RETHINKING QUANTIZED LLM FINE-TUNING: AN ADAPTIVE-GROUPING ALTERNATING ZEROTH-ORDER ALGORITHM IN CONTINUOUS MAPPING SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning quantized Large Language Models (LLMs) is crucial for model adaptation on resource-constrained edge devices, where zeroth-order (ZO) optimization has emerged as a memory-friendly alternative to first-order methods. Existing ZO fine-tuning methods typically focus on either perturbing discrete quantized weights or coarse-grained (per-tensor/channel) continuous scales. In this paper, we rethink this fundamental choice and recognize that while fine-tuning should ideally target the dequantized weights, the discrete nature of weights restricts optimization to a finite search space, and tuning coarse-grained scales is limited to a constrained continuous subspace that lacks the expressive capacity of the full parameter space. To mitigate these issues, we propose to break down scales and zero-points to a fine-grained level and fine-tune them in a continuous mapping space. Motivated by this, we propose CoZO, an adaptive-grouping alternating zeroth-order algorithm that optimizes these fine-grained parameters while adaptively balancing search-space expressiveness and estimation stability. Crucially, we provide a theoretical convergence guarantee for CoZO. Extensive experiments demonstrate that CoZO achieves superior stability and effectiveness, consistently outperforming discrete-perturbation and scale-tuning baselines under stringent memory constraints.

1 INTRODUCTION

Large language models (LLMs) (Kojima et al., 2022; Zhang et al., 2022; Touvron et al., 2023b) have demonstrated remarkable performance across a wide range of natural language processing (NLP) tasks, but the large parameter sizes and high computational demands hinder deployment on resource-constrained edge devices (Hu et al., 2022; Dettmers et al., 2021; Bommasani, 2021; Frantar & Alistarh, 2023). Model quantization (Frantar et al., 2022; Shao et al., 2023; Xiao et al., 2023) has emerged as an effective approach by mapping high-precision weights to low-bit formats, reducing storage and inference costs. Practical edge deployment demands not only model compression but also efficient adaptation of quantized models to downstream tasks, spurring increasing interest in fine-tuning quantized LLMs (Zhou et al., 2025; Xu et al., 2023; Jeon et al., 2025). Traditional first-order (FO) fine-tuning methods require storing intermediate activations and gradients during backpropagation, resulting in substantial memory overhead that limits the applicability in resource-constrained environments.

Zeroth-order (ZO) optimization methods have been successfully applied to full-parameter fine-tuning of LLMs as a memory-efficient alternative, approximating gradients via function-value perturbations using only forward evaluations. MeZO (Malladi et al., 2023) introduces ZO-SGD to LLM fine-tuning, enabling ZO optimization at inference-level memory cost. Subsequent work improves stability through variance reduction techniques such as SVRG (Gautam et al., 2024) and enhances optimization efficiency by incorporating low-rank structures (Yu et al., 2024; Chen et al., 2024) or Hessian information (Zhao et al., 2024).

Recent work has begun to explore ZO fine-tuning of quantized models under resource-constrained settings. Given discrete quantized weights, existing ZO fine-tuning methods for quantized models can

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

be broadly categorized into two strategies: discrete perturbation and coarse-grained scale optimization. The discrete perturbation applies ZO updates to quantized weights by designing perturbation schemes and update rules that are compatible with discrete value sets. QuZO (Zhou et al., 2025) introduces low-bit random perturbations combined with quantized ZO updates to directly optimize quantized weights, while ZOQO (Bar & Giryes, 2025) adopts sign-based ZO update rules that naturally align with discrete weight spaces. The coarse-grained scale optimization instead performs ZO updates over continuous per-channel or per-tensor quantization scales, thereby adapting quantized models to downstream tasks via scale parameters (Shang et al., 2025).

However, the discrete nature of quantized weights and the limited expressiveness of coarse-grained quantization scales impose fundamental structural constraints on existing ZO fine-tuning methods. Specifically, applying ZO perturbations directly to discrete quantized weights restricts updates to a finite set, and to ensure gradient unbiasedness, the perturbations must be quantization-constrained, which reduces optimization flexibility and increases estimation variance. In contrast, optimizing only coarse-grained continuous parameters, such as per-tensor or per-channel scales, alleviates discrete optimization difficulties. However, this approach restricts adaptation to a single-scalar (per-tensor/channel) subspace, hindering fine-grained task-specific adjustments. This naturally raises a key question:

Can a ZO fine-tuning method be designed that is compatible with quantized models while preserving both optimization flexibility and parameter-space expressiveness?

To address this issue, we propose CoZO, a fine-grained alternating ZO optimization method for fine-tuning quantized LLMs. Rather than applying perturbations directly to discrete quantized weights, CoZO refines the scales and zero-points in the dequantization process into more fine-grained parameterizations and performs ZO optimization in their corresponding continuous mapping spaces. Our variance analysis shows that, compared with ZO estimators based on discrete weight perturbations, estimators constructed in this continuous parameter space admit more controllable variance bounds, which helps explain the improved stability under quantization constraints. Moreover, CoZO alternately optimizes different dequantization parameters and adaptively adjusts their grouping granularity to balance the expressiveness of the continuous mapping space and the dimension-dependent variance of ZO estimation. We further establish a convergence guarantee for this alternating ZO procedure, showing that under appropriate conditions it converges to a stationary point at a rate of $O(\sqrt{d/T})$.

The main contributions of this work are as follows:

- We reveal a structural limitation in quantized ZO fine-tuning and propose CoZO, which achieves stable and efficient ZO fine-tuning by alternately optimizing fine-grained dequantization parameters in continuous space.
- We provide theoretical guarantees for CoZO by analyzing the variance properties of ZO estimators, characterizing the variance–expressiveness effect of adaptive grouping, and establishing convergence results for the alternating ZO optimization procedure.
- We evaluate CoZO on multiple large language models (OPT, Llama) and benchmark datasets. Under strict memory constraints, CoZO achieves more stable optimization and better task adaptation than ZO methods based on discrete perturbations or scale-only tuning.

2 PRELIMINARY

2.1 MODEL QUANTIZATION

For the full-precision model's weight matrix $W_f \in \mathbb{R}^m$, the process of quantization and dequantization is as follows:

$$W_q = \text{clamp}\left(\text{round}\left(\frac{W_f - \boldsymbol{Z}}{\boldsymbol{S}}\right), 0, 2^{bit} - 1\right), \quad \hat{W_f} = \boldsymbol{S} \odot W_q + \boldsymbol{Z}, \tag{1}$$

where W_q and $\hat{W}_f$ represent the quantized weights and dequantized weights. $\boldsymbol{S}$ denotes the quantized scales, $\boldsymbol{Z}$ denotes the quantized zero point, bit is the quantization bit width, $\text{clamp}(\cdot)$ is the clipping

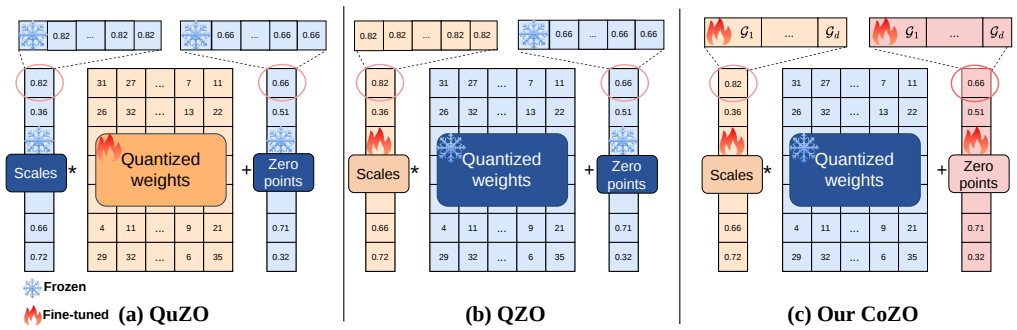


Figure 1: Overview of the CoZO framework. In contrast to QuZO and QZO, CoZO keeps the quantized weights fixed and incorporates the zero-point into the optimization, jointly grouping it with the scale in a fine-grained manner. This design enables stable and efficient fine-tuning under an alternating ZO optimization scheme.

function, $\mathrm{round}(\cdot)$ is the rounding function, and $\odot$ represents the element-wise product. Given a loss function $\mathcal{L}(\cdot)$, the optimization objective for quantized model fine-tuning can be formulated as:

$$\min_{\boldsymbol{S},\boldsymbol{Z},W_q} \mathcal{L}(\boldsymbol{S} \odot W_q + \boldsymbol{Z}). \tag{2}$$

Because $\hat{W}_f$ is not directly optimizable, the fine-tuning procedure typically adopts W_q or the coarse-grained $\boldsymbol{S}$ as the optimization variable.

2.2 Zeroth-Order Optimization

Zeroth-order optimizers estimate gradient information using only function evaluations. Among them, the well-established Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm serves as a representative ZO method and is defined as follows:

Definition 2.1 (Simultaneous Perturbation Stochastic Approximation (Spall, 2002)). Given the model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and the loss $\mathcal{L}(\cdot)$, the gradient estimated by SPSA under batch $\mathcal{B}$ is:

$$\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) = \frac{\mathcal{L}(\boldsymbol{\theta}+\epsilon u;\mathcal{B}) - \mathcal{L}(\boldsymbol{\theta}-\epsilon u;\mathcal{B})}{2\epsilon} u \approx uu^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}), \tag{3}$$

where ϵ is the perturbation scale, and u is a random variable sampled from the standard Gaussian distribution. The n-SPSA gradient estimate averages $\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})$ over n randomly sampled u.

To update discrete weights, QuZO (Zhou et al., 2025) designs a quantized ZO estimation strategy to maintain unbiasedness (Definition 2.2), along with a discrete update method for low-bit settings.

Definition 2.2 (Quantized RGE (Zhou et al., 2025)). Given model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and a loss function $\mathcal{L}(\cdot)$, Quantized RGE estimates the gradient on mini-batch $\mathcal{B}$ as:

$$\bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) = \frac{\mathcal{L}(\boldsymbol{\theta}+\epsilon u_1;\mathcal{B}) - \mathcal{L}(\boldsymbol{\theta}-\epsilon u_1;\mathcal{B})}{2\epsilon} u_2 \approx u_2 u_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}), \tag{4}$$

where ϵ is the perturbation scale, and the random variables u_1 and u_2 are generated as follows:

$$\begin{aligned} &u_1 = Q_1(u) = Q(u) \text{ with random seed } i_1, u_2 = Q_2(u) = Q(u) \text{ with random seed } i_2, \\ &Q(u_i) = \mathrm{clamp}(SQ(u),\, M_{\min},\, M_{\max}) + z_0, SQ(u) = (\lfloor s_u u \rfloor + \mathrm{Ber}(s_u u - \lfloor s_u u \rfloor)), \end{aligned} \tag{5}$$

where $Q(u)$ is a quantization function, $SQ(u)$ is a stochastic quantization process, and $\mathrm{Ber}(p)$ represents a Bernoulli distribution with probability p.

Given parameters $\boldsymbol{\theta}$ and a learning rate η, the update rule of Quantized ZO-SGD at step t is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - Q\left(\eta \bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}_t)\right), \tag{6}$$

where $\mathcal{B}_t$ represents the mini-batch data, and $\bar{\nabla}\mathcal{L}$ is the Quantized RGE gradient estimate.

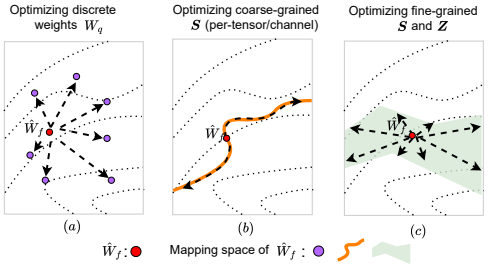


Figure 2: Mapping spaces of $\hat{W}_f$ under different optimization strategies. Optimizing discrete weights W_q restricts $\hat{W}_f$ to a discrete space, coarse-grained $\boldsymbol{S}$ enables search in a constrained continuous mapping space, while fine-grained $\boldsymbol{S}$ and $\boldsymbol{Z}$ enable a more flexible continuous mapping space.

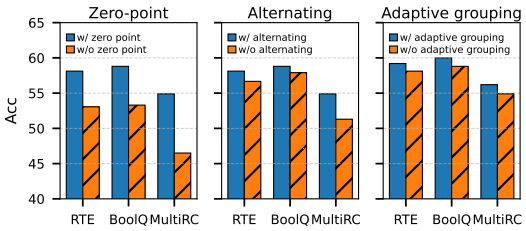


Figure 3: Ablation study on RTE, BoolQ, and MultiRC using OPT-1.3B (5k steps). We evaluate the effects of zero-point updates, alternating optimization, and adaptive grouping under a unified setting. Results show that removing any component degrades performance, while the full method consistently achieves the best results across datasets.

3 Methodology

3.1 Rethinking Fine-Tuning of Quantized LLMs

Mapping Space. To understand the optimization behavior during quantized fine-tuning, we first characterize the expressive space of the dequantized weight $\hat{W}_f$. As shown in Eq. (2), $\hat{W}_f$ is not a variable that can be directly optimized. Instead, it must be realized indirectly through the optimization of discrete weights W_q and dequantization parameters $\boldsymbol{S}$ and $\boldsymbol{Z}$. Corresponding to the two types of optimizable variables, the dequantized weights $\hat{W}_f$ are mapped into two types of mapping spaces:

Definition 3.1 (**Discrete mapping space**). Given fixed $\boldsymbol{S}$ and $\boldsymbol{Z}$, the following set is referred to as the discrete mapping space of $\hat{W}_f$ induced by $\boldsymbol{S}$ and $\boldsymbol{Z}$:

$$\mathcal{R}_d = \{\boldsymbol{S} \odot W_q + \boldsymbol{Z} \mid W_q \in \mathbb{Z}^m\}. \tag{7}$$

Definition 3.2 (**Full continuous mapping space**). Given a fixed W_q, the following set is referred to as the full continuous mapping space of $\hat{W}_f$ by W_q:

$$\mathcal{R}_c = \{\boldsymbol{S} \odot W_q + \boldsymbol{Z} \mid \boldsymbol{S}, \boldsymbol{Z} \in \mathbb{R}^m\}. \tag{8}$$

It should be noted that this space is equivalent to the one obtained by directly optimizing $\hat{W}_f$:

$$\mathcal{R}_c = \mathcal{R}_f = \left\{\hat{W}_f = \boldsymbol{S} \odot W_q + \boldsymbol{Z} \;\middle|\; \hat{W}_f \in \mathbb{R}^m\right\}. \tag{9}$$

Expressiveness Limitation. Figure 2 provides an intuitive illustration of the expressiveness differences induced by different optimization strategies. (i). For discrete optimization, existing methods (Zhou et al., 2025; Bar & Giryes, 2025) typically operate directly on the quantized weights, due to the finite bit width, the quantized weights W_q are restricted to a bounded integer set, e.g., $W_q \in \{q_{\min}, \ldots, q_{\max}\}^m$. As a result, although the discrete mapping space $\mathcal{R}_d$ is formally defined over $W_q \in \mathbb{Z}^m$, practical fine-tuning can only traverse a finite subset of $\mathcal{R}_d$ induced by the admissible quantization levels, as illustrated in Figure 2 (a). (ii). In contrast, for continuous optimization, existing methods (Shang et al., 2025; Kim et al., 2023) typically perform fine-tuning by adjusting the dequantization scale parameter $\boldsymbol{S}$. However, common quantization schemes, such as per-tensor/channel quantization, impose a structured shared parameterization on $\boldsymbol{S}$. For example, per-channel quantization enforces a shared scale within each channel (Figure 1 (b)), which restricts the reachable set of $\hat{W}_f$ to a single-scalar subset of the continuous mapping space $\mathcal{R}_c$. Consequently, much of $\mathcal{R}_c$ remains unexplored, as illustrated in Figure 2 (b).

Notably, the dequantization parameters $\boldsymbol{S}$ and $\boldsymbol{Z}$ are not constrained to follow the same scheme (e.g., per-tensor/channel) used during quantization. Consequently, allowing a finer-grained parameterization of $\boldsymbol{S}$ and $\boldsymbol{Z}$, as exemplified by the group-wise dequantization parameters in Figure 1 (c), enables

broader exploration of the continuous mapping space, thereby allowing the expressive capacity of $\hat{W}_f$ to progressively approach that of directly optimizing $\hat{W}_f$, as illustrated in Figure 2 (c).

Stability Limitation. From an optimization perspective, differences in mapping space exploration also impact the stability of ZO optimization. Discrete-perturbation-based ZO methods (Zhou et al., 2025) require quantization operations in gradient estimation, which introduces additional noise compared to continuous perturbations. Theorem 3.3 and Theorem 3.4 provide corresponding upper bounds on the variance of the gradient estimators, with proofs in Appendix D.

Theorem 3.3 (**Variance bound of SPSA**). *For the gradient estimated by Definition 2.1, the variance of the gradient estimator admits the following upper bound:*

$$\mathrm{Var}\big(\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big) \leq (d+2)\left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2. \tag{10}$$

Theorem 3.4 (**Variance bound of Quantized RGE**). *For the gradient estimated by Definition 2.2, suppose that there exist quantization errors e_i such that $u_i = u + e_i, i \in \{1,2\}$, where $u \sim \mathcal{N}(0, I_d)$. Assume that the quantization errors satisfy $\mathbb{E}\big[\|e_i\|^2\big] \leq \sigma_i^2, i \in \{1,2\}$. Then, the variance of the gradient estimator admits the following upper bound:*

$$\mathrm{Var}\big(\bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big) \leq 4\left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2\Big((d+2) + d\sigma_1^2 + \sigma_2^2 + \sigma_1^2\sigma_2^2\Big). \tag{11}$$

Remark 3.5. Compared with SPSA (Definition 2.1), the variance bound of Quantized RGE (Definition 2.2) includes extra terms induced by quantization errors. As a result, discrete optimization in high-dimensional parameter spaces may suffer from compounded instability, stemming from both the dimensionality of the optimization problem and the presence of quantization noise.

Motivation. In summary, existing quantized fine-tuning methods are limited by a constrained mapping space and unstable ZO updates under low-bit quantization. This motivates a finer-grained parameterization of the dequantization parameters, together with a continuous ZO optimization scheme that controls the variance upper bound to improve stability.

3.2 Continuous Space Zeroth-Order Fine-Tuning

Method Overview. In this subsection, we present CoZO, which parameterizes the dequantization scales and zero points through a grouping mechanism and optimizes them via alternating ZO updates. The proposed method is applicable to arbitrary quantization settings. For symmetric quantization, the zero point can be incorporated during the dequantization stage. In addition, CoZO enables efficient adaptation to different downstream tasks by switching the dequantization parameters, making it well suited for resource-constrained scenarios. The overall framework is summarized in Figure 1 (c).

Grouped Dequantization Parameterization. Let $\boldsymbol{\theta} \in \mathbb{R}^m$ denote the original dequantization parameters of the quantized model, treated as having the same dimensionality as the weights before grouping. We introduce a grouping operator $\mathcal{G}(\cdot, d)$ that partitions the dequantization parameter vector $\boldsymbol{\theta}$ into d disjoint groups. Let $\mathcal{G} : \mathbb{R}^m \to \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_d\}$ be a mapping such that each group $\mathcal{G}_i \subseteq \mathcal{G}(\boldsymbol{\theta}, d)$ consists of consecutive elements. Specifically, the size of each group is defined by:

$$|\mathcal{G}_i| = \lfloor m/d \rfloor \quad \text{for} \quad i = 1, 2, \ldots, d-1, \quad |\mathcal{G}_d| = m - (d-1)\cdot\lfloor m/d \rfloor. \tag{12}$$

All elements in each group share one parameter, i.e., $\theta_j = \theta_k$ for $\theta_j, \theta_k \in \mathcal{G}_i$. In particular, when $d = m$, this is equivalent to element-wise dequantization.

Alternating ZO Optimization. For the grouped scale and zero-point parameters, we adopt a ZO optimization scheme, use SPSA to estimate the gradients, and update the two alternately. However, Theorem 3.3 reveals an inherent trade-off: finer grouping enlarges the effective optimization space but also amplifies the zeroth-order estimation variance. To circumvent this limitation, we propose to dynamically adjust the grouping throughout optimization.

Adaptive Grouping. Our adaptive grouping scheme consists of three operations: splitting, merging, and maintaining. For the current number of groups d, the splitting operation expands the group count to $2d$ by replicating the parameters of each original group, while the merging operation reduces the group count to $d/2$ by averaging the parameters of each pair of adjacent groups. The maintaining operation keeps the current grouping structure unchanged.

To guide the choice among splitting, merging, and maintaining, we use the variability of the SPSA estimator as an indicator of the current grouping quality. Let $g_{\boldsymbol{\theta}}$ denote the scalar finite-difference

Algorithm 1 Continuous Space Zeroth-Order Fine-Tuning for Quantized LLMs (CoZO)

Input: quantization scales $\boldsymbol{S}_{\text{Init}} \in \mathbb{R}^m$, quantization zero points $\boldsymbol{Z}_{\text{Init}} \in \mathbb{R}^m$, quantized weights $W_q \in \mathbb{R}^m$, loss function $\mathcal{L}(\cdot)$, grouping operator $\mathcal{G}(\cdot)$, learning rates $\{\eta_{1t}, \eta_{2t}\}$, optimization steps T, perturbation scales $\{\epsilon_1, \epsilon_2\}$, initial number of groups d_0, adaptation interval K, quantile threshold p
Initialization: $d \leftarrow d_0$, $\boldsymbol{S} \leftarrow \mathcal{G}(\boldsymbol{S}_{\text{Init}}; d)$, $\boldsymbol{Z} \leftarrow \mathcal{G}(\boldsymbol{Z}_{\text{Init}}; d)$, $\mathcal{H}_g \leftarrow \emptyset$
for $t = 1, 2, \ldots, T$ **do**
  Sample a batch of inputs $\mathcal{B}$ and random seeds $\{seed_1, seed_2\}$
  $\boldsymbol{S}, g_S \leftarrow$ ZO_UPDATE$(\boldsymbol{S}, \epsilon_1, seed_1, \eta_{1t}, \boldsymbol{Z}, W_q, \mathcal{B}, \mathcal{L})$
  $\boldsymbol{Z}, g_Z \leftarrow$ ZO_UPDATE$(\boldsymbol{Z}, \epsilon_2, seed_2, \eta_{2t}, \boldsymbol{S}, W_q, \mathcal{B}, \mathcal{L})$
  $h_t \leftarrow \frac{1}{2}(|g_{\boldsymbol{S}}| + |g_{\boldsymbol{Z}}|)$, $\mathcal{H}_g \leftarrow \text{Append}(\mathcal{H}_g, h_t)$
  if $t \bmod K = 0$ **then**
    $\bar{h}_t \leftarrow \text{Mean}(\mathcal{H}_g[:K])$
    $q_{\text{low}} \leftarrow \text{Quantile}(\mathcal{H}_g, 1-p)$, $q_{\text{high}} \leftarrow \text{Quantile}(\mathcal{H}_g, p)$
    if $\bar{h}_t > q_{\text{high}}$ **then**
      $\boldsymbol{S}, \boldsymbol{Z}, d \leftarrow$ MERGE$(\boldsymbol{S}, \boldsymbol{Z}, d)$
    else if $\bar{h}_t < q_{\text{low}}$ **then**
      $\boldsymbol{S}, \boldsymbol{Z}, d \leftarrow$ SPLIT$(\boldsymbol{S}, \boldsymbol{Z}, d)$
    else
      $\boldsymbol{S}, \boldsymbol{Z}, d \leftarrow$ MAINTAIN$(\boldsymbol{S}, \boldsymbol{Z}, d)$
    end if
    $\mathcal{H}_g \leftarrow \emptyset$
  end if
end for

term in Definition 2.1, which satisfies $g_{\boldsymbol{\theta}} \approx u^\top \nabla \mathcal{L}(\boldsymbol{\theta}; \mathcal{B})$ (Nesterov & Spokoiny, 2017). Its second moment approximates the squared gradient norm, i.e., $\mathbb{E}[g_{\boldsymbol{\theta}}^2] \approx \|\nabla \mathcal{L}(\boldsymbol{\theta}; \mathcal{B})\|^2$. Substituting this relation into Eq. (10) gives

$$\text{Var}\big(\hat{\nabla} \mathcal{L}(\boldsymbol{\theta}; \mathcal{B})\big) \leq (d+2)\mathbb{E}[g_{\boldsymbol{\theta}}^2]. \tag{13}$$

Therefore, $g_{\boldsymbol{\theta}}^2$ serves as a practical proxy for the variability of the SPSA gradient estimator. Equation (13) also makes the variance–expressiveness trade-off explicit. Based on this observation, we maintain a historical list $\mathcal{H}_g$ to record $|g_{\boldsymbol{\theta}}|$ at each iteration. Every k steps, we compare the mean of $\mathcal{H}_g$ with its p- and $(1-p)$-quantiles, where $p > 0.5$. If the mean exceeds the p-quantile, CoZO merges groups to reduce the search dimensionality and stabilize optimization. Conversely, if the mean falls below the $(1-p)$-quantile, the optimization is considered sufficiently stable, and a splitting operation is performed to expand the optimization space and further improve the expressiveness of the grouping structure. If the mean lies between the two quantiles, the current grouping structure is maintained. The main optimization loop is shown in Algorithm 1, where the detailed implementation of ZO_UPDATE is deferred to Algorithm 2 in Appendix B.

4 CONVERGENCE

4.1 ASSUMPTIONS

Assumption 4.1 (Lipschitz condition). Assume that the gradient of the loss function $\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})$ is Lipschitz continuous with respect to $\boldsymbol{S}$ and $\boldsymbol{Z}$, with constants $L_{\boldsymbol{S}}$ and $L_{\boldsymbol{Z}}$.

Assumption 4.2 (Local r-effective rank). Let $G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t) = \max_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \|\nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \{(\boldsymbol{x}, \boldsymbol{y})\})\|$, $G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) = \max_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \|\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \{(\boldsymbol{x}, \boldsymbol{y})\})\|$.

Assume there exist matrices $\boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t) \preceq L_{\boldsymbol{S}} \boldsymbol{I}_d$ and $\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) \preceq L_{\boldsymbol{Z}} \boldsymbol{I}_d$ such that, for all $\|\boldsymbol{S} - \boldsymbol{S}_t\| \leq \eta d\, G_1, \|\boldsymbol{Z} - \boldsymbol{Z}_t\| \leq \eta d\, G_2$:

$$\nabla_{\boldsymbol{S}}^2 \mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}) \preceq \boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t), \quad \nabla_{\boldsymbol{Z}}^2 \mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}) \preceq \boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t).$$

The effective ranks of $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ satisfy $\frac{\text{tr}(\boldsymbol{H}_i(\cdot))}{\|\boldsymbol{H}_i(\cdot)\|_{\text{op}}} \leq r, i = 1, 2$.

Definition 4.3. The covariance of the mini-batch stochastic gradient with respect to $\boldsymbol{S}$ is defined as:

$$\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}) = b\Big(\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\Big] - \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top\Big),$$

with $\Sigma_{\boldsymbol{Z}}$ defined analogously.

Assumption 4.4. Assume that the maximum eigenvalues of $\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z})$ and $\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z})$ are bounded by $\lambda_{\boldsymbol{S}}$ and $\lambda_{\boldsymbol{Z}}$.

Remark 4.5. The convergence analysis is conducted under weight quantization, where scale and zero-point optimization avoids function discontinuity, making Assumption 4.1 reasonable. Moreover, since grouping reduces the parameter dimension to d, which is much smaller than the weight dimension m, the effective rank r is comparable to d in practice.

4.2 Convergence Result

This subsection establishes the convergence guarantees of Algorithm 1 under adaptive grouping. We first derive a descent lemma for alternating ZO updates with fixed grouping under spherical perturbations (Lemma 4.6). By accounting for the loss change induced by regrouping, we establish convergence for the adaptive procedure (Theorem 4.7). We then extend the analysis to Gaussian perturbations under refined assumptions (Theorem 4.8). Detailed proofs are deferred to Appendix D.

Lemma 4.6 (Descent Lemma). *Under Assumption 4.1 and Assumption 4.2, consider one iteration with the grouping structure fixed, and let $(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1})$ denote the parameters after the alternating ZO update and before regrouping. If $\boldsymbol{S}$ and $\boldsymbol{Z}$ are updated using ZO-SGD with the n-SPSA estimator and a minibatch of size b, then there exist $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$ such that*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\widetilde{\boldsymbol{S}}_{t+1},\widetilde{\boldsymbol{Z}}_{t+1}) - \mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right] &\le -\eta(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 + \|\nabla_{\boldsymbol{Z}}\mathcal{L}(\widetilde{\boldsymbol{S}}_{t+1},\boldsymbol{Z}_t)\|^2)\\ + \frac{\eta^2}{2}(L_{\boldsymbol{S}}\gamma_1\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\right] &+ L_{\boldsymbol{Z}}\gamma_2\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\widetilde{\boldsymbol{S}}_{t+1},\boldsymbol{Z}_t;\mathcal{B})\|^2\right]).\end{aligned} \tag{14}$$

Let d_t denote the number of active groups at iteration t, with $d_{\min} \le d_t \le d_{\max}$. We define the cumulative positive loss change caused by regrouping as

$$\mathcal{E}_T := \sum_{t=0}^{T-1}\mathbb{E}[\varepsilon_t], \qquad \varepsilon_t := \left[\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1}) - \mathcal{L}(\widetilde{\boldsymbol{S}}_{t+1},\widetilde{\boldsymbol{Z}}_{t+1})\right]_+. \tag{15}$$

Theorem 4.7. *Suppose Assumptions 4.1–4.4 and the uniform bounded-regrouping conditions in Appendix D hold. If $\boldsymbol{S}$ and $\boldsymbol{Z}$ are updated using ZO-SGD with the n-SPSA estimator and a minibatch of size b, and the learning rate is chosen as $\eta = O(1/\sqrt{Td_{\max}})$ and $K = \Omega(T)$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le O\left((\Delta_{\mathcal{L}}+1)\sqrt{\frac{d_{\max}}{T}} + \frac{1}{Tb} + \frac{\sqrt{d_{\max}}}{\sqrt{T}b}\right), \tag{16}$$

where $\Delta_{\mathcal{L}} := \mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0) - \mathcal{L}^*$.

Theorem 4.8. *Under Gaussian perturbations and the corresponding conditions in Appendix D, assumed uniformly over the grouping structures encountered by Algorithm 1, choosing $\eta = O(1/\sqrt{Td_{\max}})$ and $K = \Omega(T)$ gives*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le O\left((\Delta_{\mathcal{L}}+1)\sqrt{\frac{d_{\max}}{T}} + \frac{1}{Tb} + \frac{\sqrt{d_{\max}}}{\sqrt{T}b} + \frac{\exp(-\Omega(nd_{\min}))}{\sqrt{Td_{\max}}}\right). \tag{17}$$

Remark 4.9. The exponentially decaying term arises from controlling the tail behavior of Gaussian perturbations and becomes negligible for moderate n and $d_{\min}$.

5 Experiments

5.1 Setup

Datasets and Model Configurations. We evaluate model performance on a set of representative downstream benchmarks, including SST-2 (Socher et al., 2013), RTE (Dagan et al., 2005),

Table 1: Experiments on OPT-6.7B and Llama-2-7B under weight-only quantization. We report accuracy and peak GPU memory usage on downstream tasks. **Bold** indicates the best result among quantized models, and underlined indicates the second-best.

Model	Method	Order	Precision	Memory	SST-2	RTE	BoolQ	MultiRC	WIC	WSC	COPA	SQuAD
OPT-6.7B	Zero-shot	–	16 bits	–	61.24	54.87	57.30	41.70	51.25	37.50	82.00	40.62
	Fine-tuning	FO	16 bits	107.44 GB	93.12	53.43	60.10	58.90	50.00	46.15	81.00	85.97
	MeZO	ZO	16 bits	14.84 GB	94.04	65.34	68.40	61.10	61.76	56.73	79.00	77.31
	Zero-shot-Q	–	4 bits	–	61.58	56.68	58.90	41.40	50.78	50.96	81.00	49.90
	FT-Q	FO	4 bits	107.44 GB	86.47	48.74	60.10	55.70	50.47	43.27	70.00	**84.10**
	ZO-Q	ZO	4 bits	14.84 GB	84.63	55.23	52.90	58.30	52.35	55.77	70.00	72.68
	QZO	ZO	4 bits	14.00 GB	80.62	57.40	61.90	58.70	57.21	50.96	81.00	70.56
	QuZO	ZO	4 bits	33.49 GB	85.21	51.26	61.90	55.20	**58.31**	57.69	**82.00**	73.88
	CoZO (ours)	ZO	4 bits	14.00 GB	**94.38**	**64.26**	**63.60**	**60.50**	**58.31**	**62.50**	**82.00**	79.03
Llama-2-7B	Zero-shot	–	16 bits	–	58.03	62.09	66.00	42.30	50.16	36.54	79.00	60.39
	Fine-tuning	FO	16 bits	114.11 GB	92.89	80.51	77.70	49.80	64.26	63.46	75.00	89.22
	MeZO	ZO	16 bits	14.95 GB	87.04	61.37	71.10	66.80	57.37	61.54	82.00	81.97
	Zero-shot-Q	–	4 bits	–	57.34	54.51	57.90	59.20	50.78	53.85	79.00	53.60
	FT-Q	FO	4 bits	114.11 GB	**93.12**	58.48	52.00	58.80	53.76	55.77	73.00	73.88
	ZO-Q	ZO	4 bits	14.95 GB	77.87	61.73	58.10	59.20	53.76	54.81	**81.00**	69.42
	QZO	ZO	4 bits	13.98 GB	85.67	61.01	66.00	59.60	53.29	53.85	76.00	75.11
	QuZO	ZO	4 bits	31.45 GB	90.94	58.12	66.00	58.90	51.41	55.77	80.00	73.07
	CoZO (ours)	ZO	4 bits	14.01 GB	87.73	**63.17**	**66.30**	**62.10**	**58.78**	**58.65**	**81.00**	**75.87**

BoolQ (Clark et al., 2019), MultiRC (Khashabi et al., 2018), WIC (Pilehvar & Camacho-Collados, 2019), WSC (Levesque et al., 2012), COPA (Roemmele et al., 2011) and SQuAD (Rajpurkar et al., 2016). To validate the effectiveness of CoZO, we first conduct experiments on OPT-6.7B (Zhang et al., 2022) and Llama-2-7B (Touvron et al., 2023a). We subsequently extend our evaluation to a larger model (OPT-13B) to demonstrate the scalability of the proposed approach.

Baselines. We compare our proposed method against a range of established baselines. Under the full-precision setting, the compared methods include full-precision zero-shot evaluation (Zero-shot), MeZO (Malladi et al., 2023), and full fine-tuning with backpropagation (FT) (Malladi et al., 2023). Under the quantized setting, we consider direct quantization (Zero-shot-Q) (Shao et al., 2023), quantization after FT (FT-Q), quantization after MeZO (ZO-Q), QZO (Shang et al., 2025), and QuZO (Zhou et al., 2025).

5.2 Main Results

In this subsection, we report the memory footprint and accuracy of OPT-6.7B and Llama-2-7B under the weight-only quantization setting. Following standard experimental protocols, we randomly sample 1,000 examples for training and another 1,000 examples for evaluation. With the same batch size, we also report the peak GPU memory usage of each method on the SST-2 dataset. The detailed results are reported in Table 1. In addition, we conduct supplementary experiments on the 13B model to further evaluate the scalability of our method, with results summarized in Table 2. Overall, CoZO achieves competitive or improved accuracy compared to prior baselines across most settings, while remaining stable as the model scale increases. Meanwhile, compared to methods that optimize only the scale parameters, CoZO does not introduce a significant increase in GPU memory consumption. These results indicate that the proposed alternating ZO optimization framework in a continuous mapping space achieves superior overall performance compared to discrete-perturbation methods and coarse-grained continuous scale (per-channel) optimization approaches.

5.3 Optimization Dynamics of Adaptive Grouping

In this subsection, we further investigate the optimization behavior of adaptive grouping. Figure 4 compares the training loss of adaptive and fixed grouping on SST-2 using OPT-1.3B, where both settings start from two groups. Compared with fixed grouping, adaptive grouping exhibits faster loss reduction throughout training and converges to a lower loss. These results suggest that dynamically adjusting the grouping granularity enables CoZO to better balance optimization stability and parameter-space expressiveness during fine-tuning.

Table 2: Experimental results on OPT-13B under the 4-bit quantization setting.

Method	Order	SST-2	RTE	BoolQ	COPA
Zero-shot	-	58.80	59.60	59.00	80.00
Fine-tuning	FO	92.00	70.80	77.10	79.00
MeZO	ZO	91.40	66.10	67.60	88.00
Zero-shot-Q	-	59.17	52.30	57.90	75.00
FT-Q	FO	53.10	46.93	60.10	64.00
ZO-Q	ZO	64.11	54.51	59.00	57.00
QZO	ZO	86.58	53.42	59.70	80.00
QuZO	ZO	72.71	49.10	59.00	48.00
CoZO (ours)	ZO	**92.89**	**59.21**	**63.40**	**81.00**

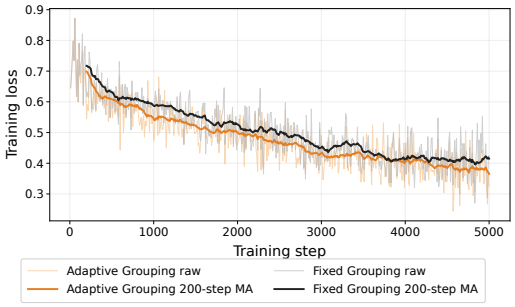


Figure 4: Training loss comparison between fixed and adaptive grouping on OPT-1.3B over SST-2.

5.4 Ablation Study

We conduct a comprehensive suite of ablation studies to systematically assess the overall effectiveness of the proposed method. We report accuracy across multiple datasets, and all experiments are conducted on OPT-1.3B with the number of training steps fixed to 5k to ensure a consistent and fair comparison. The ablations examine several key components of our design, including: (i). **CoZO w/o Zero-Point**: In the non-grouped setting, the zero-point update is frozen, and only the scale is updated. (ii). **CoZO w/o Alternating**: In the non-grouped setting, both the zero-point and scale are updated simultaneously. (iii). **CoZO w/o Adaptive Grouping**: In the alternating optimization setting, we disable adaptive grouping by setting the parameter p to 1, thereby keeping the grouping fixed during optimization. Figure 3 shows that the full CoZO consistently outperforms all ablated variants.

6 Related Work

LLM Quantization and Fine-Tuning. LLM quantization reduces storage and inference costs through low-bit representations, including post-training quantization (PTQ) (Shao et al., 2023; Xiao et al., 2023; Frantar et al., 2022) and quantization-aware training (QAT) (Jacob et al., 2018; Chen et al., 2025). Recent methods further improve low-bit robustness through adaptive scaling, grouping, or outlier-aware designs (Xiao et al., 2023; Dettmers et al., 2022; Wu et al., 2023). For downstream adaptation, parameter-efficient methods such as QA-LoRA (Xu et al., 2024) and quantization-aware variants (Xu et al., 2023) reduce trainable parameters or preserve low-bit structures. In contrast, our work focuses on full-parameter adaptation of quantized models through memory-efficient zeroth-order optimization.

Zeroth-Order Fine-Tuning. MeZO (Malladi et al., 2023) enables LLM fine-tuning using only forward evaluations, and subsequent work improves ZO optimization through sparsity (Liu et al., 2024; Guo et al., 2024), variance reduction (Gautam et al., 2024), low-rank structures (Yu et al., 2024; Chen et al., 2024), and curvature information (Zhao et al., 2024). For quantized models, QuZO (Zhou et al., 2025) and ZOQO (Bar & Giryes, 2025) directly optimize discrete quantized weights, whereas QZO (Shang et al., 2025) optimizes continuous but coarse-grained scales. CoZO differs by optimizing fine-grained scale and zero-point parameters in continuous mapping spaces with adaptive grouping.

7 Conclusion

In this paper, we rethink fine-tuning strategies for quantized LLMs and identify expressiveness limitations in both perturbing discrete quantized weights and tuning coarse-grained continuous quantization scales. To mitigate these issues, we propose CoZO, which breaks down scales and zero-points to a fine-grained level via a grouping strategy and fine-tunes them in a continuous mapping space using an alternating ZO algorithm with adaptive grouping, thereby balancing mapping-space expressiveness and estimation stability. We establish a theoretical framework to characterize the convergence behavior of CoZO, explaining why it enables efficient fine-tuning. Extensive experiments across multiple LLMs and downstream tasks show that CoZO generally outperforms existing ZO methods in terms of accuracy and memory footprint. Moreover, our results validate the scalability of CoZO to larger model sizes.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and structural refinement of the manuscript. All AI-assisted content was reviewed and verified by the authors. The authors take responsibility for the final mathematical claims, experimental results, citations, code, and manuscript content.

ETHICS STATEMENT

All experiments use publicly available benchmarks and comply with their licenses. This work does not involve new human-subject data collection, human-subject research, deployment in high-stakes decision settings, user-facing decision systems, or release of sensitive data. We have considered potential ethical risks and do not identify risks specific to this theoretical and diagnostic analysis, beyond the general computational and environmental costs of model training.

REPRODUCIBILITY STATEMENT

Section 3 details the CoZO algorithm, including grouped dequantization, alternating ZO optimization, and adaptive grouping, with additional algorithmic details provided in Appendix B. The theoretical assumptions and convergence guarantees are presented in Section 4, with detailed proofs and variance analysis in Appendix D. Section 5 describes the models, datasets, baselines, and evaluation protocols. Appendix C provides hardware configurations, implementation details, hyperparameter settings, and additional experimental results, including sensitivity analyses, to facilitate reproducibility.

REFERENCES

Noga Bar and Raja Giryes. Zoqo: Zero-order quantized optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. Efficientqat: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10081–10100, 2025.

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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

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.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

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.

Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, et al. Zeroth-order fine-tuning of llms with extreme sparsity. *arXiv preprint arXiv:2406.02913*, 2024.

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.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Hyesung Jeon, Seojune Lee, Beomseok Kang, Yulhwa Kim, and Jae-Joon Kim. Qwha: Quantization-aware walsh-hadamard adaptation for parameter-efficient fine-tuning on large language models. *arXiv preprint arXiv:2509.17428*, 2025.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 252–262, 2018.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36:36187–36207, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. *KR*, 2012(13th):3, 2012.

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.

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.

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

Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (Long and short papers)*, pp. 1267–1273, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Sifeng Shang, Jiayi Zhou, Chenyu Lin, Minxian Li, and Kaiyang Zhou. Fine-tuning quantized neural networks with zeroth-order optimization. *arXiv preprint arXiv:2505.13430*, 2025.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 2002.

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*, 2023a.

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*, 2023b.

Xiaoxia Wu, Zhewei Yao, and Yuxiong He. Zeroquant-fp: A leap forward in llms post-training w4a8 quantization using floating-point formats. *arXiv preprint arXiv:2307.09782*, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717*, 2023.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 52401–52418, 2024.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, and Hua Huang. Subzero: Random subspace zeroth-order optimization for memory-efficient llm fine-tuning. 2024.

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.

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.

Jiajun Zhou, Yifan Yang, Kai Zhen, Ziyue Liu, Yequan Zhao, Ershad Banijamali, Athanasios Mouchtaris, Ngai Wong, and Zheng Zhang. Quzo: Quantized zeroth-order fine-tuning for large language models. *arXiv preprint arXiv:2502.12346*, 2025.

A Definition of Notation

Table 3: Key Notation.

Notation	Description
$\mathcal{L}(\cdot)$	Objective function
b	Batch size
$\mathcal{B}$	A random minibatch of size b
$\hat{\nabla}\mathcal{L}(\cdot;\mathcal{B})$	SPSA gradient estimation for $\mathcal{L}$
$\bar{\nabla}\mathcal{L}(\cdot;\mathcal{B})$	Quantized RGE gradient estimation for $\mathcal{L}$
$\nabla\mathcal{L}(\cdot;\mathcal{B})$	SGD gradient estimation for $\mathcal{L}$
$\nabla\mathcal{L}(\cdot)$	True gradient for $\mathcal{L}$
η	Leaning rate
$L_{\max}$	$\max\{L_{\boldsymbol{S}}, L_{\boldsymbol{Z}}\}$
$L'_{\max}$	$\max\{\gamma_1 L_{\boldsymbol{S}}, \gamma_2 L_{\boldsymbol{Z}}\}$
$\Sigma(\boldsymbol{S}, \boldsymbol{Z})$	$b\left(\mathbb{E}\left[\nabla\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\nabla\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})^\top\right] - \nabla\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})\nabla\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})^\top\right)$
$\Sigma_{\boldsymbol{S}}(\boldsymbol{S}, \boldsymbol{Z})$	$b\Big(\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})^\top\Big] - \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})^\top\Big)$
$\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}, \boldsymbol{Z})$	$b\Big(\mathbb{E}\Big[\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})^\top\Big] - \nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})^\top\Big)$
$\lambda_{\max}$	$\max\{\lambda_{\boldsymbol{S}}, \lambda_{\boldsymbol{Z}}\}$
n	The number of random vectors u sampled in n-SPSA
d	The dimension of parameters after grouping
d_t	The number of groups active at iteration t
$d_{\min}, d_{\max}$	Minimum and maximum admissible group counts
m	Dimension of the weights
$\mathcal{G}(\cdot)$	Grouping operator
$\mathcal{G}_t$	Grouping structure active at iteration t
$\mathcal{L}_t$	Grouped objective induced by $\mathcal{G}_t$
ε_t	Positive loss increase caused by regrouping at iteration t
$\mathcal{E}_T$	Cumulative expected regrouping error $\sum_{t=0}^{T-1}\mathbb{E}[\varepsilon_t]$

B Additional Algorithm Details

In this section, we provide the detailed subroutine of ZO_Update, as shown in Algorithm 2. This subroutine is used in Algorithm 1 and performs an SPSA-based ZO update. It returns both the updated parameter and the corresponding difference term used for adaptive grouping.

C Additional Experimental Details

C.1 Implementation Details

We run all experiments on a cluster of NVIDIA A100 GPUs (80GB). We adopt OmniQuant as our base quantization method and initialize from its officially released checkpoints. To ensure a fair comparison, all baselines were re-implemented or executed using their officially released code under the same environment and, whenever possible, matched hyperparameter settings. For optimization settings, we use 20k steps for QZO, while CoZO and QuZO use 5k or 8k steps. Since QZO has a lower-dimensional optimization space, it empirically requires more iterations to achieve sufficient optimization. Using the same short training budget as the other methods may therefore under-train QZO, as shown by the OPT-1.3B results in Table 4. Moreover, setting the optimization steps of QZO to 20k is also consistent with the experimental configuration in its original paper. Unless otherwise noted, CoZO initializes both scales and zero-points with two groups and sets $K = 125$.

Table 4: Experiments on OPT-1.3B with 5k training steps. The results show that QZO is substantially under-trained under this short optimization budget.

Method	SST-2	RTE	BoolQ	MultiRC	COPA
QZO	58.94	53.07	53.30	46.50	76.00
CoZO	90.25	59.21	58.80	54.90	79.00

Algorithm 2 ZO_UPDATE($\boldsymbol{\Delta}, \epsilon, seed, \eta, \boldsymbol{\Phi}, W_q, \mathcal{B}, \mathcal{L}$)

Input: parameter vector $\boldsymbol{\Delta}$, perturbation scale ϵ, random seed $seed$, learning rate η, fixed vector $\boldsymbol{\Phi}$, quantized weights W_q, batch $\mathcal{B}$, loss function $\mathcal{L}(\cdot)$
Output: updated $\boldsymbol{\Delta}$
$\boldsymbol{\Delta} \leftarrow$ PERTURBPARAMETERS($\boldsymbol{\Delta}, \epsilon, seed$)
$\ell_+ \leftarrow \mathcal{L}(\boldsymbol{\Delta}, W_q, \boldsymbol{\Phi}; \mathcal{B})$
$\boldsymbol{\Delta} \leftarrow$ PERTURBPARAMETERS($\boldsymbol{\Delta}, -2\epsilon, seed$)
$\ell_- \leftarrow \mathcal{L}(\boldsymbol{\Delta}, W_q, \boldsymbol{\Phi}; \mathcal{B})$
$\boldsymbol{\Delta} \leftarrow$ PERTURBPARAMETERS($\boldsymbol{\Delta}, \epsilon, seed$)
$g_{\boldsymbol{\Delta}} \leftarrow (\ell_+ - \ell_-)/(2\epsilon)$
Reset random number generator with $seed$
for each $\Delta_i \in \boldsymbol{\Delta}$ **do**
    Sample $u \sim \mathcal{N}(0, 1)$
    $\Delta_i \leftarrow \Delta_i - \eta \cdot g_{\boldsymbol{\Delta}} \cdot u$
end for
Procedure: PERTURBPARAMETERS($\boldsymbol{\Delta}, \epsilon, seed$)
Reset random number generator with $seed$
for each $\Delta_i \in \boldsymbol{\Delta}$ **do**
    Sample $u \sim \mathcal{N}(0, 1)$
    $\Delta_i \leftarrow \Delta_i + \epsilon u$
end for
return $\boldsymbol{\Delta}, g_{\boldsymbol{\Delta}}$

C.2 HYPERPARAMETERS

In the experiments with OPT and Llama, we adopt a training strategy similar to that described in (Socher et al., 2013) for classification tasks, where logits corresponding to the label words are extracted and optimized using the cross-entropy loss. For multiple-choice tasks, we retain only the correct candidate and apply teacher forcing on the correct examples. The loss is computed solely on the tokens corresponding to the candidate, excluding the prompt. OPT and Llama share the same set of hyperparameters, which are reported in Table 5.

Table 5: Main hyperparameter settings for all tasks.

Method	**Batch size**	**Learning rate**	**Steps**	**Perturbation scale ϵ**
FT	$\{8, 16\}$	1e−5	1k	/
MeZO	$\{8, 16\}$	5e−6	20k	1e−3
QZO	$\{8, 16\}$	{1e−7, 2e−7}	20k	1e−3
QuZO	$\{8, 16\}$	{1e−6, 1e−7}	{5k, 8k}	1e−3
CoZO	$\{8, 16\}$	{1e−7, 2e−7}	{5k, 8k}	1e−3

C.3 MORE RESULTS

To further verify the consistency of the proposed CoZO algorithm across different model scales, we conduct additional experiments on the OPT-2.7B model. The results are reported in Table 6.

Table 6: Experiments on OPT-2.7B under weight-only quantization. **Bold** indicates the best result among quantized models, and underlined indicates the second-best.

Model	Method	Order	Precision	Memory	SST-2	RTE	BoolQ	COPA	CB	MultiRC
OPT-2.7B	Zero-shot	–	16 bits	–	56.19	54.15	51.10	71.00	41.10	50.00
	Fine-tuning	FO	16 bits	43.75 GB	92.55	60.29	60.10	78.00	69.50	94.64
	MeZO	ZO	16 bits	7.15 GB	93.35	58.48	63.00	75.00	61.00	64.29
	Zero-shot-Q	–	4 bits	–	56.65	53.79	48.10	70.00	41.60	51.79
	FT-Q	FO	4 bits	43.75 GB	91.74	52.71	60.10	74.00	**66.40**	**89.29**
	ZO-Q	ZO	4 bits	7.15 GB	92.20	54.15	54.20	67.00	57.60	60.71
	QZO	ZO	4 bits	6.08 GB	80.04	53.79	59.20	71.10	56.80	66.07
	QuZO	ZO	4 bits	14.83 GB	65.94	54.15	**62.20**	71.00	58.20	66.07
	CoZO (ours)	ZO	4 bits	6.33 GB	**92.66**	**57.40**	**62.20**	**76.00**	59.50	67.86

Table 7: Training time comparison of different methods on SST-2 using OPT-6.7B.

Method	**Training Time**
Fine-tuning	10:00
MeZO	4:19:18
FT-Q	10:00
ZO-Q	4:19:18
QZO	4:07:24
QuZO	6:03:12
CoZO	3:18:33

C.4 Training Time Comparison

To further evaluate the training efficiency of different methods, we compare their training time on the SST-2 dataset using OPT-6.7B. As shown in Table 7, CoZO requires the shortest training time among all ZO-based methods while achieving higher accuracy than the other methods. These results demonstrate that CoZO not only improves fine-tuning performance but also provides clear advantages in training efficiency.

C.5 Hyperparameter Sensitivity

We further investigate the sensitivity of CoZO to the hyperparameters associated with adaptive grouping, including the initial number of groups d_0 and the quantile threshold p. Specifically, when varying d_0, we fix $p = 0.9$, while when varying p, we fix $d_0 = 2$. All experiments are conducted on OPT-1.3B for 5k optimization steps with a fixed learning rate of 1×10^{-7}, while keeping all other settings unchanged.

As shown in Tables 8 and 9, CoZO remains stable over a relatively wide range of hyperparameter values. For the quantile threshold p, the performance is largely unchanged when p varies from 0.7 to 0.9, while a smaller threshold leads to moderate degradation on SST-2 and WIC, likely because more aggressive splitting and merging introduces more frequent changes to the grouping structure. For the initial number of groups d_0, CoZO exhibits little sensitivity across the tested range from 2 to 16, suggesting that the adaptive grouping mechanism can effectively adjust the grouping granularity during optimization.

C.6 Robustness Across Random Seeds

To explore the impact of random seed variation, we fine-tune OPT-1.3B with CoZO on downstream NLP tasks using three different seeds. The results of each run, together with the mean and standard deviation, are presented in Table 10. The results demonstrate that CoZO maintains relatively consistent

Table 8: Sensitivity analysis of the quantile threshold p on OPT-1.3B. The initial number of groups is fixed to $d_0 = 2$.

Threshold p	SST-2	CB	WIC	WSC
0.9	85.89	71.43	58.15	57.69
0.8	85.89	71.43	58.15	58.65
0.7	85.89	71.43	58.15	58.65
0.6	83.26	71.43	55.64	58.65
0.5	83.37	71.43	55.64	58.65

Table 9: Sensitivity analysis of the initial number of groups d_0 on OPT-1.3B. The quantile threshold is fixed to $p = 0.9$.

Initial groups d_0	SST-2	CB	WIC	WSC
2	85.89	71.43	58.15	57.69
4	87.73	71.43	58.15	58.65
8	87.73	71.43	58.15	58.65
16	87.73	71.43	58.15	58.65

performance across different random seeds, indicating the robustness of the proposed method to randomness in the experimental setting.

C.7 Generalization to Broader Model Architectures

To evaluate the generalization ability of the proposed method across other large language models, we conduct additional experiments on the Qwen2-7B model with the SST-2, CB, and RTE datasets. As shown in Table 11, the results demonstrate that our method achieves strong generalization across different model architectures.

C.8 Comparison with QA-LoRA

We further compare CoZO with QA-LoRA (Xu et al., 2024), a representative first-order quantization-aware parameter-efficient fine-tuning method. We choose QA-LoRA rather than QLoRA (Dettmers et al., 2023) because the role of quantization in the two settings is fundamentally different. QLoRA primarily introduces low-bit quantization to reduce the memory overhead of first-order parameter-efficient fine-tuning, while the trainable LoRA adapters remain in higher precision and the adapted model is ultimately operated in a higher-precision computation space. In contrast, CoZO directly optimizes the quantized model through zeroth-order estimation and preserves the low-bit quantized structure throughout both fine-tuning and deployment. QA-LoRA is therefore a more appropriate first-order baseline for our setting, since both QA-LoRA and CoZO are designed to retain the quantized model representation after adaptation.

We conduct the comparison on LLaMA-2-7B under the 4-bit setting. For QA-LoRA, we set the LoRA rank and LoRA alpha to 8 and 16, respectively. Peak GPU memory usage is measured on SST-2 under the same hardware environment. As shown in Table 12, CoZO outperforms QA-LoRA by 3.67 and 5.05 percentage points on SST-2 and RTE, respectively, while QA-LoRA performs better on CB. Averaged over the three tasks, CoZO improves accuracy by 2.21 percentage points and reduces peak GPU memory usage by 1.29 GB. Overall, CoZO achieves better task performance with a smaller memory footprint while preserving the low-bit quantized model throughout optimization.

Table 10: Experiments with three different random seeds using CoZO on OPT-1.3B. The average results are reported as mean ± standard deviation.

Seed	SST-2	CB	WIC	WSC
0	88.65	69.64	58.78	54.81
1	89.68	69.64	57.84	53.85
2	89.56	71.43	57.99	53.85
Average	89.30 ± 0.57	70.24 ± 1.03	58.20 ± 0.50	54.17 ± 0.56

Table 11: Experiments on Qwen2-7B

Method	Precision	SST-2	CB	RTE
Zero-Shot	16 bits	59.17	41.07	67.87
MeZO	16 bits	89.45	96.43	84.12
ZO-Q	4 bits	66.06	82.14	79.78
Zero-Shot-Q	4 bits	64.56	42.86	63.18
QZO	4 bits	87.16	89.29	83.03
QuZO	4 bits	87.16	82.14	80.51
CoZO	4 bits	**88.65**	**92.86**	**83.39**

C.9 Projection-Space Analysis of Zero-Point Optimization

The benefit of optimizing the zero-point can be further understood from the reachable space of dequantized weights. Consider a quantized weight group with fixed integer weights $\mathbf{w}_q$. When only the scale is optimized, the dequantized weights can be expressed as

$$\hat{\mathbf{w}} = s\mathbf{w}_q, \tag{18}$$

such that all attainable weight variations are restricted to $\mathrm{span}\{\mathbf{w}_q\}$. In particular, for entries satisfying $w_{q,i} = 0$, adjusting the scale cannot change their dequantized values. By additionally optimizing the zero-point, the dequantized weights become

$$\hat{\mathbf{w}} = s\mathbf{w}_q + z\mathbf{1}, \tag{19}$$

which expands the reachable space from $\mathrm{span}\{\mathbf{w}_q\}$ to $\mathrm{span}\{\mathbf{w}_q, \mathbf{1}\}$, introducing an additional translation direction independent of the quantized weights. Therefore, zero-point optimization not only introduces additional trainable parameters, but also provides complementary weight-adjustment directions that cannot be represented by scale-only optimization.

To further verify this analysis, we conduct a fixed-code projection experiment on OPT-1.3B. With $\mathbf{w}_q$ kept fixed, we compare the optimal projection errors under scale-only and joint scale–zero parameterizations. As shown in Figure 5, introducing the zero-point consistently reduces the projection error across all Transformer layers. More importantly, the reduction is generally more pronounced for zero-coded weights, i.e., $w_{q,i} = 0$. At these positions, scale-only optimization provides no effective local adjustment direction, whereas the translation freedom introduced by the zero-point can still directly modify the dequantized values. These results demonstrate that zero-point optimization systematically expands the reachable dequantized weight space under fixed quantized codes and provides a complementary adjustment direction to scale-only parameterization.

D Proofs

D.1 Analysis of Variance

D.1.1 Variance Analysis of SPSA

Theorem D.1 (**Variance bound of SPSA**). *For the gradient estimated by Definition 2.1, the variance of the gradient estimator admits the following upper bound:*

$$\mathrm{Var}\big(\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big) \le (d+2)\left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2.$$

Table 12: Comparison between CoZO and QA-LoRA on Llama-2-7B under the 4-bit quantization setting. Peak GPU memory usage is measured on SST-2.

Method	Precision	Memory	SST-2	CB	RTE	Avg.
CoZO (ours)	4 bits	**14.01 GB**	**87.73**	67.56	**63.17**	**72.82**
QA-LoRA	4 bits	15.30 GB	84.06	**69.64**	58.12	70.61

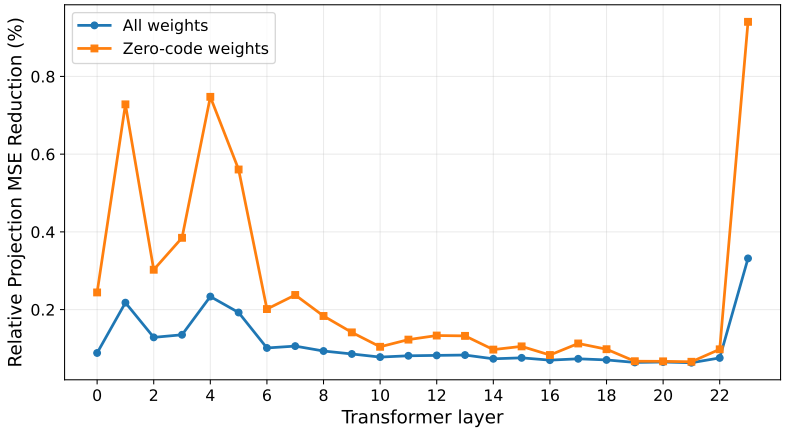


Figure 5: Relative reduction in the optimal projection MSE after introducing zero-point optimization on OPT-1.3B. We report the layer-wise results for all linear weights and zero-coded weights ($w_{q,i} = 0$). Zero-point optimization consistently reduces the projection error across Transformer layers, with more pronounced improvements for zero-coded weights where scale-only optimization cannot directly adjust the dequantized values.

Proof. For the gradient estimated by Definition 2.1, we have,

$$\begin{aligned}\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) &= \frac{\mathcal{L}(\boldsymbol{\theta}+\epsilon u;\mathcal{B})-\mathcal{L}(\boldsymbol{\theta}-\epsilon u;\mathcal{B})}{2\epsilon}u \\ &\approx uu^T\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}),\end{aligned} \tag{20}$$

Let $\hat{g}_1(\boldsymbol{\theta}) = \hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})$ without loss of generality, and its variance can be expressed as,

$$\mathrm{Var}\big(\hat{g}_1(\boldsymbol{\theta})\big) := \mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta}) - \mathbb{E}[\hat{g}_1(\boldsymbol{\theta})]\|^2\Big] \le \mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta})\|^2\Big], \tag{21}$$

Substituting Eq. (20) into $\mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta})\|^2\Big]$, we have,

$$\mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta})\|^2\Big] = \mathbb{E}\Big[\big\|uu^T\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big\|^2\Big] = \mathbb{E}\Big[\|u\|^2\,(u^T\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\Big]. \tag{22}$$

By the fourth-order moment formula of the Gaussian distribution, we can obtain,

$$\mathbb{E}\Big[\|u\|^2\,(u^T\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\Big] = (d+2)\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2. \tag{23}$$

So we have,

$$\mathrm{Var}\big(\hat{g}_1(\boldsymbol{\theta})\big) := \mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta}) - \mathbb{E}[\hat{g}_1(\boldsymbol{\theta})]\|^2\Big] \le \mathbb{E}\Big[\|\hat{g}_1(\boldsymbol{\theta})\|^2\Big] = (d+2)\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2. \tag{24}$$

D.1.2 Variance Analysis of QuZO

Theorem D.2 (Variance bound of quantized RGE). *For the gradient estimated by Definition 2.2, suppose that there exist quantization errors e_i such that $u_i = u + e_i, i \in \{1,2\}$, where $u \sim \mathcal{N}(0, I_d)$. Assume that the quantization errors satisfy $\mathbb{E}\big[\|e_i\|^2\big] \le \sigma_i^2, i \in \{1,2\}$. Then, the variance of the gradient estimator admits the following upper bound:*

$$\mathrm{Var}\big(\bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big) \le 4\,\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2\left((d+2) + d\sigma_1^2 + \sigma_2^2 + \sigma_1^2\sigma_2^2\right).$$

Proof. For the gradient estimated by Definition 2.2, we have,

$$\begin{aligned}\bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) &= \frac{\mathcal{L}(\boldsymbol{\theta}+\epsilon u_1;\mathcal{B})-\mathcal{L}(\boldsymbol{\theta}-\epsilon u_1;\mathcal{B})}{2\epsilon}u_2 \\ &\approx u_2 u_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}),\end{aligned} \tag{25}$$

Let $\hat{g}_2(\boldsymbol{\theta}) = \bar{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B})$ without loss of generality, and its variance can be expressed as,

$$\mathrm{Var}\big(\hat{g}_2(\boldsymbol{\theta})\big) := \mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta}) - \mathbb{E}[\hat{g}_2(\boldsymbol{\theta})]\|^2\Big] \le \mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big], \tag{26}$$

Substituting Eq. (25) into $\mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big]$, we have,

$$\mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big] = \mathbb{E}\Big[\big\|u_2 u_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big\|^2\Big]. \tag{27}$$

Since the perturbation passes through low-bit quantization, there will be quantization errors for both u_1 and u_2. We denote the perturbation before quantization as:

$$u_1 = u + e_1, \quad u_2 = u + e_2 \tag{28}$$

Since $\mathbb{E}[u_i] = u$, we have,

$$\mathbb{E}[e_i] = 0, \quad i = 1, 2 \tag{29}$$

Since $u_2 u_1^\top \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) = u_2(u_1^\top, \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))$, we have,

$$\|\hat{g}_2(\boldsymbol{\theta})\|^2 = \|u_2\|^2 \left(u_1^\top \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right)^2 = \|u+e_2\|^2 \left((u+e_1)^\top \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right)^2. \tag{30}$$

Let $\mathbb{E}\|e_i\|^2 \le \sigma_i^2$ for $i \in \{1,2\}$, and u is independent of (e_1, e_2). Then we have,

$$\begin{aligned}\mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big] &= \mathbb{E}\Big[\big\|u_2 u_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\big\|^2\Big] \\ &= \mathbb{E}\Big[\|u_2\|^2 \left(u_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right)^2\Big] \\ &= \mathbb{E}\Big[\|u+e_2\|^2 \left((u+e_1)^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right)^2\Big].\end{aligned} \tag{31}$$

Using the inequalities $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and $(x+y)^2 \le 2x^2 + 2y^2$, we obtain,

$$\|u+e_2\|^2 \left((u+e_1)^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right)^2 \le 4\Big(\|u\|^2 + \|e_2\|^2\Big)\Big((u^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2 + (e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\Big). \tag{32}$$

Taking expectation on both sides and expanding the product yields,

$$\begin{aligned}\mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big] \le 4\Big(&\mathbb{E}\big[\|u\|^2 (u^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] + \mathbb{E}\big[\|u\|^2 (e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] \\ &+ \mathbb{E}\big[\|e_2\|^2 (u^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] + \mathbb{E}\big[\|e_2\|^2 (e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big]\Big).\end{aligned} \tag{33}$$

By the fourth-order moment formula of the Gaussian distribution, we have,

$$\mathbb{E}\big[\|u\|^2 (u^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] = (d+2)\left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2. \tag{34}$$

For the second term, using the independence between u and e_1 and the fact that $\mathbb{E}\|u\|^2 = d$, we obtain,

$$\begin{aligned}\mathbb{E}\big[\|u\|^2 (e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] &= \mathbb{E}\|u\|^2 \cdot \mathbb{E}\big[(e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] \\ &\le d\,\mathbb{E}\|e_1\|^2 \left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2 \le d\sigma_1^2 \left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2,\end{aligned} \tag{35}$$

where we used $(e_1^T x)^2 \le \|e_1\|^2 \|x\|^2$. Similarly, for the third term, we have,

$$\mathbb{E}\big[\|e_2\|^2 (u^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] \le \sigma_2^2 \left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2, \tag{36}$$

where we used $\mathbb{E}[(u^T x)^2] = \|x\|^2$ for $u \sim \mathcal{N}(0, I_d)$. For the last term, we have,

$$\mathbb{E}\big[\|e_2\|^2 (e_1^T \nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B}))^2\big] \le \mathbb{E}\big[\|e_2\|^2 \|e_1\|^2\big] \left\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\right\|^2. \tag{37}$$

Since e_1 and e_2 are independent, then,

$$\mathbb{E}\big[\|e_2\|^2 \|e_1\|^2\big] \le \sigma_2^2 \sigma_1^2. \tag{38}$$

Combining the above results yields,

$$\mathbb{E}\Big[\|\hat{g}_2(\boldsymbol{\theta})\|^2\Big] \leq 4\,\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2\,\Big((d+2)+d\sigma_1^2+\sigma_2^2+\sigma_1^2\sigma_2^2\Big). \tag{39}$$

Therefore, we obtain,

$$\mathrm{Var}\big(\hat{g}_2(\boldsymbol{\theta})\big) \leq 4\,\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2\,\Big((d+2)+d\sigma_1^2+\sigma_2^2+\sigma_1^2\sigma_2^2\Big). \tag{40}$$

By setting $\sigma = \sigma_2^2 + \sigma_1^2\sigma_2^2$, we obtain,

$$\mathrm{Var}\big(\hat{g}_2(\boldsymbol{\theta})\big) \leq 4\,\|\nabla\mathcal{L}(\boldsymbol{\theta};\mathcal{B})\|^2\,\Big((d+2)+d\sigma_1^2+\sigma\Big). \tag{41}$$

D.2 Proof of Theorem 4.7

Lemma D.3. $\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\right] = \|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 + \frac{1}{b}\mathrm{tr}\left(\Sigma(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right)$.

Proof. By the definition of $\Sigma(\boldsymbol{S}_t,\boldsymbol{Z}_t)$, we have,

$$\Sigma(\boldsymbol{S}_t,\boldsymbol{Z}_t) = b\Big(\mathbb{E}\big[\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})^\top\big] - \nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)^\top\Big). \tag{42}$$

Dividing both sides of the equation by b and taking the trace yields,

$$\begin{aligned}\frac{1}{b}\,\mathrm{tr}(\Sigma(\boldsymbol{S}_t,\boldsymbol{Z}_t)) &= \mathrm{tr}\big(\mathbb{E}\big[\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})^\top\big]\big) - \mathrm{tr}\big(\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)^\top\big)\\ &= \mathbb{E}\big[\mathrm{tr}\big(\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})^\top\big)\big] - \mathrm{tr}\big(\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\,\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)^\top\big)\\ &= \mathbb{E}\big[\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\big] - \|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2.\end{aligned} \tag{43}$$

By simple rearrangement of terms, we obtain,

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\right] = \|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 + \frac{1}{b}\mathrm{tr}\left(\Sigma(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right). \tag{44}$$

Lemma D.4. *Let $\mathcal{B}$ be a random minibatch of size b. Then, the gradient norm of ZO is*

$$\mathbb{E}\left[\left\|\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B}))\right\|^2\right] = \frac{d+n-1}{n}\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\|^2\right], \tag{45}$$

$$\mathbb{E}\left[\left\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B}))\right\|^2\right] = \frac{d+n-1}{n}\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\|^2\right]. \tag{46}$$

Proof. Noting that as $\epsilon \to 0$ limit, we have,

$$\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B}) = \frac{1}{bn}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{B}}\sum_{i\in[n]}\boldsymbol{u}_i\boldsymbol{u}_i^\top\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x},\boldsymbol{y})\}\big). \tag{47}$$

Taking expectation over the batch b and the $\boldsymbol{u}_i$, we have $\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\right] = \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})$, so $\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})$ is an unbiased estimator of the gradient. Computing the second moment, we have,

$$\begin{aligned}&\mathbb{E}\Big[\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\Big]\\ &= \frac{1}{b^2n^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\sum_{i,j\in[n]}\mathbb{E}\Bigg[\Big(\boldsymbol{u}_i\boldsymbol{u}_i^\top\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)\Big)\\ &\quad\times\Big(\boldsymbol{u}_j\boldsymbol{u}_j^\top\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)\Big)^\top\Bigg].\end{aligned} \tag{48}$$

Let v_1, v_2 be two arbitrary vectors. when $i \neq j$, we have that,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_i,\boldsymbol{u}_j}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top \boldsymbol{u}_j\boldsymbol{u}_j^\top\right] &= \mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top\right]\mathbb{E}_{\boldsymbol{u}_j}\left[\boldsymbol{u}_j\boldsymbol{u}_j^\top\right]\\ &= \mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top\right] I\\ &= v_1 {v_2}^\top,\end{aligned} \tag{49}$$

and when $i = j$, we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top \boldsymbol{u}_i\boldsymbol{u}_i^\top\right] &= \mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{u}^{\otimes 4}\right](v_1, v_2)\\ &= \frac{3d}{d+2}\mathrm{Sym}(I^{\otimes 2})(v_1, v_2)\\ &= \frac{d}{d+2}\cdot v_1^\top v_2 \cdot I + \frac{2d}{d+2}\cdot v_1 v_2^\top.\end{aligned} \tag{50}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\Big[\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\,\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\Big]\\ &= \frac{1}{b^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\left(\frac{n-1}{n}+\frac{2d}{n(d+2)}\right)\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)\,\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)^\top\Big]\\ &\quad+\frac{d}{n(d+2)}\cdot\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)^\top\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)\Big]\,\boldsymbol{I}.\end{aligned} \tag{51}$$

Next, observe that when $(x_1, y_1) \neq (x_2, y_2)$, we have,

$$\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)\,\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)^\top\Big] = \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\,\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top. \tag{52}$$

In the case where $(x_1, y_1) = (x_2, y_2)$, we have,

$$\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)\,\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)^\top\Big] = \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\,\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}). \tag{53}$$

Therefore,

$$\begin{aligned}&\frac{1}{b^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\mathbb{E}\Big[\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}\big)\,\nabla_{\boldsymbol{S}}\mathcal{L}\big(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}\big)^\top\Big]\\ &= \nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\,\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \frac{1}{b}\,\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}).\end{aligned} \tag{54}$$

Hence, Eq. (51) can be expressed as:

$$\begin{aligned}&\mathbb{E}\Big[\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\,\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\Big]\\ &= \left(1+\frac{d-2}{n(d+2)}\right)\left(\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\,\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top+\frac{1}{b}\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z})\right)\\ &\quad+\frac{d}{n(d+2)}\,\boldsymbol{I}\cdot\left(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\|^2+\frac{1}{b}\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}))\right).\end{aligned} \tag{55}$$

Taking the trace of both sides of the above equation, we obtain,

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\right\|^2\right] &= \left(1+\frac{d^2+d-2}{n(d+2)}\right)\left(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\|^2+\frac{1}{b}\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}))\right)\\ &= \frac{d+n-1}{n}\cdot\mathbb{E}\Big[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\|^2\Big].\end{aligned} \tag{56}$$

Similarly, we can obtain the result for $\boldsymbol{Z}$.

Lemma D.5. *For a twice continuously differentiable function $\mathcal{L}$ on the line segment connecting θ_t and θ_{t+1}, let $\Delta = \theta_{t+1} - \theta_t$. Then, the following equality holds:*

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) =&\mathcal{L}(\theta_t) + \nabla\mathcal{L}(\theta_t)^T\Delta\\ &+\int_0^1(1-\lambda)\Delta^T\nabla^2\mathcal{L}(\theta_t+\lambda\Delta)\Delta\,d\lambda.\end{aligned} \tag{57}$$

Proof. Define $h(\lambda) = \mathcal{L}(\theta_t + \lambda\Delta)$. By the chain rule, we have

$$\begin{aligned} h'(\lambda) &= \nabla\mathcal{L}(\theta_t + \lambda\Delta)^T\Delta, \\ h''(\lambda) &= \Delta^T\nabla^2\mathcal{L}(\theta_t + \lambda\Delta)\Delta. \end{aligned} \tag{58}$$

Applying the fundamental theorem of calculus and changing the order of integration yields

$$\begin{aligned} h(1) - h(0) &= \int_0^1 h'(\lambda)\, d\lambda \\ &= h'(0) + \int_0^1 \int_0^\lambda h''(s)\, ds\, d\lambda \\ &= h'(0) + \int_0^1 (1-s)h''(s)\, ds. \end{aligned} \tag{59}$$

Substituting the expressions for $h'(0)$ and $h''(s)$ completes the proof.

Lemma D.6 (Dimension-Free Rate). *Assume the loss exhibits local r-effective rank (Assumption 4.2). if $\boldsymbol{S}$, $\boldsymbol{Z}$, are all updated at each iteration using ZO-SGD with the n-SPSA estimator and a minibatch of size b, then there exist constants $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$ such that the expected loss decrease can be bounded as:*

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)|\boldsymbol{S}_t\right] - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t) \leq -\eta||\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)||^2 + \frac{L_{\boldsymbol{S}}}{2}\eta^2\gamma_{\mathbf{1}}\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2\right], \tag{60}$$

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1})|\boldsymbol{Z}_t\right] - \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) &\leq -\eta\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|^2 \\ &\quad + \frac{L_{\boldsymbol{Z}}}{2}\eta^2\gamma_{\mathbf{2}}\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2\right]. \end{aligned} \tag{61}$$

Proof. By Lemma D.5, we have that,

$$\begin{aligned} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) &= \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) + \nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^\top(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t) \\ &\quad + \int_0^1 (1-\lambda)\left(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t\right)^\top\nabla^2_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \lambda\boldsymbol{Z}_{t+1} + (1-\lambda)\boldsymbol{Z}_t)\left(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t\right)\mathrm{d}\lambda. \end{aligned} \tag{62}$$

Next, note that,

$$\begin{aligned} \|\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t\| &= \eta\left\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\right\| \\ &\overset{(a)}{\leq} \eta\sqrt{d}\cdot\frac{1}{bn}\sum\left|\boldsymbol{u}_i^\top\nabla_{\boldsymbol{Z}}\mathcal{L}\big(\boldsymbol{S}, \boldsymbol{Z}; \{(\boldsymbol{x}, \boldsymbol{y})\}\big)\right| \\ &\overset{(b)}{\leq} \eta\, d\, G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \end{aligned} \tag{63}$$

where (a) is due to sampling $\boldsymbol{u}$ from a spherical distribution with radius $1/\sqrt{d}$, (b) is a consequence of Assumption 4.2. Therefore $\|\lambda\boldsymbol{Z}_{t+1} + (1-\lambda)\boldsymbol{Z}_t - \boldsymbol{Z}_t\| \leq \lambda\,\eta\, d\, G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$. By the assumption we have the upper bound $\nabla^2_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \lambda\boldsymbol{Z}_{t+1} + (1-\lambda)\boldsymbol{Z}_t) \preceq \boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$, and thus,

$$\begin{aligned} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) &\leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) + \nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^\top(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t) \\ &\quad + \frac{1}{2}(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t)^\top\boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t) \\ &= \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta\,\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^\top\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \\ &\quad + \frac{1}{2}\eta^2\,\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top\boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}). \end{aligned} \tag{64}$$

Taking the conditional expectation with respect to $\boldsymbol{Z}_t$, the formula for the covariance of our ZO estimate $\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})$, yields,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) \mid \boldsymbol{Z}_t] &\leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta \left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&\quad + \frac{\eta^2}{2}\Big\langle \boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\Big[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\, \hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top\Big] \Big\rangle \\
&\overset{(a)}{=} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta \left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&\quad + \frac{\eta^2}{2}\cdot\frac{d}{n(d+2)}\left(\left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 + \frac{1}{b}\operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\right)\operatorname{tr}(\boldsymbol{H}(\boldsymbol{Z}_t)) \\
&\quad + \frac{\eta^2}{2}\left(1 + \frac{d-2}{n(d+2)}\right)\Bigg(\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{Z}_t)^\top \boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{Z}_t) \\
&\qquad\qquad + \frac{1}{b}\Big\langle \Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\Big\rangle\Bigg).
\end{aligned} \tag{65}$$

where (a) is due to plug in Eq. (55). By assumption, the Hessian upper bound $\boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$ satisfies $\|\boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|_{\mathrm{op}} \leq L_{\boldsymbol{Z}}$ and $\operatorname{tr}(\boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)) \leq L_{\boldsymbol{Z}}\, r$. Thus,

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) \mid \boldsymbol{Z}_t] \leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta \left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&+ \frac{L_{\boldsymbol{Z}}}{2}\eta^2\left(\frac{dr + d - 2}{n(d+2)} + 1\right)\left(\left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 + \frac{1}{b}\operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\right) \\
&= \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta \left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&\quad + \frac{L_{\boldsymbol{Z}}}{2}\eta^2\left(\frac{dr + d - 2}{n(d+2)} + 1\right)\mathbb{E}\Big[\left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\right\|^2\Big].
\end{aligned} \tag{66}$$

Similarly, we can obtain the result for $\boldsymbol{S}$.

Lemma D.7. *Under Assumption 4.1 and Assumption 4.2, if $\boldsymbol{S}$, $\boldsymbol{Z}$, are all updated at each iteration using ZO-SGD with the n-SPSA estimator and a minibatch of size b, then there exist $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$ such that the expected decrease in the loss can be bounded as:*

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\right] &\leq -\eta(||\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)||^2 + ||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)||^2) \\
&\quad + \frac{\eta^2}{2}(L_{\boldsymbol{S}}\gamma_1\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2\right] \\
&\quad + L_{\boldsymbol{Z}}\gamma_2\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2\right]).
\end{aligned} \tag{67}$$

Proof. By combining Eq. (60) and (61), and taking expectations, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\right] &\leq -\eta(||\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)||^2 + ||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)||^2) \\
&\quad + \frac{\eta^2}{2}(L_{\boldsymbol{S}}\gamma_1\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2\right] \\
&\quad + L_{\boldsymbol{Z}}\gamma_2\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2\right]).
\end{aligned} \tag{68}$$

Theorem D.8. *Assumptions 4.1-4.4 are all satisfied, if $\boldsymbol{S}$, $\boldsymbol{Z}$, are all updated at each iteration using ZO-SGD with the n-SPSA estimator and a minibatch of size b, and the learning rate is chosen as $\eta = O(\frac{1}{\sqrt{Td}})$, then there exist $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$, such that the following inequality holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq O(\frac{\sqrt{d}}{\sqrt{T}} + \frac{1}{Tb} + \frac{\sqrt{d}}{\sqrt{T}b}).$$

Proof. For $\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2$, we can bound it as follows:

$$\mathbb{E}[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2] \leq 2||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)||^2 + 2||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)||^2. \tag{69}$$

By leveraging the Lipschitz continuity of $\boldsymbol{S}$, we obtain,

$$\begin{aligned}\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 &\le 2L_{\boldsymbol{S}}^2\|\boldsymbol{S}_{t+1}-\boldsymbol{S}_t\|^2+2||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)||^2\\ &= 2L_{\boldsymbol{S}}^2\|\eta\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2+2||\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)||^2.\end{aligned} \tag{70}$$

Taking expectations on both sides yields and let $\eta \le \min\{\frac{1}{L_{\max}},\frac{1}{L'_{\max}}\}$, where $L'_{\max}=\max\{\gamma_1 L_{\boldsymbol{S}},\gamma_2 L_{\boldsymbol{Z}}\}$, we have,

$$\begin{aligned}\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le{}& 2L_{\boldsymbol{S}}^2\eta^2\mathbb{E}\left[\|\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\right]+2\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\\ \overset{(a)}{=}{}& \frac{2L_{\boldsymbol{S}}^2\eta^2(d+n-1)}{n}\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t;\mathcal{B})\|^2\right]+2\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\\ \overset{(b)}{=}{}& \frac{2L_{\boldsymbol{S}}^2\eta^2(d+n-1)}{n}\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2+2\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\\ &+\frac{2L_{\boldsymbol{S}}^2\eta^2(d+n-1)}{nb}tr\left(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right)\\ \le{}& 2\left[\frac{2L_{\boldsymbol{S}}^2(d+n-1)}{nL_{max}^2}\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2+\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\right]\\ &+\frac{2L_{\boldsymbol{S}}^2\eta^2(d+n-1)}{nb}tr\left(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right),\end{aligned} \tag{71}$$

where (a) follows from Lemma D.4,and (b) follows from Lemma D.3. Adding $\|\nabla_{\boldsymbol{S}}L(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2$ to the inequality yields, By adding $\|\nabla_{\boldsymbol{S}}L(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2$ to the inequality, denoting $\frac{2L_{\boldsymbol{S}}^2(d+n-1)}{nL_{max}^2}$ as C_1 and denoting $\frac{2L_{\boldsymbol{S}}^2(d+n-1)}{n}$ as C_2, we have,

$$\begin{aligned}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 &= \|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2+\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2\\ &\le (1+C_1)\left(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2+\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\right)+\frac{\eta^2C_2}{b}tr\left(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right).\end{aligned} \tag{72}$$

By applying Lemma D.3, Eq. (67) can be written as,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1})-\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)] \le{}& -\frac{\eta}{2}\Big(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2+\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t)\|^2\Big)\\ &+\frac{\eta^2}{2b}\Big[L_{\boldsymbol{S}}\,\gamma_1\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))+L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big].\end{aligned} \tag{73}$$

Substituting Eq. (72) into Eq. (73), we obtain,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1})-\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)] \le{}& -\frac{\eta}{2}\left[\frac{1}{1+C_1}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2-\frac{\eta^2C_2}{b(1+C_1)}\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))\right]\\ &+\frac{\eta^2}{2b}\Big[L_{\boldsymbol{S}}\gamma_1\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))+L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big].\end{aligned} \tag{74}$$

By rearranging the above inequality, we obtain,

$$\begin{aligned}\frac{\eta}{2(1+C_1)}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le{}& \mathbb{E}[\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)-\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1})]+\frac{\eta^3C_2}{2b(1+C_1)}\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))\\ &+\frac{\eta^2}{2b}\Big[L_{\boldsymbol{S}}\gamma_1\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))+L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big].\end{aligned} \tag{75}$$

By dividing both sides of the inequality by $\frac{\eta}{2(1+C_1)}$, we have,

$$\begin{aligned}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le{}& \frac{2(1+C_1)}{\eta}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)-\mathcal{L}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1})]+\frac{\eta^2C_2}{b}\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))\\ &+\frac{\eta(1+C_1)}{b}\Big[L_{\boldsymbol{S}}\gamma_1\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))+L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big].\end{aligned} \tag{76}$$

Summing both sides of the inequality with respect to t from 0 to T-1, we have,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le \frac{2(1+C_1)}{\eta T}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}(\boldsymbol{S}_T,\boldsymbol{Z}_T)] + \frac{1}{T}\sum_{t=0}^{T-1}\frac{\eta^2 C_2}{b}\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)) + \sum_{t=0}^{T-1}\frac{\eta(1+C_1)}{bT}\Big[L_{\boldsymbol{S}}\gamma_1\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big]. \tag{77}$$

Since $\mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}(\boldsymbol{S}_T,\boldsymbol{Z}_T)\le\mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}^*$, where $\mathcal{L}^*=\min_{\boldsymbol{S},\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})$, taking expectations on both sides of the inequality yields,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le \frac{2(1+C_1)}{\eta T}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}^*] + \frac{1}{T}\sum_{t=0}^{T-1}\frac{\eta^2 C_2}{b}\mathbb{E}\Big[\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t))\Big] + \sum_{t=0}^{T-1}\frac{\eta(1+C_1)}{bT}\mathbb{E}\Big[L_{\boldsymbol{S}}\gamma_1\,\mathrm{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t,\boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_t))\Big]. \tag{78}$$

Due to Assumption 4.4, we obtain,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{S}_t,\boldsymbol{Z}_t)\|^2 \le \frac{2(1+C_1)}{\eta T}(\mathcal{L}(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}^*) + \frac{\eta^2 C_2 d\lambda_{\boldsymbol{S}}}{b} + \frac{d\eta(1+C_1)}{b}\mathbb{E}\Big[L_{\boldsymbol{S}}\gamma_1\lambda_{\boldsymbol{S}} + L_{\boldsymbol{Z}}\,\gamma_2\lambda_{\boldsymbol{Z}}\Big]. \tag{79}$$

By choosing $\eta = O(\frac{1}{\sqrt{Td}})$, the above gives the fixed-grouping bound used as a one-step building block below.

D.3 Variance–Expressiveness Effect of Adaptive Grouping

We first formalize the mechanism underlying the adaptive grouping rule in Section 3.2. For a fixed local variability level $v \ge 0$, define the dimension-dependent SPSA variance factor

$$\mathcal{V}(d;v) := (d+2)v. \tag{80}$$

Lemma D.9 (Effect of regrouping on the SPSA variance factor). *For an admissible merge with even d and a split from d to $2d$, respectively,*

$$\mathcal{V}(d/2;v) = \left(\frac{d}{2}+2\right)v = \frac{d+4}{2d+4}\,\mathcal{V}(d;v), \qquad \mathcal{V}(2d;v) = (2d+2)v = \frac{2d+2}{d+2}\,\mathcal{V}(d;v). \tag{81}$$

Hence, $\mathcal{V}(d/2;v)\le\mathcal{V}(d;v)\le\mathcal{V}(2d;v)$, with strict inequalities whenever $v>0$. Moreover, SPLIT *is a refinement of the current grouping, so its reachable continuous mapping space contains that of the pre-split grouping, while* MERGE *imposes the reverse constraint.*

D.4 Adaptive-Grouping Extension

We now complete the proof of Theorem 4.7 for the adaptive grouping procedure in Algorithm 1. Since the dimension of the grouped parameterization changes over time, we first make the active grouped objective explicit.

Let $\mathcal{G}_t = \{\mathcal{G}_{t,1},\dots,\mathcal{G}_{t,d_t}\}$ be the grouping structure active at iteration t. Define the broadcast matrix $\boldsymbol{A}_t \in \{0,1\}^{m\times d_t}$ by $(\boldsymbol{A}_t)_{ji}=1$ if the j-th dequantization parameter belongs to group $\mathcal{G}_{t,i}$ and $(\boldsymbol{A}_t)_{ji}=0$ otherwise. For grouped parameters $\boldsymbol{S},\boldsymbol{Z}\in\mathbb{R}^{d_t}$, their full-dimensional representations are $\boldsymbol{A}_t\boldsymbol{S}$ and $\boldsymbol{A}_t\boldsymbol{Z}$. We therefore write the objective under the active grouping as

$$\mathcal{L}_t(\boldsymbol{S},\boldsymbol{Z}) = \mathcal{L}((\boldsymbol{A}_t\boldsymbol{S})\odot W_q + \boldsymbol{A}_t\boldsymbol{Z}), \tag{82}$$

where the dependence on the input batch is omitted for brevity. Let $\mathcal{F}_t$ denote the sigma-field generated by all randomness before sampling the minibatch and perturbation directions at iteration t. Then $\mathcal{G}_t$, d_t, $\boldsymbol{S}_t$, and $\boldsymbol{Z}_t$ are $\mathcal{F}_t$-measurable. We use $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and write $\Sigma_{\boldsymbol{S},t}$ and $\Sigma_{\boldsymbol{Z},t}$ for Definition 4.3 applied to $\mathcal{L}_t$.

Assumption D.10 (Uniform regularity under adaptive grouping). Let $\mathfrak{G}$ denote the set of admissible grouping structures, with $d_{\min} \le d(\mathcal{G}) \le d_{\max}$. We assume that Assumptions 4.1–4.4 hold uniformly for all $\mathcal{G} \in \mathfrak{G}$ with common constants $L_{\boldsymbol{S}}, L_{\boldsymbol{Z}}, \lambda_{\boldsymbol{S}}, \lambda_{\boldsymbol{Z}}$, and that the corresponding local effective rank satisfies $r_{\mathcal{G}} \le r_{\max}$.

Let $(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1})$ denote the parameters after completing the alternating ZO update but before regrouping, and let $(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1})$ denote the parameters after SPLIT, MERGE, or MAINTAIN. Define

$$\varepsilon_t = \Big[\mathcal{L}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) - \mathcal{L}_t(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1})\Big]_+, \qquad \mathcal{E}_T = \sum_{t=0}^{T-1} \mathbb{E}[\varepsilon_t], \tag{83}$$

which is the formal active-grouping interpretation of Eq. (15).

Conditioned on $\mathcal{F}_t$, the grouping structure is fixed during the alternating ZO update. Hence, the one-step inequality in the proof of Theorem 4.7 can be applied with d replaced by d_t. Let

$$C_{1,t} = \frac{2L_{\boldsymbol{S}}^2(d_t + n - 1)}{nL_{\max}^2}, \qquad C_{2,t} = \frac{2L_{\boldsymbol{S}}^2(d_t + n - 1)}{n}, \tag{84}$$

and let $\gamma_{1,t} = \Theta(r_t/n)$ and $\gamma_{2,t} = \Theta(r_t/n)$, where r_t is the local effective rank under $\mathcal{G}_t$. Denoting the parameters after the alternating update but before regrouping by $(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1})$, we obtain

$$\begin{aligned}\frac{\eta}{2(1+C_{1,t})}\left\|\nabla\mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t)\right\|^2 \le{}& \mathbb{E}_t\Big[\mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t) - \mathcal{L}_t(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1})\Big] \\ &+ \frac{\eta^3 C_{2,t}}{2b(1+C_{1,t})}\mathbb{E}_t[\operatorname{tr}(\Sigma_{\boldsymbol{S},t}(\boldsymbol{S}_t, \boldsymbol{Z}_t))] \\ &+ \frac{\eta^2}{2b}\mathbb{E}_t\Big[L_{\boldsymbol{S}}\gamma_{1,t}\operatorname{tr}(\Sigma_{\boldsymbol{S},t}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ &\qquad + L_{\boldsymbol{Z}}\gamma_{2,t}\operatorname{tr}\Big(\Sigma_{\boldsymbol{Z},t}(\widetilde{\boldsymbol{S}}_{t+1}, \boldsymbol{Z}_t)\Big)\Big].\end{aligned} \tag{85}$$

Recall the regrouping error in Eq. (83). By the definition of the positive part, the following inequality holds pathwise, without requiring the regrouping decision to be independent of the ZO update:

$$\mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t) - \mathcal{L}_t(\widetilde{\boldsymbol{S}}_{t+1}, \widetilde{\boldsymbol{Z}}_{t+1}) \le \mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t) - \mathcal{L}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) + \varepsilon_t. \tag{86}$$

Substituting Eq. (86) into Eq. (85), taking total expectations, and using the tower property gives a valid one-step bound for the adaptive procedure.

Under Assumption D.10, define

$$\bar{C}_1 = \max_{d_{\min}\le d\le d_{\max}} \frac{2L_{\boldsymbol{S}}^2(d+n-1)}{nL_{\max}^2}, \qquad \bar{C}_2 = \max_{d_{\min}\le d\le d_{\max}} \frac{2L_{\boldsymbol{S}}^2(d+n-1)}{n}, \tag{87}$$

and $\bar{\gamma}_i = \sup_t \gamma_{i,t} < \infty$ for $i \in \{1, 2\}$. Since $C_{1,t} \le \bar{C}_1$ and $C_{2,t} \le \bar{C}_2$, summing over $t = 0, \dots, T-1$ yields

$$\begin{aligned}\frac{\eta}{2(1+\bar{C}_1)}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t)\right\|^2 \le{}& \mathcal{L}_0(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}^* + \mathcal{E}_T \\ &+ \frac{\eta^3\bar{C}_2}{2b}\sum_{t=0}^{T-1}\mathbb{E}\operatorname{tr}(\Sigma_{\boldsymbol{S},t}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ &+ \frac{\eta^2}{2b}\sum_{t=0}^{T-1}\mathbb{E}\Big[L_{\boldsymbol{S}}\bar{\gamma}_1\operatorname{tr}(\Sigma_{\boldsymbol{S},t}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ &\qquad + L_{\boldsymbol{Z}}\bar{\gamma}_2\operatorname{tr}\Big(\Sigma_{\boldsymbol{Z},t}(\widetilde{\boldsymbol{S}}_{t+1}, \boldsymbol{Z}_t)\Big)\Big].\end{aligned} \tag{88}$$

The loss terms telescope because the post-regrouping objective at iteration t is exactly the objective used to initialize iteration $t+1$. Moreover, Assumption 4.4 and $d_t \le d_{\max}$ imply

$$\operatorname{tr}(\Sigma_{\boldsymbol{S},t}) \le d_{\max}\lambda_{\boldsymbol{S}}, \qquad \operatorname{tr}(\Sigma_{\boldsymbol{Z},t}) \le d_{\max}\lambda_{\boldsymbol{Z}}. \tag{89}$$

Therefore, Eq. (88) gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}_t(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right\|^2 \le{}& \frac{2(1+\bar{C}_1)}{\eta T}\left(\mathcal{L}_0(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}^*+\mathcal{E}_T\right)\\ &+\frac{(1+\bar{C}_1)\eta^2\bar{C}_2 d_{\max}\lambda_{\boldsymbol{S}}}{b}\\ &+\frac{(1+\bar{C}_1)\eta d_{\max}}{b}\left(L_{\boldsymbol{S}}\bar{\gamma}_1\lambda_{\boldsymbol{S}}+L_{\boldsymbol{Z}}\bar{\gamma}_2\lambda_{\boldsymbol{Z}}\right).\end{aligned} \tag{90}$$

Following the same simplification of the dimension-dependent constants as in Theorem 4.7, and choosing $\eta = O(1/\sqrt{Td_{\max}})$, Eq. (90) gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}_t(\boldsymbol{S}_t,\boldsymbol{Z}_t)\right\|^2 \le O\left(\left(\mathcal{L}_0(\boldsymbol{S}_0,\boldsymbol{Z}_0)-\mathcal{L}^*+\mathcal{E}_T\right)\sqrt{\frac{d_{\max}}{T}}+\frac{1}{Tb}+\frac{\sqrt{d_{\max}}}{\sqrt{T}b}\right), \tag{91}$$

which proves Theorem 4.7.

We next characterize $\mathcal{E}_T$ using the structure of the regrouping operations. For MAINTAIN, both the grouping structure and grouped parameters remain unchanged, so $\varepsilon_t = 0$. For SPLIT, each current group is partitioned into two consecutive subgroups and both subgroups inherit the original shared scale and zero point. Hence,

$$\boldsymbol{A}_{t+1}\boldsymbol{S}_{t+1} = \boldsymbol{A}_t\widetilde{\boldsymbol{S}}_{t+1}, \qquad \boldsymbol{A}_{t+1}\boldsymbol{Z}_{t+1} = \boldsymbol{A}_t\widetilde{\boldsymbol{Z}}_{t+1}, \tag{92}$$

which leaves the dequantized weights and the loss unchanged. Thus, split also satisfies $\varepsilon_t = 0$.

For an admissible merge with even d_t, define the pairwise averaging projection $\boldsymbol{P}_t \in \mathbb{R}^{d_t\times d_t}$ as a block-diagonal matrix whose 2×2 blocks equal

$$\frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}. \tag{93}$$

Then $\boldsymbol{P}_t^\top = \boldsymbol{P}_t$, $\boldsymbol{P}_t^2 = \boldsymbol{P}_t$, and $\|\boldsymbol{P}_t\|_{\mathrm{op}} = 1$. Let

$$\delta_{\boldsymbol{S},t} = \left\|(\boldsymbol{P}_t-\boldsymbol{I})\widetilde{\boldsymbol{S}}_{t+1}\right\|, \qquad \delta_{\boldsymbol{Z},t} = \left\|(\boldsymbol{P}_t-\boldsymbol{I})\widetilde{\boldsymbol{Z}}_{t+1}\right\|. \tag{94}$$

The merged representation is equivalent, in the pre-merge coordinate system, to assigning the pairwise averaged value to both members of each adjacent pair. Consequently,

$$\mathcal{L}_{t+1}(\boldsymbol{S}_{t+1},\boldsymbol{Z}_{t+1}) = \mathcal{L}_t(\boldsymbol{P}_t\widetilde{\boldsymbol{S}}_{t+1},\boldsymbol{P}_t\widetilde{\boldsymbol{Z}}_{t+1}). \tag{95}$$

Moreover, the merge displacement is exactly determined by the disagreement between adjacent groups:

$$\delta^2_{\boldsymbol{S},t} = \frac{1}{2}\sum_{j=1}^{d_t/2}\left(\widetilde{S}_{t+1,2j-1}-\widetilde{S}_{t+1,2j}\right)^2, \qquad \delta^2_{\boldsymbol{Z},t} = \frac{1}{2}\sum_{j=1}^{d_t/2}\left(\widetilde{Z}_{t+1,2j-1}-\widetilde{Z}_{t+1,2j}\right)^2. \tag{96}$$

Suppose the pre-regrouping iterates satisfy the standard bounded-iterate condition

$$\|\widetilde{\boldsymbol{S}}_{t+1}\|_\infty \le B_{\boldsymbol{S}}, \qquad \|\widetilde{\boldsymbol{Z}}_{t+1}\|_\infty \le B_{\boldsymbol{Z}}. \tag{97}$$

Since the gradients are Lipschitz continuous, they are bounded on this bounded domain. Thus, there exist finite constants $G_{\boldsymbol{S}}$ and $G_{\boldsymbol{Z}}$ such that the corresponding partial gradient norms are bounded by $G_{\boldsymbol{S}}$ and $G_{\boldsymbol{Z}}$ over all admissible groupings. Applying Assumption 4.1 first to the scale update induced by the merge and then to the zero-point update gives

$$\varepsilon_t \le G_{\boldsymbol{S}}\delta_{\boldsymbol{S},t} + \frac{L_{\boldsymbol{S}}}{2}\delta^2_{\boldsymbol{S},t} + G_{\boldsymbol{Z}}\delta_{\boldsymbol{Z},t} + \frac{L_{\boldsymbol{Z}}}{2}\delta^2_{\boldsymbol{Z},t}. \tag{98}$$

Eq. (96) and $d_t \le d_{\max}$ further imply

$$\delta_{\boldsymbol{S},t} \le \sqrt{d_{\max}} B_{\boldsymbol{S}}, \qquad \delta_{\boldsymbol{Z},t} \le \sqrt{d_{\max}} B_{\boldsymbol{Z}}. \tag{99}$$

Therefore, every merge satisfies the uniform bound

$$\varepsilon_t \le e_m := G_{\boldsymbol{S}} \sqrt{d_{\max}} B_{\boldsymbol{S}} + G_{\boldsymbol{Z}} \sqrt{d_{\max}} B_{\boldsymbol{Z}} + \frac{d_{\max}}{2} \left(L_{\boldsymbol{S}} B_{\boldsymbol{S}}^2 + L_{\boldsymbol{Z}} B_{\boldsymbol{Z}}^2 \right) < \infty. \tag{100}$$

This bound is structural: merge is an averaging projection rather than an extrapolation, and its loss effect is controlled entirely by the disagreement between the adjacent groups being merged.

If regrouping is performed every K iterations and $K \ge cT$ for a constant $c > 0$, the number of regrouping operations, and hence the number of merge operations, satisfies

$$N_T^{\text{merge}} \le \left\lceil \frac{T}{K} \right\rceil \le \left\lceil \frac{1}{c} \right\rceil. \tag{101}$$

Combining this inequality with Eq. (100) yields

$$\mathcal{E}_T \le N_T^{\text{merge}} e_m \le \left\lceil \frac{1}{c} \right\rceil e_m = O(1), \tag{102}$$

which, together with Theorem 4.7, shows that $K \ge cT$ is one sufficient condition for preserving the fixed-grouping rate. Since the optimization budget T is specified in advance, the adaptation interval K can be chosen as a function of T. Thus, $K = \Omega(T)$ specifies an asymptotic scheduling condition rather than a fixed numerical value. The experimental choice $K = 125$ applies to the reported finite training budgets and does not imply that K remains fixed as T increases.

More generally, Eq. (98) shows that $K = \Theta(T)$ is not necessary for rate preservation. Let $\mathcal{M}_T$ be the set of iterations at which merge is performed. If

$$\sum_{t \in \mathcal{M}_T} \mathbb{E}\left[\delta_{\boldsymbol{S},t} + \delta_{\boldsymbol{Z},t} + \delta_{\boldsymbol{S},t}^2 + \delta_{\boldsymbol{Z},t}^2 \right] = O(1), \tag{103}$$

then Eq. (98) directly implies $\mathcal{E}_T = O(1)$, even when K is fixed.

D.5 Proof of Theorem 4.8

Under Gaussian noise, similar to MeZO (Malladi et al., 2023), we consider the following modified version of the local r-effective rank assumption, where the upper bound on the Hessian is measured over a ball with a radius twice as large as the one in Assumption 4.2.

Assumption D.11 (Local r-effective rank, Gaussian). Let

$$G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t) = \max_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} \left\| \nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \{(\boldsymbol{x}, \boldsymbol{y})\}) \right\|, \tag{104}$$

$$G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) = \max_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} \left\| \nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \{(\boldsymbol{x}, \boldsymbol{y})\}) \right\|. \tag{105}$$

There exist matrices $\boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)$ and $\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$ such that:

1. For all $\boldsymbol{S}$ or $\boldsymbol{Z}$ such that

$$\|\boldsymbol{S} - \boldsymbol{S}_t\| \le 2\eta d\, G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t), \|\boldsymbol{Z} - \boldsymbol{Z}_t\| \le 2\eta d\, G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t),$$

   we have,

$$\nabla_{\boldsymbol{S}}^2 \mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}) \preceq \boldsymbol{H}(\boldsymbol{S}_t, \boldsymbol{Z}_t), \nabla_{\boldsymbol{Z}}^2 \mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}) \preceq \boldsymbol{H}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t).$$

2. The effective rank of $\boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)$ and $\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$, i.e.,

$$\frac{\operatorname{tr}(\boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t))}{\|\boldsymbol{H}_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|_{\text{op}}}, \frac{\operatorname{tr}(\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))}{\|\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|_{\text{op}}},$$

   is at most r.

Assumption D.12. $G_1(\boldsymbol{S}, \boldsymbol{Z}) \le M_1$ and $G_2(\boldsymbol{S}, \boldsymbol{Z}) \le M_2$.

Lemma D.13. *Let $\boldsymbol{u}_i \sim \mathcal{N}(0, \boldsymbol{I})$ i.i.d. Then,*

$$\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\right] = \left(1+\frac{1}{n}\right)\cdot\left(\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \frac{1}{b}\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z})\right) + \frac{1}{n}\boldsymbol{I}\cdot\left(\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\|^2 + \frac{1}{b}\operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z}))\right), \tag{106}$$

$$\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\right] = \left(1+\frac{1}{n}\right)\cdot\left(\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \frac{1}{b}\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z})\right) + \frac{1}{n}\boldsymbol{I}\cdot\left(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\|^2 + \frac{1}{b}\operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S},\boldsymbol{Z}))\right). \tag{107}$$

Proof. As in the proof of Lemma D.4, we have that in the $\epsilon \to 0$ limit.

$$\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\right] = \frac{1}{b^2n^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\sum_{i,j\in[n]}\mathbb{E}\left[(\boldsymbol{z}_i\boldsymbol{z}_i^\top\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\}))(\boldsymbol{z}_j\boldsymbol{z}_j^\top\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\}))^\top\right]. \tag{108}$$

Let v_1, v_2 be two arbitrary vectors.when $i \neq j$, we have that,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_i,\boldsymbol{u}_j}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top \boldsymbol{u}_j\boldsymbol{u}_j^\top\right] &= \mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top\right]\mathbb{E}_{\boldsymbol{u}_j}\left[\boldsymbol{u}_j\boldsymbol{u}_j^\top\right]\\ &= \mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top\right] I\\ &= v_1 v_2^\top.\end{aligned} \tag{109}$$

and when $i = j$, we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_i}\left[\boldsymbol{u}_i\boldsymbol{u}_i^\top v_1 v_2^\top \boldsymbol{u}_i\boldsymbol{u}_i^\top\right] &= \mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{u}^{\otimes 4}\right](v_1,v_2)\\ &= 3\mathrm{Sym}(I^{\otimes 2})(v_1,v_2)\\ &= v_1^\top v_2\cdot I + 2v_1v_2^\top.\end{aligned} \tag{110}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\right]\\ &= \frac{1}{b^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\left(\frac{n-1}{n}+\frac{2}{n}\right)\mathbb{E}\left[\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\})^\top\right]\\ &+ \frac{1}{n}\cdot\mathbb{E}\left[\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(x_1,\boldsymbol{y}_1)\})^\top\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\})\right]\boldsymbol{I}.\end{aligned} \tag{111}$$

In the proof of Lemma D.4, we show that,

$$\begin{aligned}&\frac{1}{b^2}\sum_{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2)\in\mathcal{B}}\mathbb{E}\left[\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_1,\boldsymbol{y}_1)\})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\{(\boldsymbol{x}_2,\boldsymbol{y}_2)\})^\top\right]\\ &= \nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \frac{1}{b}\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z}).\end{aligned} \tag{112}$$

Plugging this yields,

$$\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z};\mathcal{B})^\top\right] = \left(\frac{n+1}{n}\right)\cdot\left(\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})^\top + \frac{1}{b}\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z})\right) + \frac{1}{n}\boldsymbol{I}\cdot\left(\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S},\boldsymbol{Z})\|^2 + \frac{1}{b}\operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S},\boldsymbol{Z}))\right). \tag{113}$$

Similarly, we can obtain the result for $\boldsymbol{S}$.

Lemma D.14 (Standard $\mathcal{X}^2$-tail bound (Malladi et al., 2023)). *Let $\mathcal{Z}$ be a $\mathcal{X}^2$ random variable with $\mathcal{K}$ degrees of freedom. Then,*

$$\Pr[\mathcal{Z} \geq \mathcal{K} + e] \leq \exp\left(-\min\left(\frac{e^2}{16k}, \frac{e}{16}\right)\right). \tag{114}$$

Lemma D.15. *Under Assumption 4.1 and Assumption D.11, if $\boldsymbol{S}$, $\boldsymbol{Z}$, are all updated at each iteration using ZO-SGD with the n-SPSA estimator and a minibatch of size b, then there exist $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$ such that the expected decrease in the loss can be bounded as:*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\right] \leq -\eta\left(\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 + \|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|^2\right) \\
&+ \frac{\eta^2}{2}\left(L_{\boldsymbol{S}}\gamma_1\mathbb{E}\left[\|\nabla_{\boldsymbol{S}}\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2\right] + L_{\boldsymbol{Z}}\gamma_2\mathbb{E}\left[\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2\right]\right) \\
&+ \eta^2\exp(-\Omega(nd))\left(L_{\boldsymbol{Z}}G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}}G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right).
\end{aligned} \tag{115}$$

Proof. Let $\mathcal{A}$ be the event that $\|\boldsymbol{Z} - \boldsymbol{Z}_t\| \leq 2\eta d G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)$. On $\mathcal{A}$, we have that,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) \leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^\top\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \\
+ \frac{1}{2}\eta^2\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}).
\end{aligned} \tag{116}$$

Under Assumption 4.1, we have,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) \leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^\top\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \\
+ \frac{1}{2}\eta^2\mathcal{L}_{\boldsymbol{Z}}\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2.
\end{aligned} \tag{117}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1})|\boldsymbol{Z}_t\right] \\
&\leq \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta\left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&+ \frac{\eta^2}{2}\left\langle\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top \cdot \mathbf{1}\left(\mathcal{A}\right)\right]\right\rangle \\
&+ \frac{1}{2}\eta^2\mathcal{L}_{\boldsymbol{Z}}\mathbb{E}\left[\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \cdot \mathbf{1}\left(\neg\mathcal{A}\right)\|^2\right] \\
&= \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) - \eta\left\|\nabla_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\right\|^2 \\
&+ \frac{\eta^2}{2}\left\langle\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top\right]\right\rangle \\
&+ \frac{\eta^2}{2}\left\langle\mathcal{L}_{\boldsymbol{Z}}\boldsymbol{I} - \boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top \cdot \mathbf{1}\left(\neg\mathcal{A}\right)\right]\right\rangle.
\end{aligned} \tag{118}$$

The latter term can be bounded as follows:

$$\begin{aligned}
&\frac{\eta^2}{2}\left\langle\mathcal{L}_{\boldsymbol{Z}}\boldsymbol{I} - \boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top \cdot \mathbf{1}\left(\neg\mathcal{A}\right)\right]\right\rangle \\
&\leq \eta^2\mathcal{L}_{\boldsymbol{Z}}\mathbb{E}\left[\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2 \cdot \mathbf{1}\left(\neg\mathcal{A}\right)\right] \\
&\leq \eta^2\mathcal{L}_{\boldsymbol{Z}}\mathbb{E}\left[\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^4\right]^{\frac{1}{2}} Pr\left[\neg\mathcal{A}\right]^{\frac{1}{2}}.
\end{aligned} \tag{119}$$

The gradient estimate $\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})$ satisfies,

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\nabla}_{\boldsymbol{Z}}\mathcal{L}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\|^4\right] &\leq \frac{1}{n}\sum\mathbb{E}\left[\left|\boldsymbol{u}_i^\top\nabla\mathcal{L}_{\boldsymbol{Z}}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\right|^4 \cdot \|\boldsymbol{u}_i\|^4\right] \\
&\leq \mathbb{E}\left[\left|\boldsymbol{u}^\top\nabla\mathcal{L}_{\boldsymbol{Z}}(\boldsymbol{S}, \boldsymbol{Z}; \mathcal{B})\right|^4\right]^{\frac{1}{2}}\left[\|\boldsymbol{u}\|^8\right]^{\frac{1}{2}} \\
&\leq \sqrt{105}(d+6)^2 G_2(\boldsymbol{S}, \boldsymbol{Z})^4,
\end{aligned} \tag{120}$$

where we have plugged in explicit formulas for moments of Gaussian and $\mathcal{X}^2$ random variables. Next, note that on the event $\neg\mathcal{A}$, we have,

$$2\eta d G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) \leq \|\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t\| = \eta \left\| \hat{\nabla}_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \right\| \leq \eta \cdot \frac{1}{n} \sum_{i\in[n]} \|\boldsymbol{u}_i\|^2 G(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t). \tag{121}$$

Therefore,

$$\Pr[\neg\mathcal{A}] \leq \Pr\left[\sum_{i\in[n]} \|\boldsymbol{u}_i\|^2 \geq 2nd \right]. \tag{122}$$

Since $\sum_{i\in[n]} \|\boldsymbol{u}_i\|$ is a $\mathcal{X}^2$ random variable with nd degrees of freedom, by Lemma D.14, we thus have that,

$$\Pr[\neg\mathcal{A}] \leq \exp\left(-\frac{nd}{16}\right). \tag{123}$$

So,

$$\begin{aligned} &\frac{\eta^2}{2} \left\langle \mathcal{L}_{\boldsymbol{Z}} \boldsymbol{I} - \boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t), \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B}) \hat{\nabla}_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})^\top \cdot \mathbf{1}(\neg\mathcal{A}) \right] \right\rangle \\ &\leq \eta^2 \mathcal{L}_{\boldsymbol{Z}} 105^{\frac{1}{4}} (d+6) G(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 \exp\left(-\frac{nd}{32}\right) \\ &= \eta^2 \mathcal{L}_{\boldsymbol{Z}} G(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 \exp\left(-\Omega(nd)\right). \end{aligned} \tag{124}$$

Plugging in Eq. *equation* 113, along with the fact that $\|\boldsymbol{H}_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|_{op} \leq L_{\boldsymbol{Z}}$ and $tr\left(\|\boldsymbol{H}(W_t, A_{t+1}, B_{t+1})\|_{op}\right) \leq L_{\boldsymbol{Z}} r$. Letting $\gamma_2 = \frac{r+n+1}{n}$, we have,

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) | \boldsymbol{Z}_t\right] - \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) &\leq -\eta \|\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|^2 \\ &\quad + \frac{L_{\boldsymbol{Z}}}{2} \eta^2 \gamma_2 \mathbb{E}\left[\|\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2\right] \\ &\quad + \eta^2 L_{\boldsymbol{Z}} G(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 \exp(-\Omega(nd)). \end{aligned} \tag{125}$$

Similarly, we have,

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t) | \boldsymbol{S}_t\right] - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t) &\leq -\eta \|\nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 + \frac{L_{\boldsymbol{S}}}{2} \eta^2 \gamma_1 \mathbb{E}\left[\|\nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2\right] \\ &\quad + \eta^2 L_{\boldsymbol{S}} G(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2 \exp(-\Omega(nd)). \end{aligned} \tag{126}$$

By combining Eq. (125) and (126) and taking expectations, we obtain,

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1}) - \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)] &\leq -\eta \left(\|\nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 + \|\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)\|^2 \right) \\ &\quad + \frac{\eta^2}{2} \big(L_{\boldsymbol{S}} \gamma_1 \mathbb{E}\left[\|\nabla_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t; \mathcal{B})\|^2 \right] \\ &\quad + L_{\boldsymbol{Z}} \gamma_2 \mathbb{E}\left[\|\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t; \mathcal{B})\|^2 \right] \big) \\ &\quad + \eta^2 \exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2 \right). \end{aligned} \tag{127}$$

Theorem D.16. *If $\boldsymbol{S}$, $\boldsymbol{Z}$, are all updated at each iteration using ZO-SGD with the n-SPSA estimator and a minibatch of size b, and the learning rate is chosen as $\eta = O(\frac{1}{\sqrt{Td}})$, then there exist $\gamma_1 = \Theta(r/n)$ and $\gamma_2 = \Theta(r/n)$, such that the following inequality holds:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq O\left(\frac{\sqrt{d}}{\sqrt{T}} + \frac{1}{Tb} + \frac{\sqrt{d}}{\sqrt{T}b} + \frac{\exp(-\Omega(nd))}{\sqrt{Td}}\right).$$

Proof. It is readily observed that, in the case of Gaussian noise, Lemma D.15 contains an additional term $\eta^2 \exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right)$ compared with Lemma D.7. Thus, we can derive the following result by following the proof strategy of Theorem 4.7:

$$\begin{aligned}\frac{\eta}{2(1+C_1)}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq \mathbb{E}[\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t) - \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1})] + \frac{\eta^3 C_2}{2b(1+C_1)}\operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ + \frac{\eta^2}{2b}\Big[L_{\boldsymbol{S}}\gamma_1 \operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2 \operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\Big] \\ + \eta^2 \exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right).\end{aligned} \tag{128}$$

By dividing both sides of the inequality by $\frac{\eta}{2(1+C_1)}$, we have,

$$\begin{aligned}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq \frac{2(1+C_1)}{\eta}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t) - \mathcal{L}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_{t+1})] + \frac{\eta^2 C_2}{b}\operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ + \frac{\eta(1+C_1)}{b}\Big[L_{\boldsymbol{S}}\gamma_1 \operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2 \operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\Big] \\ + 2\eta(1+C_1) \exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right).\end{aligned} \tag{129}$$

Summing both sides of the inequality with respect to t from 0 to T-1, we have,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq \frac{2(1+C_1)}{\eta T}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}(\boldsymbol{S}_T, \boldsymbol{Z}_T)] + \frac{1}{T}\sum_{t=0}^{T-1}\frac{\eta^2 C_2}{b}\operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) \\ + \sum_{t=0}^{T-1}\frac{\eta(1+C_1)}{bT}\Big[L_{\boldsymbol{S}}\gamma_1 \operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2 \operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\Big] \\ + \sum_{t=0}^{T-1}\frac{2\eta(1+C_1)}{T}\exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right).\end{aligned} \tag{130}$$

Since $\mathcal{L}(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}(\boldsymbol{S}_T, \boldsymbol{Z}_T) \leq \mathcal{L}(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}^*$, where $\mathcal{L}^* = \min_{\boldsymbol{S},\boldsymbol{Z}} \mathcal{L}(\boldsymbol{S}, \boldsymbol{Z})$, taking expectations on both sides of the inequality yields,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq \frac{2(1+C_1)}{\eta T}\mathbb{E}[\mathcal{L}(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}^*] + \frac{1}{T}\sum_{t=0}^{T-1}\frac{\eta^2 C_2}{b}\mathbb{E}\Big[\operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t))\Big] \\ + \sum_{t=0}^{T-1}\frac{\eta(1+C_1)}{bT}\mathbb{E}\Big[L_{\boldsymbol{S}}\gamma_1 \operatorname{tr}(\Sigma_{\boldsymbol{S}}(\boldsymbol{S}_t, \boldsymbol{Z}_t)) + L_{\boldsymbol{Z}}\,\gamma_2 \operatorname{tr}(\Sigma_{\boldsymbol{Z}}(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t))\Big] \\ + \sum_{t=0}^{T-1}\frac{2\eta(1+C_1)}{T}\exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} G_2(\boldsymbol{S}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2\right).\end{aligned} \tag{131}$$

Due to Assumption 4.4 and D.12, we obtain,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{S}_t, \boldsymbol{Z}_t)\|^2 \leq \frac{2(1+C_1)}{\eta T}(\mathcal{L}(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}^*) + \frac{\eta^2 C_2 d\lambda_{\boldsymbol{S}}}{b} \\ + \frac{d\eta(1+C_1)}{b}\mathbb{E}\Big[L_{\boldsymbol{S}}\gamma_1\lambda_{\boldsymbol{S}} + L_{\boldsymbol{Z}}\,\gamma_2\lambda_{\boldsymbol{Z}}\Big] \\ + 2\eta(1+C_1)\exp(-\Omega(nd)) \left(L_{\boldsymbol{Z}} M_2^2 + L_{\boldsymbol{S}} M_1^2\right).\end{aligned} \tag{132}$$

By choosing $\eta = O(\frac{1}{\sqrt{Td}})$, then the convergence rate as shown in above is $O(\frac{\sqrt{d}}{\sqrt{T}} + \frac{1}{Tb} + \frac{\sqrt{d}}{\sqrt{Tb}} + \frac{\exp(-\Omega(nd))}{\sqrt{Td}})$.

Since the optimization budget T is known in advance, we choose $K = \Omega(T)$. Under a uniform per-merge loss bound, this ensures $\mathcal{E}_T = O(1)$. By choosing $\eta = \Theta(1/\sqrt{Td_{\max}})$, the convergence rate becomes $O(\frac{\sqrt{d_{\max}}}{\sqrt{T}} + \frac{1}{Tb} + \frac{\sqrt{d_{\max}}}{\sqrt{Tb}} + \frac{\exp(-\Omega(nd_{\min}))}{\sqrt{Td_{\max}}})$.

D.6 Adaptive Grouping under Gaussian Perturbations

We finally extend the Gaussian analysis above to the time-varying grouping sequence of Algorithm 1. Conditioned on $\mathcal{F}_t$, the active grouping is fixed, so the Gaussian one-step bound established in the proof of Theorem 4.8 applies with d replaced by d_t. Relative to Eq. (85), this introduces the additional term

$$\eta^2 \exp(-\Omega(nd_t)) \left(L_{\boldsymbol{Z}} G_2(\widetilde{\boldsymbol{S}}_{t+1}, \boldsymbol{Z}_t)^2 + L_{\boldsymbol{S}} G_1(\boldsymbol{S}_t, \boldsymbol{Z}_t)^2 \right). \tag{133}$$

The regrouping step is handled by exactly the same pathwise inequality in Eq. (86), so its only effect is again the cumulative error $\mathcal{E}_T$. Under the uniform Gaussian assumptions and the bounds $G_1 \le M_1$, $G_2 \le M_2$, we have $d_t \ge d_{\min}$ and hence

$$\exp(-\Omega(nd_t)) \le \exp(-\Omega(nd_{\min})). \tag{134}$$

Following Eq. (90), the extra Gaussian contribution after summing and dividing by the descent coefficient is therefore bounded by

$$2\eta(1+\bar{C}_1) \exp(-\Omega(nd_{\min})) \left(L_{\boldsymbol{Z}} M_2^2 + L_{\boldsymbol{S}} M_1^2 \right). \tag{135}$$

Choosing $\eta = O(1/\sqrt{Td_{\max}})$ and $K = \Omega(T)$ yields

$$\begin{aligned} &\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla \mathcal{L}_t(\boldsymbol{S}_t, \boldsymbol{Z}_t) \right\|^2 \\ &\le O\left((\mathcal{L}_0(\boldsymbol{S}_0, \boldsymbol{Z}_0) - \mathcal{L}^* + 1) \sqrt{\frac{d_{\max}}{T}} + \frac{1}{Tb} + \frac{\sqrt{d_{\max}}}{\sqrt{T}b} + \frac{\exp(-\Omega(nd_{\min}))}{\sqrt{Td_{\max}}} \right), \end{aligned} \tag{136}$$

which proves Theorem 4.8.