;\xi)-\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top),\mathbf{U}\mathbf{V}^\top\right\rangle dt\right|^2\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\overset{(2)}{\leq}\mathbb{E}_{\mathbf{V},\xi}\left[\frac{1}{2}\int_{-1}^{1}\left|\left\langle\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top;\xi)-\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top),\mathbf{U}\mathbf{V}^\top\right\rangle\right|^2 dt\cdot\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^2\right]\\
&\overset{(3)}{\leq}\mathbb{E}_{\mathbf{V},\xi}\left[\frac{1}{2}\int_{-1}^{1}\left\|\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top;\xi)-\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top)\right\|_F^2 dt\cdot\left\|\mathbf{U}\mathbf{V}^\top\right\|_F^4\right]\\
&\overset{(4)}{=}\frac{1}{2}\int_{-1}^{1}\mathbb{E}_{\mathbf{V},\xi}\left[\left\|\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top;\xi)-\nabla\mathcal{F}(\mathbf{W}+t\mu\mathbf{U}\mathbf{V}^\top)\right\|_F^2\right]dt\cdot\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^4\right]\\
&\overset{(5)}{\leq}\frac{1}{2}\int_{-1}^{1}\sigma^2 dt\cdot\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^4\right]\\
&\overset{(6)}{=}(nr+2)\sigma^2,
\end{aligned}
\tag{64}
$$

where (1) uses the fundamental theorem of calculus; (2) uses the Jensen's inequality; (3) uses the Cauchy-Schwarz inequality; (4) uses $\|\mathbf{U}\mathbf{V}^\top\|_F=\|\mathbf{V}\|_F$ in Lemma 1; (5) uses Assumption 2; (6) uses the fourth-order moment identity of standard Gaussian variables: $\mathbb{E}_{\mathbf{V}}\left[\|\mathbf{V}\|_F^4\right]=(nr+2)$.

Substitute a), b) and c) into Eq.(61), we can get:

$$
\mathbb{E}_{\xi,\mathbf{V}}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W},\mathbf{U},\mathbf{V};\xi)\right\|_F^2\right]\leq 3(nr+2)(\|\mathcal{P}_{\mathbf{U}}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2+\sigma^2)+\frac{3L^2\mu^2nr(nr+2)(nr+4)}{4}.
\tag{65}
$$

□

Lemma 6. ***One-step Descent.*** *Under Assumptions 1 - 3, at the t-th iteration, there exist inequality such that:*

$$
\begin{aligned}
&\mathcal{F}(\mathbf{W}_{t+1})-\mathcal{F}(\mathbf{W}_t)\\
\leq&\frac{\eta}{4r}\left\|\nabla\mathcal{F}(\mathbf{W}_t)\right\|_F^2+\left(\frac{3(nr+2)L\eta^2}{2r^2}-\frac{\eta}{r}\right)\left\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}))\right\|_F^2\\
&+\left(\frac{3L\eta^2+2r\eta}{2r}\right)\left(\frac{L^2\mu^2n(nr+2)(nr+4)}{4}\right)+\frac{3L\eta^2(nr+2)}{2r^2}\sigma^2.
\end{aligned}
\tag{66}
$$

Proof. By L-smoothness in Assumption 1, we can get:

$$
\mathcal{F}(\mathbf{W}_{t+1};\xi_t)\leq\mathcal{F}(\mathbf{W}_t;\xi_t)-\frac{\eta}{r}\langle\nabla\mathcal{F}(\mathbf{W}_t;\xi_t),\widehat{\nabla}\mathcal{F}(\mathbf{W}_t,\mathbf{U}_t,\mathbf{V}_t;\xi_t)\rangle+\frac{L\eta^2}{2r^2}\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t,\mathbf{U}_t,\mathbf{V}_t;\xi_t)\|_F^2.
\tag{67}
$$

Taking the conditional expectation with respect to ξ_t, we can get:

$$
\mathcal{F}(\mathbf{W}_{t+1})\leq\mathcal{F}(\mathbf{W}_t)-\frac{\eta}{r}\langle\nabla\mathcal{F}(\mathbf{W}_t),\widehat{\nabla}\mathcal{F}(\mathbf{W}_t,\mathbf{U}_t,\mathbf{V}_t)\rangle+\frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t}\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t,\mathbf{U}_t,\mathbf{V}_t;\xi_t)\|_F^2.
\tag{68}
$$

Simplifying the above equation, we can get:

$$
\begin{aligned}
&\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \\
&\leq -\frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&= -\frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad - \frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle \\
&\overset{(1)}{\leq} -\frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle\right\|_F^2\right), \qquad (69)
\end{aligned}
$$

where (1) uses the Young's inequality: $ab \leq a^2 + \frac{b^2}{4}$.

Taking the conditional expectation with respect to $\mathbf{V}_t$, we have

$$
\begin{aligned}
&\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \\
&\leq -\frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbb{E}_{\mathbf{V}_t}\left[\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\right]\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t, \mathbf{V}_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \mathbb{E}_{\mathbf{V}_t}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle\right\|_F^2\right]\right) \\
&\overset{(1)}{=} -\frac{\eta}{r}\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t, \mathbf{V}_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \mathbb{E}_{\mathbf{V}_t}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle\right\|_F^2\right]\right) \\
&= -\frac{\eta}{r}\langle \mathcal{P}_{U_t^\perp}(\nabla\mathcal{F}(\mathbf{W}_t)) + \mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t)), \mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\rangle + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t, \mathbf{V}_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \mathbb{E}_{\mathbf{V}_t}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle\right\|_F^2\right]\right) \\
&\overset{(2)}{=} -\frac{\eta}{r}\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 + \frac{L\eta^2}{2r^2}\mathbb{E}_{\xi_t, \mathbf{V}_t}\left[\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t, \xi_t)\|_F^2\right] \\
&\quad + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \mathbb{E}_{\mathbf{V}_t}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t) - \langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\rangle \mathbf{U}_t\mathbf{V}_t^\top\rangle\right\|_F^2\right]\right) \\
&\overset{(3)}{\leq} -\frac{\eta}{r}\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 + \frac{\eta}{r}\left(\frac{1}{4}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \frac{L^2\mu^2 nr(nr+2)(nr+4)}{4}\right) \\
&\quad + \frac{L\eta^2}{2r^2}\left(3(nr+2)(\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 + \sigma^2) + \frac{3L^2\mu^2 nr(nr+2)(nr+4)}{4}\right) \\
&= \frac{\eta}{4r}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \left(\frac{3(nr+2)L\eta^2}{2r^2} - \frac{\eta}{r}\right)\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&\quad + \left(\frac{3L\eta^2 + 2r\eta}{2r}\right)\left(\frac{L^2\mu^2 n(nr+2)(nr+4)}{4}\right) + \frac{3L\eta^2(nr+2)}{2r^2}\sigma^2, \qquad (70)
\end{aligned}
$$

where (1) uses Lemma 3; (2) uses the property of orthogonal projection; (3) uses Lemma 4 and Lemma 5. □

A.3 Theorems

Theorem 2. *Under Assumptions 1, 2 and 3, run our algorithm for $t = 0, ..., T-1$ steps with period ν, and at each step using the rank-r two-point estimator $\widehat{\nabla}\mathcal{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}, \xi)$, where $\mathbf{U}^\top\mathbf{U} = I_r$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Let $\varepsilon_{exploit,t} := \|\cos\Theta(\mathbf{U}_{exploit,t}, \mathbf{A}_{t,p})\|_2$, $I = \min\{m, n\}$, $\Sigma_t = diag(\{s_{t,i}\}_{i=1}^I)$ is the singular matrix by SVD decomposition for $\nabla\mathcal{F}(\mathbf{W}_t)$ and $\mathbf{A}_{t,p}$ is the*

corresponding top-p left singular vectors. Suppose the step size satisfies $0 < \eta < \frac{1}{4L(nr+2)}$, then the following inequality holds:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 \leq \frac{8\Delta}{3\eta T} + \sigma^2 + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \frac{5}{3}G_1(r-p)\cdot G_2(p), \quad (71)$$

where $G_1(r-p) = (1-\frac{r-p}{m-q})$, $G_2(p) = \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{i=1}^{I} s_{t,i}^2 - \varepsilon_{exploit,t}^2\sum_{i=1}^{p} s_{t,i}^2\right)$.

Proof. By Lemma 6, we can get:

$$\begin{aligned}
&\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \\
&\leq \frac{\eta}{4r}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 + \left(\frac{3(nr+2)L\eta^2}{2r^2} - \frac{\eta}{r}\right)\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&\quad + \left(\frac{3L\eta^2+2r\eta}{2r}\right)\left(\frac{L^2\mu^2 n(nr+2)(nr+4)}{4}\right) + \frac{3L\eta^2(nr+2)}{2r^2}\sigma^2. \\
&\overset{(1)}{\leq} \left(\frac{3(nr+2)L\eta^2}{2r^2} - \frac{3\eta}{4r}\right)\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 - \left(\frac{3(nr+2)L\eta^2}{2r^2} - \frac{\eta}{r}\right)\|\mathcal{P}_{U_t^\perp}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&\quad + \left(\frac{3L\eta^2+2r\eta}{2r}\right)\left(\frac{L^2\mu^2 n(nr+2)(nr+4)}{4}\right) + \frac{3L\eta^2(nr+2)}{2r^2}\sigma^2, \qquad (72)
\end{aligned}$$

where (1) uses the orthogonality between $\mathbf{U}_t$ and $\mathbf{U}_t^\perp$.

then, rearranging the inequality yields, let $0 < \eta < \frac{r}{4L(nr+2)}$, thus $\frac{3(nr+2)L\eta^2}{2r^2} - \frac{3\eta}{4r} \leq 0$, then we can get:

$$\begin{aligned}
&\left(\frac{3\eta}{4r} - \frac{3(nr+2)L\eta^2}{2r^2}\right)\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 \\
&\leq \mathcal{F}(\mathbf{W}_t) - \mathcal{F}(\mathbf{W}_{t+1}) + \left(\frac{\eta}{r} - \frac{3(nr+2)L\eta^2}{2r^2}\right)\|\mathcal{P}_{U_t^\perp}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&\quad + \left(\frac{3L\eta^2+2r\eta}{2r}\right)\left(\frac{L^2\mu^2 n(nr+2)(nr+4)}{4}\right) + \frac{3L\eta^2(nr+2)}{2r^2}\sigma^2. \qquad (73)
\end{aligned}$$

Since the following expression increases monotonically with η, we can get:

$$\frac{\frac{\eta}{r} - \frac{3(nr+2)L\eta^2}{2r^2}}{\frac{3\eta}{4r} - \frac{3(nr+2)L\eta^2}{2r^2}} < \frac{5}{3}. \quad (74)$$

Similarly, based on monotonicity, we can obtain:

$$\frac{3\eta}{4r} - \frac{3(nr+2)L\eta^2}{2r^2} > \frac{3\eta}{8r}, \quad (75)$$

$$\frac{3L\eta^2(nr+2)}{2r^2} < \frac{3\eta}{8r}. \quad (76)$$

By $nr > 1$ we can get $L\eta < \frac{r}{12}$, then we can get:

$$\frac{3L\eta^2+2r\eta}{2r} \leq \frac{9\eta}{8r} \quad (77)$$

Simplifying the above inequality, we get:

$$\begin{aligned}
&\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 \\
&\leq \frac{8r}{3\eta}\left(\mathcal{F}(\mathbf{W}_t) - \mathcal{F}(\mathbf{W}_{t+1})\right) + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \sigma^2 + \frac{5}{3}\|\mathcal{P}_{\mathbf{U}_t^\perp}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2. \qquad (78)
\end{aligned}$$

By summing over $t = 0, ..., T-1$ and divide both sides by T, we obtain:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 \\
&\leq \frac{8r(\mathcal{F}(\mathbf{W}_0)-\mathcal{F}(\mathbf{W}_T))}{3\eta T}+\sigma^2+\frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4)+\frac{5}{3}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{P}_{\mathbf{U}_t^\perp}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2 \\
&\overset{(1)}{\leq} \frac{8\Delta}{3\eta T}+\sigma^2+\frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4)+\frac{5}{3}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{P}_{\mathbf{U}_t^\perp}(\nabla\mathcal{F}(\mathbf{W}))\|_F^2, \qquad (79)
\end{aligned}$$

where (1) uses the Assumption 3.

For term $\|\mathcal{P}_{\mathbf{U}_t^\perp}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2$:

We first perform an SVD decomposition for $\nabla\mathcal{F}(\mathbf{W}_t)$:

$$\nabla\mathcal{F}(\mathbf{W}_t)=\mathbf{A}_t\mathbf{\Sigma}_t\mathbf{B}_t^\top=\sum_{i=1}^{I}a_{t,i}\cdot s_{t,i}\cdot(b_{t,i})^\top, \qquad (80)$$

where $I=\min\{m,n\}$; $\mathbf{A}_t=\{a_{t,i}\}_{i=1}^m$ is the left singular matrix; $\mathbf{\Sigma}_t=\text{diag}(\{s_{t,i}\}_{i=1}^I)$ is the singular matrix; $\mathbf{B}_t=\{b_{t,i}\}_{i=1}^n$ is the right singular matrix.

Based on the orthogonality of $\mathbf{A}_t$ and $\mathbf{B}_t$, we can obtain:

$$\begin{aligned}
&\|\mathcal{P}_{\mathbf{U}_t^\perp}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&=\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2-\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&=\left\|\sum_{i=1}^{I}a_{t,i}\cdot s_{t,i}\cdot(b_{t,i})^\top\right\|_F^2-\left\|\mathbf{U}_t\mathbf{U}_t^\top\cdot\sum_{i=1}^{I}a_{t,i}\cdot s_{t,i}\cdot(b_{t,i})^\top\right\|_F^2 \\
&\overset{(1)}{=}\sum_{i=1}^{I}s_{t,i}^2\cdot(a_{t,i}^\top a_{t,i})\cdot(b_{t,i}^\top b_{t,i})-\sum_{i=1}^{I}s_{t,i}^2\cdot(\mathbf{U}_t\mathbf{U}_t^\top\cdot a_{t,i})^\top(\mathbf{U}_t\mathbf{U}_t^\top\cdot a_{t,i})\cdot(b_{t,i})^\top b_{t,i} \\
&=\sum_{i=1}^{I}s_{t,i}^2-\sum_{i=1}^{I}s_{t,i}^2\cdot(\mathbf{U}_t\mathbf{U}_t^\top\cdot a_{t,i})^\top(\mathbf{U}_t\mathbf{U}_t^\top\cdot a_{t,i}) \\
&\overset{(2)}{=}\sum_{i=1}^{I}s_{t,i}^2-\sum_{i=1}^{I}s_{t,i}^2\cdot\left\|\mathbf{U}_t^\top\cdot a_{t,i}\right\|_F^2 \\
&=\sum_{i=1}^{I}s_{t,i}^2\cdot\left(1-\left\|\mathbf{U}_t^\top\cdot a_{t,i}\right\|_F^2\right), \qquad (81)
\end{aligned}$$

where (1) uses the properties of the Frobenius inner product of rank-one matrices; (2) use the column orthogonality of $\mathbf{U}_t$.

We can be obtained:

$$\|\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2=\sum_{i=1}^{I}s_{t,i}^2\cdot\left\|\mathbf{U}_t^\top\cdot a_{t,i}\right\|_F^2. \qquad (82)$$

1) **Quantifiable analysis for the exploit part**

In the ideal case, the exploitation step recovers the top-p left singular subspace:

$$\mathbf{A}_{t,p}:=[a_{t,1},\ldots,a_{t,p}]. \qquad (83)$$

If $\mathbf{U}_{\text{exploit},t}=\mathbf{A}_{t,p}$, then $\mathbf{U}_{\text{exploit},t}^\top\cdot\mathbf{A}_{t,p}=\mathbf{I}_p$, we can obtain:

$$\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2=\sum_{i=1}^{p}s_{t,i}^2. \qquad (84)$$

Thus, the residual after exploitation is

$$\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 = \sum_{i>p} s_{t,i}^2. \tag{85}$$

When $U_{\text{exploit},t}$ is not exact but only approximates the top-p subspace, let $\Theta(\cdot,\cdot)$ represent the angles in all directions spanned by these two subspaces. We measure the similarity using cosine:

$$\varepsilon_{\text{exploit},t} := |\cos\Theta(\mathbf{U}_{\text{exploit},t}, \mathbf{A}_{t,p})|, \tag{86}$$

then we have

$$\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \geq \varepsilon_{\text{exploit},t}^2 \sum_{i=1}^{p} s_{t,i}^2, \tag{87}$$

and consequently,

$$\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \leq \sum_{i>p} s_{t,i}^2 + (1-\varepsilon_{\text{exploit},t}^2)\sum_{i=1}^{p} s_{t,i}^2. \tag{88}$$

2) **Quantifiable analysis for the explore part**

The fundamental objective of exploration is to scavenge additional informative components from $\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))$.

When the exploration subspace $\mathbf{U}_{\text{explore},t}$ is randomly sampled within the orthogonal complement $\mathbf{U}_{\text{exploit},t}^{\perp}$, we have:

$$\mathbb{E}_{\mathbf{U}_{\text{explore},t}}\left[\|\mathcal{P}_{\mathbf{U}_{\text{explore},t}}(\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t)))\|_F^2\right] = \frac{r-p}{m-q}\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2. \tag{89}$$

Consequently, the expected norm of the residual gradient projected onto the orthogonal complement of the combined subspace $\mathbf{U}_t$ is given by:

$$\begin{aligned}
&\|\mathcal{P}_{\mathbf{U}_t^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 \\
&= \|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2 - \mathbb{E}\|\mathcal{P}_{\mathbf{U}_{\text{explore},t}}(\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t)))\|_F^2 \\
&= (1-\frac{r-p}{m-q})\|\mathcal{P}_{\mathbf{U}_{\text{exploit},t}^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2.
\end{aligned} \tag{90}$$

Substituting this back into Eq. (88), we arrive at the bound for the expected approximation error:

$$\|\mathcal{P}_{\mathbf{U}_t^{\perp}}(\nabla\mathcal{F}(\mathbf{W}_t)))\|_F^2 \leq (1-\frac{r-p}{m-q})\left(\sum_{i>p} s_{t,i}^2 + (1-\varepsilon_{\text{exploit},t}^2)\sum_{i=1}^{p} s_{t,i}^2\right). \tag{91}$$

Substituting back into Eq. (79) yields:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{F}(\mathbf{W}_t)\|_F^2 \\
&\leq \frac{8\Delta}{3\eta T} + \sigma^2 + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \frac{5}{3}\cdot\frac{1}{T}\sum_{t=0}^{T-1}(1-\frac{r-p}{m-q})\left(\sum_{i>p} s_{t,i}^2 + (1-\varepsilon_{\text{exploit},t}^2)\sum_{i=1}^{p} s_{t,i}^2\right) \\
&= \frac{8\Delta}{3\eta T} + \sigma^2 + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \frac{5}{3}\cdot\frac{1}{T}\sum_{t=0}^{T-1}(1-\frac{r-p}{m-q})\left(\sum_{i=1}^{I} s_{t,i}^2 - \varepsilon_{\text{exploit},t}^2\sum_{i=1}^{p} s_{t,i}^2\right) \\
&= \frac{8\Delta}{3\eta T} + \sigma^2 + \frac{3}{4}L^2\mu^2 nr(nr+2)(nr+4) + \frac{5}{3}\cdot G_1(p)\cdot G_2(r-p),
\end{aligned} \tag{92}$$

where $G_1(p) = \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{i=1}^{I} s_{t,i}^2 - \varepsilon_{\text{exploit},t}^2\sum_{i=1}^{p} s_{t,i}^2\right)$, $G_2(r-p) = (1-\frac{r-p}{m-q})$.

This formulation G_1 and G_2 explicitly decouples the distinct roles of the two mechanisms: **Exploitation** determines the initial magnitude of the residual $\frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{i=1}^{I} s_{t,i}^2 - \varepsilon_{\text{exploit},t}^2\sum_{i=1}^{p} s_{t,i}^2\right)$ by identifying the dominant singular vectors; **Exploration** further reduces this residual by a consistent scaling factor $(1-\frac{r-p}{m-q})$ through proactive searching within the orthogonal complement. □

Corollary 2. ***Discussion regarding*** $\varepsilon_{exploit}$*: Under Assumptions 1 and 2, At the* k*-th interval, let* $\rho = \max_{t=k\nu,\ldots,(k+1)\nu-1}\{\|\mathbf{W}_j - \mathbf{W}_{j+1}\|_F\}$, $\Delta\mathbf{W}_z = \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t)$ *and* $\sigma_z = \sigma_p(\Delta\mathbf{W}_z) - \sigma_{p+1}(\Delta\mathbf{W}_z)$ *is the singular value gap. We bound the similarity between the estimated subspace* $\mathbf{A}_{z,p}$ *and the optimal subspace* $\mathbf{A}_p$ *as following:*

$$|\cos\Theta(\mathbf{A}_p, \mathbf{A}_{z,p})| \geq \sqrt{1 - \frac{L^2\eta^2(\mu\nu\sqrt{nr} + \nu^2\cdot\rho)^2}{r^2\sigma_z^2}} \tag{93}$$

Proof. a At the k-th interval, we define the average historical ZO gradient as:

$$\Delta\mathbf{W}_z = \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t), \tag{94}$$

where its left singular vector is represented as $\mathbf{A}_z$.

The average historical true gradient is:

$$\Delta\mathbf{W}_f = \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\nabla\mathcal{F}(\mathbf{W}_t), \tag{95}$$

where its left singular vector is represented as $\mathbf{A}_f$.

Based on the fact that scaling by a positive number does not alter any singular vector properties, it follows that left singular vector of $\frac{\eta\nu}{r}\nabla\mathcal{F}(\mathbf{W})$ is same to $\nabla\mathcal{F}(\mathbf{W})$.

Let $\alpha = \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\left(\mathbb{E}_{\xi_t,\mathbf{V}_t}\left[\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t)\right] - \nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)$, according to Wedin's perturbation theorem, we can obtain:

$$|\sin\Theta(\mathbf{A}_p, \mathbf{A}_{z,p})| \leq \frac{\|\mathbf{A}_{z,p}^\top\alpha\|_F}{\sigma_p(\Delta\mathbf{W}_z) - \sigma_{p+1}(\Delta\mathbf{W}_z)} \tag{96}$$

Next, we will process the numerator term $\|\mathbf{A}_{z,p}^\top\alpha\|_F$:

$$\begin{aligned}
&\|\mathbf{A}_{z,p}^\top\alpha\|_F \\
&\overset{(1)}{=} \|\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\alpha\|_F \\
&= \left\|\frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\left(\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbb{E}_{\xi_t,\mathbf{V}_t}\left[\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t; \xi_t)\right] - \mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)\right\|_F \\
&\overset{(2)}{=} \left\|\frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\left(\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbb{E}_{\mathbf{V}_t}\left[\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)\right] - \mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)\right\|_F \\
&\overset{(3)}{=} \left\|\frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\left(\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathcal{P}_{\mathbf{U}_t}(\nabla\mathcal{F}_{\mu,\mathbf{U}_t}(\mathbf{W}_t)) - \mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)\right\|_F \\
&\overset{(4)}{=} \left\|\frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\left(\mathcal{P}_{\mathbf{A}_{z,p}}(\nabla\mathcal{F}_{\mu,\mathbf{U}_t}(\mathbf{W}_t)) - \mathcal{P}_{\mathbf{A}_{z,p}}\left(\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)\right)\right\|_F \\
&\overset{(5)}{\leq} \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\underbrace{\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\nabla\mathcal{F}_{\mu,\mathbf{U}_t}(\mathbf{W}_t) - \nabla\mathcal{F}(\mathbf{W}_t)\right)\right\|_F}_{a)} + \frac{\eta}{r}\sum_{t=k\nu}^{(k+1)\nu-1}\underbrace{\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\nabla\mathcal{F}(\mathbf{W}_t) - \nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right)\right\|_F}_{b)}
\end{aligned} \tag{97}$$

where (1) uses the in Lemma 1; (2) uses the Assumption 2; (3) uses Lemma 2; (4) uses the row orthogonality of $\mathbf{U}_t$ and $\mathbf{A}_{z,p}$. □

For term a), we perform the following operations:

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\nabla\mathcal{F}_{\mu,\mathbf{U}_t}(\mathbf{W}_t)-\nabla\mathcal{F}(\mathbf{W}_t)\right)\right\|_F^2\\
=&\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\mathbb{E}_{\mathbf{V}_t}[\nabla\mathcal{F}(\mathbf{W}_t+\mu\mathbf{U}_t\mathbf{V}_t^\top)]-\nabla\mathcal{F}(\mathbf{W}_t)\right)\right\|_F^2\\
\overset{(1)}{\leq}&\mathbb{E}_{\mathbf{V}_t}\left[\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\nabla\mathcal{F}(\mathbf{W}_t+\mu\mathbf{U}_t\mathbf{V}_t^\top)-\nabla\mathcal{F}(\mathbf{W}_t)\right)\right\|_F^2\right]\\
\overset{(2)}{\leq}&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\left(\mathbf{U}_t\mathbf{V}_t^\top\right)\right\|_F^2\right]\\
=&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\mathrm{tr}\left(\mathbf{V}_t\mathbf{U}_t^\top\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbf{U}_t\mathbf{V}_t^\top\right)\right]\\
=&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\mathrm{tr}\left(\mathbf{V}_t\mathbf{U}_t^\top\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbf{U}_t\mathbf{V}_t^\top\right)\right]\\
\overset{(3)}{=}&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\mathrm{tr}\left((\mathbf{U}_t^\top\mathbf{A}_{z,p}\mathbf{A}_{z,p}^\top\mathbf{U}_t)(\mathbf{V}_t^\top\mathbf{V}_t)\right)\right]\\
\overset{(4)}{\leq}&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\mathrm{tr}\left(\mathbf{V}_t^\top\mathbf{V}_t\right)\right]\\
=&L^2\mu^2\cdot\mathbb{E}_{\mathbf{V}_t}\left[\|\mathbf{V}_t\|_F^2\right]\\
\overset{(5)}{=}&L^2\mu^2nr,
\end{aligned}\tag{98}$$

where (1) use convexity of norm operations; (2) uses Assumption 1; (3) uses the cyclic property of the trace; (4) uses the Poincaré Separation Theorem; (5) uses the properties of the second moment of the standard normal distribution.

For term b), we perform the following operations:

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\Big(\nabla\mathcal{F}(\mathbf{W}_t)-\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\Big)\right\|_F\\
\overset{(1)}{\leq}&\left\|\nabla\mathcal{F}(\mathbf{W}_t)-\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\right\|_F\\
\overset{(2)}{\leq}&L\left\|\mathbf{W}_t-\mathbf{W}_{(k+1)\nu}\right\|_F\\
=&L\left\|\sum_{j=t}^{(k+1)\nu-1}(\mathbf{W}_j-\mathbf{W}_{j+1})\right\|_F.
\end{aligned}\tag{99}$$

where (1) uses the non-expansiveness of orthogonal projection; (2) is derived from the L-smoothness assumption (Assumption 1).

We assume that the upper bound of a single gradient update is ρ, then:

$$\left\|\mathcal{P}_{\mathbf{A}_{z,p}}\Big(\nabla\mathcal{F}(\mathbf{W}_t)-\nabla\mathcal{F}(\mathbf{W}_{(k+1)\nu})\Big)\right\|_F\leq L\nu\cdot\rho\tag{100}$$

Substitute a) and b) into Eq.(96):

$$|\sin\Theta(\mathbf{A}_p,\mathbf{A}_{z,p})|\leq\frac{L\eta(\mu\nu\sqrt{nr}+\nu^2\cdot\rho)}{r(\sigma_p(\Delta\mathbf{W}_z)-\sigma_{p+1}(\Delta\mathbf{W}_z))}.\tag{101}$$

We denote $\sigma_z=\sigma_p(\Delta\mathbf{W}_z)-\sigma_{p+1}(\Delta\mathbf{W}_z)$, then:

$$|\sin\Theta(\mathbf{A}_p,\mathbf{A}_{z,p})|\leq\frac{L\eta(\mu\nu\sqrt{nr}+\nu^2\cdot\rho)}{r\sigma_z}\tag{102}$$

Next, we calculate the similarity:

$$|\cos\Theta(\mathbf{A}_p,\mathbf{A}_{z,p})|\geq\sqrt{1-\frac{L^2\eta^2(\mu\nu\sqrt{nr}+\nu^2\cdot\rho)^2}{r^2\sigma_z^2}}\tag{103}$$

A.4 Descent Property of the Objective Function

Lemma 7. *Proof.* To derive the descent bound from the expectation form, we rely on two core properties of the ZO estimator constructed via Gaussian smoothing: the first moment (expected direction) and the second moment (variance bound).

For the Expected Gradient Signal (First Term): Based on the properties of Gaussian smoothing, the expectation of the ZO estimator equals the gradient of the smoothed function within the subspace $\mathcal{G}$ (Lemma 2):

$$\mathbb{E}[\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)] = \mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}_{\mu,\mathbf{U}_t}(\mathbf{W}_t)), \tag{104}$$

where $\mathcal{P}_{\mathcal{G}} = \mathbf{U}_t\mathbf{U}_t^T$ is the orthogonal projection operator onto $\mathcal{G}$. Under the L-smoothness assumption, the smoothing error is bounded by $\mathcal{O}(\mu)$. Thus, the inner product becomes:

$$-\eta \cdot \mathbb{E}\left[\left\langle \nabla\mathcal{F}(\mathbf{W}_t), \widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)\right\rangle\right] = -\eta \left\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\rangle + \mathcal{O}(\eta\mu). \tag{105}$$

Since $\mathcal{P}_{\mathcal{G}}$ is an orthogonal projection matrix, it is idempotent and symmetric ($\mathcal{P}_{\mathcal{G}} = \mathcal{P}_{\mathcal{G}}^T\mathcal{P}_{\mathcal{G}}$). The inner product naturally reduces to the squared Frobenius norm:

$$\left\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\rangle = \left\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathcal{P}_{\mathcal{G}}^T\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\rangle = \left\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\|_F^2. \tag{106}$$

Therefore, the first term simplifies to $-\eta \left\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\|_F^2 + \mathcal{O}(\eta\mu)$.

For the Variance Bound (Second Term) According to standard subspace ZO optimization theory, the second moment of the estimator in an r-dimensional subspace is bounded by the projected gradient norm and the smoothing error (Lemma 5):

$$\mathbb{E}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)\right\|_F^2\right] \leq \mathcal{O}(\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\|_F^2) + \mathcal{O}(L^2\mu^2). \tag{107}$$

Substituting this into the second term of the descent lemma:

$$\frac{L\eta^2}{2} \cdot \mathbb{E}\left[\left\|\widehat{\nabla}\mathcal{F}(\mathbf{W}_t, \mathbf{U}_t, \mathbf{V}_t)\right\|_F^2\right] \leq \mathcal{O}\left(L\eta^2 \left\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\|_F^2\right) + \mathcal{O}\left(L^3\eta^2\mu^2\right). \tag{108}$$

Synthesis: Combining the bounds from the first term and the second term, we obtain the full descent inequality:

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \leq \mathcal{O}\left((L\eta^2 - \eta)\left\|\mathcal{P}_{\mathcal{G}}(\nabla\mathcal{F}(\mathbf{W}_t))\right\|_F^2\right) + \mathcal{O}\left(L^3\eta^2\mu^2 + \eta\mu\right). \tag{109}$$

□

B Experimental Supplementary

B.1 Baseline Descriptions

Detailed descriptions of all baselines are provided below:

- **Zero-shot** evaluates the pretrained model directly without any task-specific adaptation.
- **In-Context Learning** [10] performs inference by prepending a small number of task-specific demonstrations to the input prompt, without updating model parameters.
- **Fine-tuning with Adam** [18] represents a FO optimization baseline, where all model parameters are updated using the Adam optimizer under the same training data budget.
- **FT-LoRA** [14] LoRA is an efficient fine-tuning strategy that freezes pre-trained model weights and injects trainable low-rank matrices into each layer of the Transformer architecture. By optimizing only these small rank-decomposition matrices instead of the entire high-dimensional weight space, LoRA significantly reduces the number of trainable parameters and GPU memory requirements.
- **MeZO** [22] is a memory-efficient ZO optimization method for fine-tuning large language models without backpropagation. It estimates gradients using symmetric random perturbations and requires only two forward passes per update step. By performing in-place parameter perturbations and avoiding the storage of intermediate activations and backward graphs, MeZO achieves a memory footprint comparable to inference.
- **MeZO-LoRA** [22] extends MeZO by restricting ZO optimization to low-rank adapter parameters introduced by LoRA, while keeping the pretrained backbone frozen. By reducing the effective optimization dimensionality, MeZO-LoRA significantly improves the stability and efficiency of gradient estimation.

- **HiZOO** [45] is a Hessian-informed ZO optimization method designed to improve the stability and efficiency of fine-tuning large language models under memory constraints. By incorporating diagonal Hessian information, HiZOO adaptively scales random perturbations and parameter updates to account for heterogeneous curvature across different parameter dimensions.
- **SubZero** [39] is a random subspace ZO optimization method designed to reduce the variance and improve the efficiency of ZO fine-tuning for large language models. Instead of performing random perturbations in the full parameter space, SubZero restricts gradient estimation to a low-dimensional random subspace, significantly reducing the effective perturbation dimension.
- **LOZO** [7] is a low-rank ZO optimization method designed to improve the efficiency of memory-constrained fine-tuning for large language models. By restricting ZO gradient estimation to a low-rank subspace, LOZO captures the intrinsic low-rank structure commonly observed in LLM gradients, thereby significantly reducing estimation variance.

B.2 The Hyperparameter Settings

Table 6 summarizes the hyperparameter grids used for AHZO and all baselines in our experiments on the Qwen and OPT models. Following the training setup of [7, 22], we train 20,000 steps for ZO method. For FT and FT-LoRA, we use one-tenth as many steps as ZO methods.

Table 6: The hyperparameter grids for experiments.

Qwen2.5-7B		
Experiment	**Hyperparameters**	**Values**
AHZO	Batch size	16
	Rank (r)	{4, 8}
	Interval (ν)	{50, 100, 200, 300}
	ϵ	{1e−2, 5e−2}
	Learning rate	{1e−3, 5e−4, 1e−4}
	p	{2, 4, 6}
LOZO	Batch size	16
	Rank (r)	{4, 8}
	Interval (ν)	{50, 100, 200,300}
	ϵ	1e−3
	Learning rate	{1e−6, 1e−7}
HiZOO	Batch size	16
	ϵ	1e−3
	Learning rate	1e−6
	Hessian smooth type	1e−8
SubZero	Batch size	16
	Gauss Rank	24
	ϵ	1e−3
	Learning rate	{1e−6, 1e−7}
	update interval	1000
	perturbation	two side
MeZO	Batch size	16
	ϵ	1e−3
	Learning rate	{1e−7, 5e−7}
MeZO-LoRA	Batch size	16
	(r, α)	(4, 16)
	ϵ	1e−2
	Learning rate	1e−5
FT-LoRA	Batch size	8
	(r, α)	(4, 16)
	Learning rate	1e−5
FT	Batch size	8
	Learning rate	1e−5

OPT-13B		
Experiment	**Hyperparameters**	**Values**
AHZO	Batch size	16
	Rank (r)	{4, 8}
	Interval (ν)	{50, 100, 200,300}
	ϵ	{1e−2, 5e−2}
	Learning rate	{1e−3, 5e−4, 1e−4}
	p	{2, 4, 6}
LOZO	Batch size	16
	Rank (r)	{2, 4, 8}
	Interval (ν)	{50, 100, 200, 300}
	ϵ	{1e−3, 1e−4}
	Learning rate	{1e−6, 1e−7}
HiZOO	Batch size	16
	ϵ	1e−3
	Learning rate	1e−6
	Hessian smooth type	1e−8
SubZero	Batch size	16
	Gauss Rank	24
	ϵ	1e−3
	Learning rate	{1e−6, 1e−7}
	update interval	1000
	perturbation	two side
MeZO	Batch size	16
	ϵ	1e−3
	Learning rate	{1e−7, 1e−6, 5e−7}
MeZO-LoRA	Batch size	16
	(r, α)	(8, 16)
	ϵ	1e−3
	Learning rate	{1e−5, 5e−5, 1e−4}
FT-LoRA	Batch size	8
	(r, α)	(4, 16)
	Learning rate	{1e−5, 5e−5, 8e−5}
FT	Batch size	8
	Learning rate	{1e−5, 5e−5, 8e−5}

B.3 Memory Consumption

As shown in Tables 7 and 8, AHZO achieves a memory footprint comparable to ZO (batch size = 16) and FO (batch size=8) methods designed for memory efficiency across all tasks on Qwen2.5-7B and OPT-13B, while remaining substantially more memory-efficient than FT.

Table 7: GPU memory consumption (GB) of different optimization methods on Qwen2.5-7B.

Task	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
FT	80.0	121.8	153.1	123.7	85.3	80.9	116.8	76.2	127.7	139.7	149.6	114.1
FT-LoRA	30.6	68.5	58.0	74.3	33.1	29.2	48.6	19.5	78.8	91	98.7	57.3
MeZO	16.4	22.4	23.7	37.4	17.3	17.9	56.9	15.6	26.9	34.0	53.1	29.2
MeZO-LoRA	17.4	22.2	23.6	37.5	16.8	18.0	56.7	15.4	27.1	33.9	54.5	29.4
HiZOO	40.0	46.6	50.0	54.6	41.3	40.0	77.9	39.6	46.9	52.2	69.2	50.8
SubZero	21.4	27.1	25.6	59.8	21.4	21.4	59.4	21.3	36.5	54.9	78.7	38.9
LOZO	21.0	25.3	30.0	46.5	21.0	21.0	59.5	20.9	31.2	46.5	77.4	36.4
AHZO	20.1	25.6	30.7	46.8	23.5	21.5	60.5	22.3	32.1	46.8	78.5	37.1

Table 8: GPU memory consumption (GB) of different optimization methods on OPT-13B.

Task	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
FT	103.0	140.6	167.0	203.1	107.4	108.5	198.0	101.2	109.8	219.0	176.1	148.5
FT-LoRA	38.4	78.6	109.7	142.5	40.9	41.9	107.6	30.5	41.6	175.7	109.9	83.4
MeZO	26.8	34.1	40.1	44.5	27.4	27.2	52.6	25.5	33.8	41.0	74.5	38.9
MeZO-LoRA	27.2	34.2	40.4	45.0	27.3	27.5	52.9	25.2	33.6	41.2	74.9	39.0
HiZOO	57.2	66.0	72.7	77.0	59.0	58.1	104.1	57.1	64.5	82.9	135.6	75.8
SubZero	29.1	39.5	48.6	45.3	28.4	28.2	63.3	27.4	36.4	48.3	73.7	42.6
LOZO	26.9	35.9	41.1	43.8	28.6	28.0	51.9	26.9	34.8	47.2	76.1	40.1
AHZO	27.3	35.8	41.5	43.9	29.7	28.8	51.3	27.8	35.5	48.8	74.8	40.5

B.4 Wall-clock Time

Tables 9 and 10 report the practical wall-clock runtime ($\nu = 100$) comparison between AHZO and LOZO across all tasks on Qwen2.5-7B and OPT-13B. The results show that since these decompositions are performed periodically rather than at every iteration, the overall runtime overhead remains minimal and is negligible in practice.

Table 9: Wall-clock time consumption ($hour$ (h) & $minute$ (m)) for training on Qwen2.5-7B.

Task	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
LOZO	4h31m	8h1m	9h9m	10h47m	4h1m	4h27m	19h49m	2h34m	6h49m	7h10m	12h43m	8h11m
AHZO	4h40m	8h22m	9h32m	11h1m	4h3m	4h46m	19h52m	2h46m	7h0m	7h17m	12h51m	8h23m

Table 10: Wall-clock time consumption ($hour$ (h) & $minute$ (m)) for training on OPT-13B.

Task	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
LOZO	5h8m	11h2m	14h11m	16h8m	6h11m	5h58m	34h9m	4h10m	10h38m	11h7m	20h21m	12h38m
AHZO	5h19m	11h14m	14h21m	16h16m	6h19m	6h6m	34h18m	4h16m	10h49m	11h18m	20h28m	12h48m

Table 11 further reports the wall-clock runtime breakdown of AHZO and LOZO on Qwen2.5-32B. Across all evaluated tasks, the forward and perturbation costs of AHZO remain nearly identical to those of LOZO, indicating that our method does not introduce additional overhead to the dominant computation.

Table 11: Wall-clock runtime comparison between AHZO and LOZO on Qwen2.5-32B (seconds).

Component	BoolQ		SST-2		WiC		WSC		Average	
	AHZO	LOZO	AHZO	LOZO	AHZO	LOZO	AHZO	LOZO	AHZO	LOZO
Forward	120874	120841	15531	15533	21887	21903	37142	37170	48859	48862
Perturb	6160	6215	6025	6067	6026	6069	6023	6067	6058	6105
SVD	466	0	462	0	463	0	463	0	464	0
QR	22	0	22	0	22	0	22	0	22	0
Total	127522	127057	22040	21600	28399	27972	43651	43236	55403	54966

B.5 Additional Sensitivity Analysis

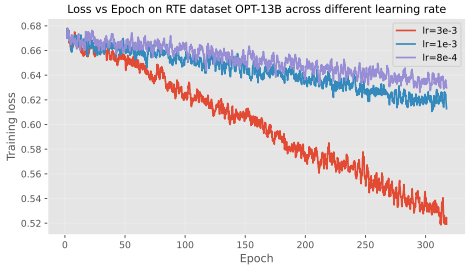

Figure 5: Sensitivity Analysis for learning rate

B.6 Additional Supplementary Experiments

We have added Sparse MeZO (S-MeZO) to the OPT-13B evaluation in Table 12. These results demonstrate that AHZO remains competitive against a recent sparse ZO fine-tuning method and that its gains are not limited to comparisons with earlier baselines.

Table 12: Comparison with S-MeZO on OPT-13B.

Method	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
S-MeZO	92.3	63.5	69.6	66.4	61.1	58.8	**63.3**	87.0	80.8	77.9	31.9	68.4
AHZO	**93.7**	**74.7**	**75.0**	**74.6**	**64.4**	**62.5**	**62.7**	**90.0**	**81.8**	**85.7**	**32.0**	**72.5**

Since the official implementation of TeZO is not publicly available, we compare AHZO with the results reported in the TeZO paper under the same OPT-13B setting and a matched training budget of 15,000 updates. As shown below in Table 13, AHZO achieves a higher average score of 69.05 compared with 67.19 for TeZO.

Table 13: Comparison of AHZO and TeZO on OPT-13B.

Method	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
TeZO	90.2	61.1	69.6	65.1	**63.5**	54.3	56.8	**87.0**	81.2	80.7	29.6	67.19
AHZO	**91.2**	**70.4**	**71.4**	**67.8**	**62.5**	**56.4**	**57.5**	**87.0**	**81.3**	**82.9**	**31.1**	**69.05**

For LOZO on Qwen, we initially omitted $r = 2$ because it performed consistently worse than $r = 4$ and $r = 8$ in our preliminary validation experiments. To make the comparison more complete, we have now included the result for $r = 2$ and will report it alongside the other rank settings. Similarly, for SubZero, we have added the original rank setting $r = 48$ in addition to $r = 24$ (Table 14).

Table 14: Performance comparison of different rank configurations on Qwen2.5-7B.

Method	SST2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Average
LOZO (reported)	90.6	86.6	**96.4**	84.3	**73.1**	70.2	87.9	90.0	76.6	87.5	60.9	82.2
LOZO ($r = 2$)	89.1	86.6	89.3	84.5	72.1	67.6	85.0	90.0	75.7	84.9	59.1	80.4
SubZero (reported)	88.1	88.4	89.3	84.1	72.1	69.3	87.7	90.0	74.6	86.9	59.6	80.9
SubZero ($r = 48$)	89.2	87.7	87.5	84.3	71.2	68.0	87.6	87.0	76.3	87.7	62.0	80.8
AHZO	**93.8**	**89.5**	**96.4**	**84.8**	**73.1**	**73.2**	**88.3**	**93.0**	**77.5**	**89.4**	**63.0**	**83.8**

We conducted five-seed (0,1,2,3,4) experiments for AHZO, LOZO, and SubZero on CB and WSC using OPT-13B and Qwen2.5-7B under the same experimental protocol. Table 15 reports the result of each seed, the mean, and the sample standard deviation. AHZO achieves the highest mean performance across all four task–model settings.

B.7 Empirical Validation of the Alignment Conditions.

We report the singular-value spectra σ_z of current FO-gradient matrices in Table 16, and we report the empirical within-period drift (ρ) in Table 17.

Table 15: Five-seed results on CB and WSC across OPT-13B and Qwen2.5-7B.

Task	Model	Method	0	1	2	3	4	Mean	Std
CB	OPT-13B	LOZO	0.696	0.518	0.536	0.661	0.518	0.586	0.086
		SubZero	0.714	0.679	0.661	0.679	0.679	0.682	0.020
		AHZO	0.714	0.696	0.714	0.679	0.679	**0.696**	**0.018**
	Qwen2.5-7B	LOZO	0.964	0.893	0.929	0.911	0.911	0.921	0.027
		SubZero	0.893	0.875	0.893	0.893	0.857	0.882	0.016
		AHZO	0.964	0.946	0.964	0.946	0.929	**0.950**	**0.015**
WSC	OPT-13B	LOZO	0.635	0.615	0.365	0.635	0.538	0.558	0.115
		SubZero	0.625	0.615	0.394	0.577	0.413	0.525	0.112
		AHZO	0.644	0.596	0.433	0.596	0.606	**0.575**	**0.082**
	Qwen2.5-7B	LOZO	0.731	0.635	0.673	0.663	0.731	0.687	0.043
		SubZero	0.721	0.712	0.663	0.663	0.673	0.687	0.028
		AHZO	0.731	0.702	0.702	0.721	0.702	**0.712**	**0.014**

Table 16: Singular-value gap of FO gradient (500 step).

Dataset/Model	Layer	0	100	200	300	400	500	Mean
Copa/Qwen2.5-7B	down_proj L27	3.99	3.97	3.93	3.83	3.79	3.73	3.87
	q_proj L18	2.42	2.38	2.37	2.41	2.42	2.24	2.37
	k_proj L16	2.16	2.15	2.13	2.08	2.07	2.00	2.10
	v_proj L17	1.96	1.92	1.90	1.88	1.87	1.82	1.89
SST2/opt-13b	down_proj L16	12.09	10.48	10.88	9.93	10.38	9.22	10.50
	q_proj L17	11.50	10.26	10.24	9.62	9.76	8.93	10.05
	k_proj L18	11.42	10.20	10.22	9.72	9.89	8.61	10.01
	v_proj L15	11.35	10.04	10.24	9.41	9.68	8.62	9.89

Table 17: Parameter drift (every 5k steps).

Dataset/Model	View	5k	10k	15k	20k	Avg
Copa/Qwen2.5-7B	overall	0.30	0.25	0.31	0.28	0.29
	q_proj	0.09	0.07	0.09	0.08	0.08
	down_proj	0.20	0.17	0.21	0.19	0.19
SST2/opt-13b	overall	0.43	0.48	0.59	0.62	0.53
	q_proj	0.14	0.16	0.20	0.21	0.18
	down_proj	0.28	0.32	0.39	0.41	0.35

C Limitations

Compared with standard MeZO, AHZO requires storing an additional small low-rank matrix, resulting in a modest increase in memory overhead. In addition, AHZO introduces periodic matrix decomposition operations, leading to additional computational overhead and a slight increased overall runtime.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction sections accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are provided in Section 5, and the complete proofs are provided in the Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Code link and experimental details are provided in Section 6 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and code are provided in the code link of Section 6

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix 6 and Section B

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Similar studies do not conduct statistical significance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments are conducted on a machine equipped with a cluster of NVIDIA A100-SXM4-80GB GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our method does not involve any privacy or security concerns.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All the datasets and models used in this paper are publicly available.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the codes, datasets, and models used in this paper are publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: We utilize the LLM only for grammar checking.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.