Rethinking Gradient Approximation in Quantization: A Zeroth-Order Expectation Perspective

Wenkang Wang[1] **Jia Zhang**[1] **Zelin Wei**[1] **Dongxu Liu**[1]
Jiabao Li[1] **Wanli Shi**[1] **Bin Gu**[1*]
[1]School of Artificial Intelligence, Jilin University, China
{wangwk25, jiazhang25, weizl26, liudx9924, lijb26}@mails.jlu.edu.cn
{wanli_shi, gubin}@jlu.edu.cn

Abstract

As large language models (LLMs) continue to scale, quantization has become a key technique for efficient deployment. However, multi-bit quantization employs non-differentiable rounding, hindering gradient optimization. Existing methods rely on heuristic surrogate gradients (e.g., STE), which work empirically but lack a unified theory. To address this challenge, we propose the Zeroth-Order Expectation Gradient (ZOE-Grad), a zeroth-order expectation view of quantization surrogate gradients. Specifically, we establish an equivalence between the expectations of zeroth-order gradient estimators and surrogate gradients, providing a unified zeroth-order interpretation of widely used surrogates. Under this framework, STE corresponds to a degenerate, discontinuous perturbation that ignores the local geometry of quantization boundaries, limiting its expressiveness. In contrast, continuous perturbation-based surrogate gradients capture boundary-local information and produce smoother, more structured gradients. We further establish bounded-error convergence guarantees for the quantized objective. Through simulations under multiple perturbation distributions, we verify that the proposed framework accurately captures the relationship between the expectations of zeroth-order gradient estimators and surrogate gradients. Furthermore, extensive experiments on OPT-1.3B, OPT-6.7B, LLaMA-2-7B, and Qwen3-8B show that continuous ZOE-Grad surrogates consistently outperform STE across diverse LLM architectures. Our code is available at `https://github.com/wangwk699/ZOE_Grad`.

1 Introduction

LLMs have driven advances in natural language understanding and generation across research and industry [40, 3, 45]. However, larger models improve performance but incur high deployment-time computation and memory costs, which limit accessibility on constrained hardware [2]. To mitigate these challenges, quantization reduces memory usage and accelerates inference while largely preserving model performance [33, 38, 11].

The current quantization methods for LLMs can be broadly categorized into Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). PTQ methods [33, 23, 19] quantize pre-trained LLMs without retraining, enabling efficient deployment but often degrading accuracy at very low bit-widths. QAT methods [9, 22, 16] improve accuracy retention by incorporating quantization during training, but they incur higher computational overhead [16]. Despite their differences, both PTQ and QAT face a fundamental challenge: rounding in multi-bit quantization is non-differentiable, with zero or undefined gradients almost everywhere [20]. As a result, surrogate gradients are required when training quantized models or optimizing quantization parameters.

*Corresponding author.

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

Representative surrogate gradients include the straight-through estimator (STE) [1], piece-wise linear (PWL) [13], and hyperbolic tangent gradient estimator (HTGE) [29]. Among them, STE is the de facto default in PTQ and QAT due to its simplicity. Prior studies have already identified limitations of the constant, position insensitive STE gradient. For example, EWGS [17] uses the error between latent and quantized values to rescale gradients, while Nagel et al. [26] show that constant STE gradients can cause latent weights to oscillate across adjacent quantization boundaries. These observations establish that position relative to quantization boundaries matters, but they do not provide a unified constructive account of how a surrogate gradient arises from an underlying perturbation distribution or how a desired boundary dependent surrogate can be associated with such a distribution. This raises a fundamental question: is there a unified theoretical framework for systematically understanding and constructing surrogate gradients in quantized models?

To address this issue, we propose ZOE-Grad, a theoretical framework that interprets surrogate gradients in quantized optimization from a zeroth-order expectation perspective. Specifically, ZOE-Grad establishes constructive mappings in both directions between perturbation distributions and a class of boundary dependent surrogate gradients, identifies STE exactly as the expected zeroth-order estimator induced by a fixed magnitude Rademacher perturbation, and connects the perturbation choice to the approximation gap between the smoothed and original quantized objectives. This view provides a unified explanation of how perturbation support and density shape local surrogate gradients while clarifying the information limitation of STE. Building on this framework, we further provide convergence guarantees showing that optimization with surrogate gradients converges to solutions with bounded suboptimality for the original quantized objective. Overall, ZOE-Grad characterizes the surrogate design space and provides structural guidance for constructing and selecting candidate surrogate gradients.

The main contributions of this paper are summarized as follows:

- We propose ZOE-Grad, a unified zeroth-order expectation framework that establishes constructive mappings between perturbation distributions and boundary dependent surrogate gradients in both directions. The framework identifies STE as the expectation induced by a fixed magnitude Rademacher perturbation, providing a principled characterization of the surrogate design space.
- We establish a unified convergence analysis for surrogate-gradient optimization under Lipschitz smoothness and Polyak–Łojasiewicz conditions, showing that it converges to solutions with bounded suboptimality for the original quantized objective and thereby characterizing the theoretical impact of surrogate-gradient approximation.
- We empirically validate ZOE-Grad through simulations over multiple perturbation distributions, confirming the correspondence between zeroth-order expectation gradients and surrogate gradients. Experiments on OPT-1.3B, OPT-6.7B, LLaMA-2-7B, and Qwen3-8B further show that continuous ZOE-Grad surrogates consistently outperform STE.

2 Related Work

2.1 Quantization of LLMs

Quantization has shown great promise in reducing both the computational and memory overheads in LLMs [33]. Typically, it is divided into two categories: post-training quantization (PTQ) and quantization-aware training (QAT). Additionally, zero-shot quantization methods have been proposed [46], aiming to execute quantization without retraining or fine-tuning the model. Considering the quantized object, existing LLM quantization can be broadly classified into two categories: weight-only quantization and weight-activation quantization. Weight-only quantization focuses on converting the model's weights into lower bit values. For example, GPTQ [11] achieves 3/4-bit quantization through block reconstruction, while SpQR [8] protects critical weights via mixed-precision quantization, and Qlora [7] recovers the performance of quantized models via parameter-efficient fine-tuning. Weight-activation quantization, on the other hand, compresses both the model's weights and activations. Techniques such as SmoothQuant [38] and Outlier Suppression [37] achieve W8A8 quantization by managing activation outliers, while RPTQ [42] and LLM-QAT [21] achieve W4A4 quantization.

2.2 Surrogate Gradients

The rounding operation in quantization is non-differentiable, which prevents standard backpropagation from updating model parameters. A common remedy is to optimize the quantized model parameters using heuristic surrogate gradients. The most widely adopted surrogate gradient is the STE [1, 41], which replaces the derivative of round$(\cdot)$ with a constant one. While widely used, it ignores quantization effects and can become unstable at very low bit-widths due to amplified approximation errors. EWGS [17] rescales gradients using the latent to quantized error and gradient direction. Nagel et al. [26] show that constant STE gradients can cause oscillations across adjacent quantization boundaries and discusses position dependent gradient rescaling. The PWL estimator [13] propagates non-zero gradients only within the representable range of the quantizer. Differentiable Soft Quantization (DSQ) [12] utilizes a series of hyperbolic tangent functions to gradually approximate standard quantization. Magnitude-aware derivative (MAD) [32] matches STE within the representable range and applies a reciprocal form outside it. The HTGE [29] uses tanh to approximate the gradient of the rounding operator. These methods motivate surrogate gradients that retain more quantization information, while ZOE-Grad provides a unified and constructive bidirectional view by mapping perturbation distributions to surrogate gradients and, for a class of boundary dependent surrogates, mapping them back to perturbation distributions.

3 Preliminary

3.1 Problem Formulation

In this work, we focus on weight-only quantization, and consider the widely adopted symmetric uniform quantization scheme for LLMs [44]. Specifically, given the full-precision model weights $\mathbf{W}_f$, the quantization and dequantization processes are defined as:

$$\mathbf{W}_{\mathrm{INT}} = \text{clamp}\left(\text{round}\left(\frac{\mathbf{W}_f}{\mathbf{s}}\right), -2^{N-1}, 2^{N-1}-1\right), \tag{1}$$

$$\widehat{\mathbf{W}}_f = \mathbf{s} \odot \mathbf{W}_{\mathrm{INT}}, \tag{2}$$

where $\mathbf{W}_{\mathrm{INT}}$ and $\hat{\mathbf{W}}_f$ denote the quantized and dequantized weights, respectively. Here, $\mathbf{s}$ is the scaling factor, N is the target bit-width, $\mathrm{clamp}(\cdot)$ denotes the clipping operator, $\mathrm{round}(\cdot)$ is the rounding function, and $\odot$ represents element-wise multiplication. In weight-only quantization, the choice of the scaling factor $\mathbf{s}$ is critical for preserving model accuracy. Common strategies include MinMax (or dynamic) quantization [15], which determines $\mathbf{s}$ based on the range of actual inputs at each layer. In contrast, several recent works treat the scaling factor $\mathbf{s}$ as a learnable parameter and optimize it via gradient-based methods using heuristic surrogate gradients [9, 33]. Consider a neural network with objective function $L\left(\mathbf{W}_f\right)$. For notational simplicity, we omit the network inputs. Following prior work, we treat the scale factor $\mathbf{s}$ as the optimization variable. Thus, the resulting optimization problem can be expressed as:

$$\min_{\mathbf{s}\in\mathbb{R}^d} \mathcal{L}(\mathbf{s}) = L\left(\hat{\mathbf{W}}_f\right). \tag{3}$$

3.2 Zeroth-Order Optimization

When optimizing the objective $\mathcal{L}(\mathbf{s})$ in Eq. (3), the rounding function in Eq. (1) obstructs gradient backpropagation. To address this issue, we analyze it using zeroth-order techniques, a widely used class of gradient-free methods that has been well studied in neural networks [24, 25]. Because the rounding operator is applied element-wise to the weight tensor, it suffices to study surrogate gradients through the corresponding scalar function on a single weight entry. Specifically, for a scalar $\delta > 0$, the two-point gradient estimator [27] of the rounding function r is defined as:

$$G_r^2(u; z, \delta) = \frac{r(u+\delta z) - r(u-\delta z)}{2\delta} z, \tag{4}$$

where z is a scalar random perturbation drawn from a symmetric, zero-mean distribution λ with a finite second moment.

3.3 Boundary Points

Notably, the rounding function has discontinuities at all half-integer values, which we refer to as boundary points and denote by $\mathcal{B} = \{b_k = k + \frac{1}{2} : k \in \mathbb{Z}\}$. Furthermore, we introduce the counting function

$$M(u, \delta z) := \#\big(\mathcal{B} \cap (u - \delta z, u + \delta z)\big) \quad (z \geq 0), \tag{5}$$

which counts the number of boundary points crossed within a symmetric interval centered at u. One can verify that

$$r(u + \delta z) - r(u - \delta z) = M(u, \delta z) \qquad (z \geq 0), \tag{6}$$

and for a more general z

$$r(u + \delta z) - r(u - \delta z) = \operatorname{sign}(z) M\left(u, \delta |z|\right). \tag{7}$$

4 The Proposed Framework: ZOE-Grad

4.1 Zeroth-Order Expectation Gradient

Substituting Eq. (7) into Eq. (4) and taking the expectation over the random perturbation yields

$$\mathbb{E}_{z\sim\lambda}\left[G_r^2(u; z, \delta)\right] = \frac{1}{2\delta} \int_{-\infty}^{\infty} |z| M\left(u, \delta|z|\right) \lambda(z) dz. \tag{8}$$

Furthermore, assuming that the distribution λ is even, we have

$$\mathbb{E}_{z\sim\lambda}\left[G_r^2(u; z, \delta)\right] = \frac{1}{\delta} \int_{0}^{\infty} z M(u, \delta z) \lambda(z) dz. \tag{9}$$

From the definition of $M(u, \delta z)$, it follows that the expectation of the zeroth-order gradient estimator for the round function is an even and periodic function with a period of 1. Moreover, we assume bounded perturbations with $|z| \leq C$ for some constant $C > 0$. Under this assumption, we define the local index set

$$\mathcal{K}_u := \{k \in \mathbb{Z} : |u - b_k| < \delta C\}. \tag{10}$$

Since the boundary points in $\mathcal{B}$ are spaced by 1, the interval $(u - \delta C, u + \delta C)$ contains only finitely many elements of $\mathcal{B}$, and hence $\mathcal{K}_u$ is finite. Then, for $z > 0$, $M(u, \delta z)$ can be further expressed as:

$$M(u, \delta z) = \sum_{k \in \mathcal{K}_u} \mathbf{1}_{\{|u - b_k| < \delta z\}}. \tag{11}$$

With this definition, the expectation of the zeroth-order estimator can be written as a finite sum:

$$\mathbb{E}_{z\sim\lambda}[G_r^2(u; z, \delta)] = \frac{1}{\delta} \sum_{k \in \mathcal{K}_u} \int_{|u - b_k|/\delta}^{C} z\, \lambda(z)\, dz. \tag{12}$$

This structure highlights an important locality property: only quantization boundaries within a neighborhood of radius δC around u contribute to the gradient estimate.

4.2 Connection to Surrogate Gradients

In this section, we show that the expectation of the zeroth-order estimator in Eq. (12) can be expressed in an equivalent surrogate gradient form.

Theorem 4.1. *Let λ be a distribution supported on $[-C, C]$ with probability density function $\lambda(z)$ (so that $\lambda(z) = 0$ for $|z| > C$). Assume the integrals $\int_0^C z\lambda(z)dz$ and $\int_0^C z^2\lambda(z)dz$ exist and are finite. Define the constant:*

$$c := 2 \int_0^C z^2 \lambda(z) dz. \tag{13}$$

Define a new distribution $\tilde{\lambda}$ with probability density function:

$$\tilde{\lambda}(x) = \begin{cases} \dfrac{1}{c} \displaystyle\int_{|x|}^{C} z\lambda(z)dz, & |x| \leq C, \\ 0, & |x| > C. \end{cases} \tag{14}$$

Let

$$S(u) := \mathbb{E}_{z\sim\tilde{\lambda}}\left[c\sum_{k\in\mathcal{K}_u} h(u-b_k+\delta z)\right], \tag{15}$$

where $h(\cdot)$ denotes the Heaviside function. Then,

$$\mathbb{E}_{z\sim\lambda}\left[G_r^2(u;z,\delta)\right] = \frac{d}{du}S(u). \tag{16}$$

This theorem shows that the zeroth-order expectation gradient of rounding can be interpreted as the derivative of a smoothed surrogate objective defined over nearby quantization boundaries.

Proposition 4.2. ***(Normalization for smoothed rounding.)*** *Let $g(u) = \mathbb{E}_{z\sim\lambda}[G_r^2(u;z,\delta)]$ denote the raw expected gradient. For any symmetric, zero-mean perturbation distribution λ with finite variance, $g(u)$ satisfies*

$$\int_k^{k+1} g(u)\,du = c, \qquad c := \mathbb{E}_\lambda[z^2] = 2\int_0^C z^2\lambda(z)\,dz, \qquad k\in\mathbb{Z}. \tag{17}$$

This identity implies that the antiderivative $\hat{r}(u) = \int_0^u g(t)\,dt$ increases by c over each unit interval, rather than matching the unit jump of $\mathrm{round}(u)$. We therefore define the normalized surrogate gradient as

$$g_\delta(u) := \frac{1}{c}g(u) = \frac{1}{c}\mathbb{E}_{z\sim\lambda}\left[G_r^2(u;z,\delta)\right]. \tag{18}$$

Then $\int_k^{k+1} g_\delta(u)\,du = 1$, and the corresponding normalized smoothed rounding function is

$$\hat{r}(u) = \int_0^u g_\delta(t)\,dt. \tag{19}$$

Thus, in practical gradient updates, we use $g_\delta(u)$ instead of $g(u)$. The normalization is trivial when $c = 1$, such as for the uniform distribution on $[-\sqrt{3},\sqrt{3}]$, but is essential for distributions with $c \neq 1$, such as the truncated normal distribution.

Next, we present an application of Theorem 4.1 under the uniform distribution, and applications to additional distributions are detailed in Appendix A.3.

Corollary 4.3. *Assuming that the distribution $\lambda(z)$ is uniform on $[-\sqrt{3},\sqrt{3}]$ (variance $c = 1$), then according to Theorem 4.1, the normalized expectation of the zeroth-order gradient estimator can be expressed as:*

$$g_\delta(u) = \begin{cases} \sum\limits_{k\in\mathcal{K}_u} \frac{1}{4\sqrt{3}\delta}\left(3 - \frac{(u-b_k)^2}{\delta^2}\right), & \exists k, s.t. |u-b_k| < \sqrt{3}\delta, \\ 0, & \textit{otherwise.} \end{cases} \tag{20}$$

We further assume $2\sqrt{3}\delta \leq 1$, which implies that $\mathcal{K}_u$ contains at most one boundary point. Let $s(u)$ denote the boundary point closest to u. Then we have:

$$g_\delta(u) = \begin{cases} \frac{1}{4\sqrt{3}\delta}\left(3 - \frac{(u-s(u))^2}{\delta^2}\right), & \textit{if } |u-s(u)| < \sqrt{3}\delta, \\ 0, & \textit{otherwise}. \end{cases} \tag{21}$$

Next, we establish that several commonly used surrogate gradients can be recovered as expectations of zeroth-order estimators under appropriately constructed perturbation distributions.

Theorem 4.4. *Given a surrogate gradient $g(u)$ that admits the representation*

$$g(u) := \sum_{k\in\mathcal{K}_u} \psi(|u-b_k|), \tag{22}$$

where $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a continuously differentiable, non-increasing function satisfying $\psi(\delta C) = 0$. We further assume that $c := -2\delta^2\int_0^C \frac{1}{z}\psi'(z\delta)\,dz < \infty$ and define

$$\lambda(z) = \begin{cases} -\dfrac{\delta^2}{cz}\psi'(z\delta), & 0 < |z| \leq C, \\ 0, & \textit{otherwise}. \end{cases} \tag{23}$$

So that λ is a probability density function. Then, we have:

$$c\,\mathbb{E}_{z\sim\lambda}\left[G_r^2(u;z,\delta)\right] = g(u). \tag{24}$$

This theorem shows that if a surrogate gradient can be expressed as a sum of boundary-dependent terms, one can construct a perturbation distribution under which it equals the expected zeroth-order gradient estimator. Next, we will present several applications of Theorem 4.4.

Corollary 4.5. *We consider the hyperbolic tangent gradient estimator (HTGE), whose surrogate gradient is given by*

$$g(u) = \frac{t}{2}\left[1 - \tanh^2\left(t\left(u - \lfloor u \rfloor - \frac{1}{2}\right)\right)\right]. \tag{25}$$

Note that $u - \lfloor u \rfloor - \frac{1}{2}$ is mathematically equivalent to $u - s(u)$. Therefore, $g(u)$ can be written as $g(u) = \psi(|u - s(u)|)$, which corresponds to a single-boundary formulation. Hence, we set $\delta C = \frac{1}{2}$. When t is sufficiently large, $\psi(\frac{1}{2}) = \frac{t}{2}sech^2\left(\frac{t}{2}\right)$ approaches zero. Therefore, according to Theorem 4.4, the probability density function of the associated perturbation distribution λ is defined as:

$$\lambda(z) = \begin{cases} \frac{\delta^2 t^2}{c} \cdot \frac{sech^2(t\delta z)\tanh(t\delta z)}{z}, & 0 < |z| \le C, \\ \frac{\delta^3 t^3}{c}, & z = 0, \\ 0, & otherwise, \end{cases} \tag{26}$$

where $c = 2\delta^2 t^2 \int_0^{\frac{t}{2}} \frac{sech^2 u \tanh u}{u}\, du$. The detailed proof is provided in Appendix B.2.

Corollary 4.6. *The widely used STE emerges as a degenerate special case of the ZOE-Grad framework. Specifically, it corresponds to the expectation of a zeroth-order estimator induced by a discrete Rademacher perturbation, with $\mathbb{P}(z = 1) = \mathbb{P}(z = -1) = 1/2$ and $\delta = \frac{1}{2}$. Then, by Eq.* (4)*, we have*

$$\mathbb{E}_{z\sim\lambda}\left[G_r^2(u; z, \delta)\right] = \mathbb{E}_{z\sim\lambda}\left[\frac{r(u + \delta z) - r(u - \delta z)}{2\delta}\, z\right] = r\left(u + \frac{1}{2}\right) - r\left(u - \frac{1}{2}\right). \tag{27}$$

The resulting function is illustrated in Figure 2. Notably, the Rademacher perturbation can be viewed as retaining only the sign of an arbitrary continuous perturbation while discarding its magnitude, thereby ignoring positional information relative to quantization boundaries and resulting in a limited zeroth-order gradient estimator. The detailed proof is provided in Appendix B.3.

Remark 4.7. ZOE-Grad characterizes the surrogate design space rather than identifying a universally optimal perturbation distribution. Corollary 4.6 shows that the fixed magnitude Rademacher perturbation underlying STE discards perturbation magnitude information and therefore loses information about distance to nearby quantization boundaries. In contrast, continuous perturbations retain this information and induce gradients that vary with local boundary geometry. This provides structural guidance for considering continuous boundary aware perturbations as a more expressive class of candidates. For a bounded continuous perturbation supported on $[-C, C]$, the single boundary condition $2\delta C \approx 1$ further provides a theory motivated initialization for the perturbation scale δ.

More generally, selecting the optimal perturbation distribution within a parameterized family can itself be formulated as a bilevel optimization problem [10, 28]:

$$\theta^\star \in \arg\min_\theta \mathcal{L}_{\text{val}}\left(\mathbf{s}^\star(\theta)\right), \qquad \mathbf{s}^\star(\theta) \in \arg\min_{\mathbf{s}} \mathcal{L}_{\text{train}}\left(\mathbf{s}; \lambda_\theta\right), \tag{28}$$

where θ parameterizes the family or shape of the perturbation distribution λ_θ, and $\mathbf{s}^\star(\theta)$ is the quantization scale vector obtained by minimizing the training objective under the surrogate induced by λ_θ. Solving this bilevel problem is beyond the scope of this work. In our experiments, we instead use the lightweight pilot fine tuning procedure in Section 6.2.3 to select among the candidate surrogate gradients before full training.

5 Convergence Analysis

In this section, we relate the original quantized objective to the smoothed objective induced by surrogate gradients. Let $\mathcal{L}(\mathbf{s})$ be the objective with the discrete quantizer, and let $\mathcal{L}'(\mathbf{s})$ be its smoothed counterpart obtained by replacing rounding with the smoothed rounding function in Eq. 19.

For each normalized weight u, define its distance to the nearest quantization boundary as $\text{dist}(u, \mathcal{B}) := |u - s(u)|$, where $s(u)$ denotes the closest boundary point. The full notation and technical assumptions are deferred to Appendix C. We first bound the gap between $\mathcal{L}(\mathbf{s})$ and $\mathcal{L}'(\mathbf{s})$.

Theorem 5.1. *Under Assumption 5.1 (a) and the single-boundary condition $2\delta C \leq 1$, for any scale parameter $\mathbf{s}$, the difference between the quantized objective and its smoothed counterpart satisfies*

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \leq G_L \|\mathbf{s}\|_\infty \cdot \Psi_\lambda(\gamma(\mathbf{s}))\,, \tag{29}$$

where

$$\gamma(\mathbf{s}) := \min_j \operatorname{dist}(u_j(\mathbf{s}), \mathcal{B}), \qquad \mathcal{B} := \left\{k + \tfrac{1}{2} : k \in \mathbb{Z}\right\}, \tag{30}$$

and Ψ_λ is the distribution-dependent error function given by

$$\Psi_\lambda(d) = \begin{cases} \frac{1}{2} - \int_b^{b+d} g_\delta(t)\,dt, & \text{surrogate gradient based on ZOE-Grad,} \\ \frac{1}{2} - \frac{\tanh(td)}{2}, & \text{existing surrogate gradient,} \end{cases} \tag{31}$$

with $b \in \mathcal{B}$ denoting an arbitrary quantization boundary, and $g_\delta(\cdot)$ denoting the normalized surrogate gradient defined in Eq. 18.

Remark 5.2. For the ZOE-Grad surrogate gradient, the error function $\Psi_\lambda(d)$ equals $1/2$ at $d = 0$ and vanishes as $d \to \delta C$. HTGE exhibits analogous behavior: $\Psi_\lambda(d)$ equals $1/2$ at $d = 0$ and approaches 0 as $d \to \frac{1}{2}$ when t is sufficiently large. Explicit error functions for the uniform and truncated normal perturbation distributions are provided in Appendix C.4.

Theorem 5.3. ***(Linear convergence for the original quantized objective).*** *Suppose Assumptions C.2.1 and C.2.2 hold, and let $\{\mathbf{s}_k\}_{k=0}^K$ denote the sequence of iterates generated by the update rule. If the stepsize satisfies $\eta \leq \frac{1}{M}$, then the last iterate $\mathbf{s}_K$ satisfies*

$$\mathbb{E}[\mathcal{L}(\mathbf{s}_K) - \mathcal{L}^\star] \leq (1 - \eta\mu)^K \left(\mathcal{L}'(\mathbf{s}_0) - \mathcal{L}'^\star\right) + \frac{B^2}{2\mu} + \frac{M\eta}{2\mu}\sigma + \mathbb{E}[\Delta_\lambda(\mathbf{s}_K)] + \Delta_\lambda(\mathbf{s}^\star), \tag{32}$$

where the distribution-dependent error term is now defined as

$$\Delta_\lambda(\mathbf{s}) = G_L \|\mathbf{s}\|_\infty \Psi_\lambda(\gamma(\mathbf{s}))\,, \tag{33}$$

with Ψ_λ given by Theorem 5.1, and the definitions of G_L, B, M, μ, and σ provided in Assumption C.2.

6 Experiments

6.1 Numerical Verification of ZOE-Grad

In this section, we verify the core theory of ZOE-Grad through numerical simulations. Specifically, we test whether Monte Carlo estimates match the analytic surrogate gradients characterized by Theorem 4.1 under given perturbation distributions, and whether the perturbation distributions constructed by Theorem 4.4 recover the corresponding surrogate gradients.

6.1.1 Verification of Theorem 4.1: From Perturbation to Surrogate Gradient

We verify Theorem 4.1 under uniform and truncated normal perturbations by comparing analytic surrogate gradients with Monte Carlo estimates computed from $N = 2000$ samples.

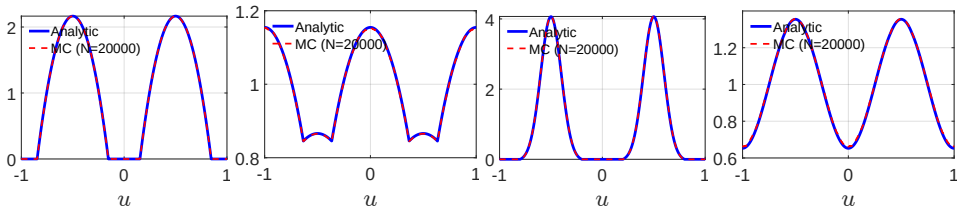
For uniform perturbations, we use the closed-form expressions in (20) and (21). Figure 1(a) and Figure 1(b) show results for $\delta = 0.2$ and $\delta = 0.5$, corresponding to the single- and multi-boundary regimes of $\mathcal{K}_u$, respectively. For truncated normal perturbations, we use the analytic expression in Appendix A.3.2 with $C = 3$ and report results for $\delta = 0.1$ and $\delta = 0.3$ in Figure 1(c) and Figure 1(d), respectively. Additional results for other choices of δ are provided in Appendix D.1.

Overall, the Monte Carlo estimates closely match the analytic gradients predicted by Theorem 4.1.

6.1.2 Verification of Theorem 4.4: From Surrogate Gradient to Perturbation

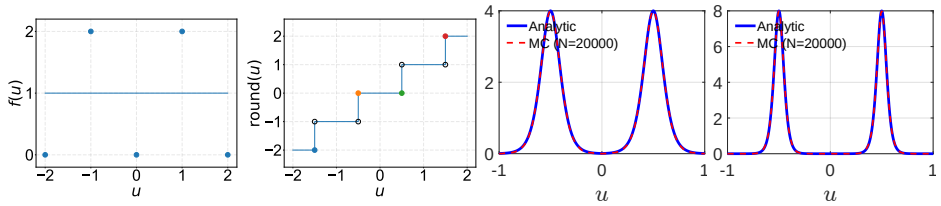
We consider Corollary 4.6 with $f(u) = \operatorname{round}(u + \frac{1}{2}) - \operatorname{round}(u - \frac{1}{2})$. Figure 2(a) shows that $f(u) = 1$ holds almost everywhere, with deviations only at integer points due to PyTorch's `torch.round` convention shown in Figure 2(b). Since integers have zero Lebesgue measure, these isolated discrepancies do not affect the expectation; hence, $f(u)$ is equivalent to STE almost everywhere. For HTGE, Figure 2(c) and Figure 2(d) report the cases $\delta = 0.1$ with $t = 8$ and $t = 16$, respectively. Additional results are deferred to Appendix D.1.

The results confirm that the perturbation distribution constructed by Theorem 4.4 accurately recovers the target surrogate gradient in expectation.



(a) $\delta = 0.2$ (b) $\delta = 0.5$ (c) $\delta = 0.1$ (d) $\delta = 0.3$

Figure 1: Comparison of perturbation distributions. The first two panels correspond to uniform perturbations, while the last two correspond to truncated Gaussian perturbations with $C = 3$.



(a) $f(u)$ (b) `round(u)` (c) $\delta = 0.1,\ t = 8$ (d) $\delta = 0.1,\ t = 16$

Figure 2: Visualization of STE and HTGE behavior. The first two panels show $f(u)$ and PyTorch rounding behavior, while the last two show HTGE with $\delta = 0.1$ and different values of t.

6.2 Fine-tuning Large Language Models

6.2.1 Experimental Setup

Datasets and Models. To evaluate optimization performance, we consider SST-2 [34] from the GLUE benchmark, SQuAD [31] for question answering, and several SuperGLUE tasks, including RTE [5], CB [6], BoolQ [4], WSC [18], WiC [30], and MultiRC [14]. We conduct experiments on a diverse set of pretrained language models, including OPT-1.3B and OPT-6.7B [43], LLaMA-2-7B [35], and Qwen3-8B [39].

Compared Methods. To evaluate ZOE-Grad, we compare surrogate gradients instantiated within our framework, including analytic surrogates induced by uniform and truncated normal perturbations from Theorem 4.1, as well as HTGE and STE characterized by Corollary 4.5 and Corollary 4.6, respectively. We also include full-precision zero-shot evaluation (Zero-Shot) and direct quantization without training (Zero-Shot-Q) [33] as baselines.

Implementation Details. All experiments are conducted using PyTorch 2.9.1 and Hugging Face Transformers 4.57.6 on a cluster of NVIDIA A100 80GB GPUs. We adopt OmniQuant as the base quantization method and initialize from its official checkpoints when available. For Qwen3, following the OmniQuant settings, we use a learning rate of 5×10^{-3} for learnable weight clipping, batch size 1, and 20 epochs to obtain the quantized model. For all datasets and methods, we train for 5,000 steps using 1,000 labeled examples per task, and evaluate on a test set of up to 1,000 examples. We report F1 for SQuAD and accuracy for all other tasks. Because memory requirements vary across tasks, we use task specific per device batch sizes. For each task, the learning rate is selected from a predefined candidate set shared across all models and surrogate gradient methods. The complete batch sizes and learning rate candidate grid are reported in Table 10 of Appendix D.2.5. All methods use the same search space and selection protocol, although the selected learning rate may differ across model and task pairs. For ZOE-Grad, we empirically find that values of δ close to the single-boundary condition $2\delta C = 1$ tend to yield superior performance. Therefore, we use uniform perturbations with $\delta = 0.285$ and truncated normal perturbations with $\delta = 0.15$ and $C = 3$. Additional δ-sensitivity results are provided in Appendix D.2.1. For HTGE, we set $t = 16$.

6.2.2 Main Results

In this section, we compare the fine-tuning performance of different surrogate gradients on downstream tasks using the OPT family, LLaMA-2-7B, and Qwen3-8B. Table 1 reports the results

Table 1: Performance comparison of 4-bit weight-only quantization with different surrogate gradients across OPT-1.3B, OPT-6.7B, Llama-2-7B, and Qwen3-8B on downstream tasks.

Model	Method	SST-2	RTE	CB	BoolQ	WSC	WIC	MultiRC	SQuAD
OPT-1.3B	Zero-Shot	53.56	53.07	39.29	47.10	43.27	57.52	43.60	29.82
	Zero-Shot-Q	53.21	54.15	41.07	40.10	41.35	54.23	42.40	12.03
	STE	90.14	57.31	64.07	58.90	57.69	57.37	45.40	16.32
	HTGE	91.97	60.29	69.21	61.20	62.50	58.78	**60.40**	**35.84**
	Uniform	**92.32**	61.37	**73.57**	63.40	61.54	59.09	59.10	30.32
	Normal	91.51	**64.98**	71.79	**66.00**	**63.46**	**59.25**	59.10	30.18
LLaMA-2-7B	Zero-Shot	58.03	62.09	33.93	66.10	36.54	50.16	42.40	58.71
	Zero-Shot-Q	50.92	48.38	50.00	39.90	53.85	50.00	58.80	41.80
	STE	93.46	58.12	57.14	60.70	63.46	51.84	59.40	47.33
	HTGE	**95.18**	66.06	**73.21**	67.80	**73.08**	**68.81**	59.60	61.73
	Uniform	94.72	**69.70**	69.64	62.30	64.42	54.67	**62.60**	**66.69**
	Normal	94.15	67.54	**73.21**	**69.60**	65.38	53.76	60.70	66.15
OPT-6.7B	Zero-Shot	61.24	54.87	53.57	57.30	37.50	51.25	41.70	40.37
	Zero-Shot-Q	56.88	52.71	35.71	45.10	36.54	53.76	41.40	29.70
	STE	93.46	62.09	69.64	60.10	60.52	55.46	58.40	41.55
	HTGE	94.72	**70.76**	76.79	61.30	**64.42**	57.83	60.30	**54.79**
	Uniform	**94.95**	67.15	78.57	**73.80**	**64.42**	**61.29**	**60.80**	54.40
	Normal	94.72	**70.76**	**85.71**	72.80	63.46	59.25	60.40	54.38
Qwen3-8B	Zero-Shot	55.50	84.48	66.07	80.20	64.42	63.01	85.50	34.50
	Zero-Shot-Q	57.57	83.03	44.64	78.20	62.50	59.25	82.90	30.18
	STE	87.96	83.39	73.21	81.80	72.12	70.38	83.40	56.88
	HTGE	89.91	86.64	82.14	**85.60**	75.00	72.57	84.60	65.18
	Uniform	**92.89**	88.81	**89.29**	85.40	**76.92**	**73.98**	87.20	**70.29**
	Normal	91.17	**90.25**	82.14	85.10	72.12	71.00	**88.60**	64.21

under the W4A16 setting, i.e., 4-bit weight-only quantization with 16-bit activations. We further provide W3A16 results for LLaMA-2-7B and W2A16 results for OPT-6.7B in Tables 6 and 7 of Appendix D.2.2, respectively, as well as results on the larger LLaMA-2-13B model in Table 8 of Appendix D.2.3.

Across quantized models with different model sizes and bit-widths, the fine tuning results consistently show that STE underperforms the other three ZOE-based surrogates derived from continuous perturbations. This indicates that continuous ZOE surrogates provide more reliable optimization signals than the discrete STE counterpart. The observation is consistent with Corollary 4.6, which characterizes STE as a degenerate zeroth-order expectation induced by discrete perturbations.

We also measure fine tuning overhead on LLaMA-2-7B under W3A16. Averaged over the eight tasks, the fine tuning times (min:s) are 46:31 for STE, 57:04 for HTGE, 60:22 for Uniform, and 61:04 for Normal. STE is computationally cheaper because it uses the constant surrogate derivative 1, whereas the boundary dependent ZOE surrogates require evaluating analytic functions during the backward pass. This additional cost introduces no inference time overhead and represents a reasonable trade-off given their consistent accuracy improvements over STE. Task level timing results are reported in Table 9 of Appendix D.2.4.

6.2.3 Selecting Surrogate Gradients via Pilot Fine-tuning

We further observe that the best surrogate gradient varies across target LLMs and downstream tasks, so we perform lightweight pilot fine-tuning to select an appropriate surrogate before full training. Specifically, for OPT-6.7B, LLaMA-2-7B, and Qwen3-8B, we run three pilot fine-tuning experiments for 256 steps using 8, 16, and 32 labeled training examples, respectively, and evaluate them on the same test set used for full training. For each model-task pair and training size, we select the best-

Table 2: Pilot selection results for surrogate gradients using different numbers of training examples.

	Best Distribution Selected			Non-worst Distribution Selected		
Number of Training Examples	8	16	32	8	16	32
Count	10	11	13	17	20	21
Percentage (%)	41.67	45.83	54.17	70.83	83.33	87.50

Table 3: W4A4KV4 results on Qwen2.5-0.5B without rotation.

Method	R-1	R-2	R-L	R_LSum	Average
STE	28.86	9.15	22.52	22.51	20.76
Uniform	30.33	10.18	23.72	23.73	**21.99**
Normal	30.32	10.00	23.56	23.57	21.86

Table 4: W4A4KV4 results on Qwen2.5-0.5B with rotation.

Method	R-1	R-2	R-L	R_LSum	Average
RoSTE	29.50	9.60	22.74	22.75	21.15
RoUniform	31.87	11.30	24.98	24.98	**23.28**
RoNormal	31.27	10.59	23.87	23.88	22.40

performing surrogate among HTGE, Uniform, and Normal, and compare it with the best surrogate obtained from the full fine-tuning results in Table 1.

For each training set size, Table 2 reports the selection accuracy over 24 model-task pairs. As the number of pilot examples increases, the pilot procedure becomes more effective at selecting the best surrogate and avoiding the worst one. With 32 examples, it selects the best surrogate in 54.17% of cases and avoids the worst in 87.50%, suggesting that small-scale pilot fine-tuning provides an efficient criterion for surrogate selection.

6.2.4 Multiple Quantization Settings and Rotation

The main experiments above study weight-only quantization. To examine broader quantization settings and the interaction between ZOE-Grad and rotation, we follow the W4A4KV4 setting of RoSTE [36] and evaluate Qwen2.5-0.5B with quantized weights, activations, and KV-cache. We consider both the setting without rotation and the corresponding RoSTE setting with rotation. Rotation and ZOE-Grad act on complementary parts of the quantization pipeline: rotation redistributes outliers before quantization, while ZOE-Grad changes the backward signal through the rounding operator. After rotation, the transformed tensors are still quantized through element wise rounding, so the ZOE-Grad surrogate can be applied directly in the rotated coordinates.

Tables 3 and 4 report the W4A4KV4 results without and with rotation, respectively. Uniform and Normal outperform their corresponding STE baselines in both settings. Together with the weight-only quantization results in Section 6.2.2, these results show that ZOE-Grad remains effective across weight-only quantization, joint weight, activation, and KV-cache quantization, and the same joint setting combined with rotation. They also support the complementary roles of rotation and ZOE-Grad.

7 Conclusion

In this paper, we propose ZOE-Grad, a zeroth-order expectation framework that establishes an equivalence between the expectations of zeroth-order gradient estimators and surrogate gradients. Under this view, STE corresponds to a degenerate and discontinuous perturbation form that ignores the local geometry of quantization boundaries, thereby limiting its expressiveness. We also derive bounded-error convergence guarantees for the quantized objective. Experiments across diverse LLM architectures show that continuous ZOE-Grad surrogates consistently outperform STE.

Acknowledgments and Disclosure of Funding

Funding. The authors received no third-party funding or support for this work.

Competing Interests. The authors declare no competing financial interests relevant to this work.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[2] Hongzheng Chen, Jiahao Zhang, Yixiao Du, Shaojie Xiang, Zichao Yue, Niansong Zhang, Yaohui Cai, and Zhiru Zhang. Understanding the potential of fpga-based spatial acceleration for large language model inference. *ACM Transactions on Reconfigurable Technology and Systems*, 18(1):1–29, 2024.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

[4] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. 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)*, pages 2924–2936, 2019.

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

[6] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124, 2019.

[7] 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.

[8] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

[9] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

[10] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

[11] 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.

[12] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4851–4860. IEEE, 2019.

[13] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in neural information processing systems*, 29, 2016.

[14] 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)*, pages 252–262, 2018.

[15] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

[16] Tanishq Kumar, Zachary Ankner, Benjamin Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, volume 2025, pages 71833–71864, 2025.

[17] Junghyup Lee, Dohyung Kim, and Bumsub Ham. Network quantization with element-wise gradient scaling. *arXiv preprint arXiv:2104.00903*, 2021.

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

[19] Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. Qllm: Accurate and efficient low-bitwidth quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pages 15417–15439, 2024.

[20] Liyuan Liu, Chengyu Dong, Xiaodong Liu, Bin Yu, and Jianfeng Gao. Bridging discrete and backpropagation: Straight-through and beyond. *Advances in Neural Information Processing Systems*, 36:12291–12311, 2023.

[21] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 467–484, 2024.

[22] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Lifeng Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 1(4), 2024.

[23] Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. Affinequant: Affine transformation quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pages 50932–50951, 2024.

[24] 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.

[25] Bhaskar Mukhoty, Velibor Bojkovic, William De Vazelhes, Xiaohan Zhao, Giulia De Masi, Huan Xiong, and Bin Gu. Direct training of snn using local zeroth order method. *Advances in Neural Information Processing Systems*, 36:18994–19014, 2023.

[26] Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, pages 16318–16330. PMLR, 2022.

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

[28] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

[29] Zehua Pei, Xufeng Yao, Wenqian Zhao, and Bei Yu. Quantization via distillation and contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17164–17176, 2023.

[30] 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)*, pages 1267–1273, 2019.

[31] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392, 2016.

[32] Charbel Sakr, Steve Dai, Rangha Venkatesan, Brian Zimmer, William Dally, and Brucek Khailany. Optimal clipping and magnitude-aware differentiation for improved quantization-aware training. In *International conference on machine learning*, pages 19123–19138. PMLR, 2022.

[33] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pages 45472–45496, 2024.

[34] 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*, pages 1631–1642, 2013.

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

[36] Quan Wei, Chung-Yiu Yau, Hoi-To Wai, Yang Katie Zhao, Dongyeop Kang, Youngsuk Park, and Mingyi Hong. Roste: An efficient quantization-aware supervised fine-tuning approach for large language models. *arXiv preprint arXiv:2502.09003*, 2025.

[37] Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances in Neural Information Processing Systems*, 35:17402–17414, 2022.

[38] 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*, pages 38087–38099. PMLR, 2023.

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[40] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 72–77, 2019.

[41] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

[42] Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. Rptq: Reorder-based post-training quantization for large language models. *arXiv preprint arXiv:2304.01089*, 2023.

[43] 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.

[44] Xiandong Zhao, Ying Wang, Xuyi Cai, Cheng Liu, and Lei Zhang. Linear symmetric quantization of neural networks for low-precision integer hardware. In *International Conference on Learning Representations*, 2020.

[45] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020.

[46] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12329–12338. IEEE, 2022.

A From Perturbation Distributions to Surrogate Gradients

A.1 Proofs of Theorem 4.1

Theorem 4.1 : Let λ be a distribution supported on $[-C, C]$ with probability density function $\lambda(z)$ (so that $\lambda(z) = 0$ for $|z| > C$). Assume the integrals $\int_0^C z\lambda(z)dz$ and $\int_0^C z^2\lambda(z)dz$ exist and are finite. Define the constant:

$$c := 2\int_0^C z^2\lambda(z)dz.$$

Define a new distribution $\tilde{\lambda}$ with probability density function:

$$\tilde{\lambda}(x) = \begin{cases} \frac{1}{c}\int_{|x|}^C z\lambda(z)dz, & |x| \le C, \\ 0, & |x| > C. \end{cases}$$

Let

$$S(u) := \mathbb{E}_{z\sim\tilde{\lambda}}\left[c\sum_{k\in\mathcal{K}_u} h(u - b_k + \delta z)\right],$$

where $h(\cdot)$ denotes the Heaviside function. Then,

$$\mathbb{E}_{z\sim\lambda}\left[G_r^2(u; z, \delta)\right] = \frac{d}{du}S(u).$$

Proof. Since $\mathcal{K}_u$ is finite, the summation and expectation are interchangeable. We obtain:

$$\begin{aligned} S(u) &= c\sum_{k\in\mathcal{K}_u}\int_{-C}^C h(u - b_k + \delta z)\tilde{\lambda}(z)dz \\ &= c\sum_{k\in\mathcal{K}_u}\int_{-(u-b_k)/\delta}^C \tilde{\lambda}(z)dz \quad \text{(because } |u - b_k| < \delta C\text{)}. \end{aligned} \tag{34}$$

Differentiating with respect to u:

$$\begin{aligned} \frac{d}{du}S(u) &= c\sum_{k\in\mathcal{K}_u}\frac{d}{du}\int_{-(u-b_k)/\delta}^C \tilde{\lambda}(z)dz \\ &= \frac{c}{\delta}\sum_{k\in\mathcal{K}_u}\tilde{\lambda}\left(-\frac{u - b_k}{\delta}\right). \end{aligned} \tag{35}$$

Using the evenness of $\tilde{\lambda}$ and its definition:

$$\tilde{\lambda}\left(\frac{|u - b_k|}{\delta}\right) = \frac{1}{c}\int_{|u-b_k|/\delta}^C t\lambda(t)dt \quad \text{(since } |u - b_k|/\delta < C\text{)}. \tag{36}$$

Substituting this yields:

$$\frac{d}{du}S(u) = \frac{1}{\delta}\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^C t\lambda(t)dt = \mathbb{E}_{z\sim\lambda}\left[G_r^2(u; z, \delta)\right]. \tag{37}$$

□

A.2 Proofs of Proposition 4.2

We prove that for any symmetric, zero-mean perturbation distribution λ with finite variance, $g(u) = \mathbb{E}_\lambda[G_r^2(u; z, \delta)]$ satisfies

$$\int_0^1 g(u)\,du = c, \qquad c = \mathbb{E}_\lambda[z^2].$$

By the periodicity of g (which inherits period 1 from the rounding function), the same holds for any unit interval $[k, k + 1]$.

Proof. Using Fubini's theorem and the definition of G_r^2,

$$\int_0^1 g(u)\,du = \int_0^1 \mathbb{E}_\lambda\left[\frac{r(u+\delta z)-r(u-\delta z)}{2\delta}\,z\right]du = \mathbb{E}_\lambda\left[\frac{z}{2\delta}\int_0^1 \big(r(u+\delta z)-r(u-\delta z)\big)\,du\right]. \tag{38}$$

For any fixed $a \in \mathbb{R}$, define the 1-periodic function $h(u) = r(u) - u$. Since rounding satisfies

$$r(u+1) = r(u) + 1, \tag{39}$$

it follows that

$$h(u+1) = h(u), \tag{40}$$

which implies that h is a periodic function with period 1. Consequently,

$$\int_0^1 r(u+a)du = \int_0^1 \big(u+a+h(u+a)\big)du = \int_0^1 udu + a + \int_0^1 h(u+a)du. \tag{41}$$

Since h is 1-periodic,

$$\int_0^1 h(u+a)du = \int_0^1 h(u)du. \tag{42}$$

Similarly,

$$\int_0^1 r(u-a)du = \int_0^1 udu - a + \int_0^1 h(u)du. \tag{43}$$

Subtracting the two equations yields

$$\int_0^1 \big[r(u+a)-r(u-a)\big]du = 2a. \tag{44}$$

Letting $a = \delta z$, we have

$$\int_0^1 \big[r(u+\delta z)-r(u-\delta z)\big]du = 2\delta z. \tag{45}$$

Therefore,

$$\begin{aligned}\int_0^1 g(u)du &= \mathbb{E}_\lambda\left[\int_0^1 \frac{r(u+\delta z)-r(u-\delta z)}{2\delta}zdu\right]\\ &= \mathbb{E}_\lambda\left[\frac{z}{2\delta}\cdot 2\delta z\right]\\ &= \mathbb{E}_\lambda[z^2] = c.\end{aligned} \tag{46}$$

□

A.3 Proofs of Corollary 4.3

We present the results of applying Theorem 4.1 across different distributions of z. For practical convenience in engineering applications, we impose certain restrictions on δ for some distributions and, for unbounded distributions, satisfy the boundedness condition $|z| < C$ via truncation or reparameterization.

A.3.1 Uniform Distribution

We consider a uniform distribution with unit variance, whose probability density function is:

$$f(z) = \begin{cases}\frac{1}{2\sqrt{3}} & \text{for } z \in [-\sqrt{3}, \sqrt{3}]\\ 0 & \text{otherwise,}\end{cases} \tag{47}$$

Consequently, it is straightforward to obtain

$$\begin{aligned}
\mathbb{E}_{z\sim\lambda}\left[G_2^{(r)}(u;z,\delta)\right] &= \frac{1}{\delta}\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^{C} z\lambda(z)dz \\
&= \frac{1}{2\sqrt{3}\delta}\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^{\sqrt{3}} z dz \\
&= \frac{1}{4\sqrt{3}\delta}\sum_{k\in\mathcal{K}_u} z^2\Big|_{|u-b_k|/\delta}^{\sqrt{3}} \\
&= \begin{cases} \frac{1}{4\sqrt{3}\delta}\sum_{k\in\mathcal{K}_u}\left(3-\frac{(u-b_k)^2}{\delta^2}\right), & \exists k, s.t. |u-b_k|<\sqrt{3}\delta \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{48}$$

We further assume $2\delta C \leq 1$, which implies that $\mathcal{K}_u$ contains at most one boundary point. Let $s(u)$ denote the boundary point closest to u. Then,

$$\mathbb{E}_{z\sim\lambda}[G_r^2(u;z,\delta)] = \begin{cases} \frac{1}{4\sqrt{3}\delta}\left(3-\frac{(u-s(u))^2}{\delta^2}\right), & |u-s(u)|<\sqrt{3}\delta, \\ 0, & \text{otherwise.} \end{cases} \tag{49}$$

A.3.2 Truncated Normal Distribution on a Symmetric Interval $[-C, C]$

We consider a truncated normal distribution, whose probability density function is given by:

$$\lambda(t) = \begin{cases} \dfrac{\phi(t)}{\Phi(C)-\Phi(-C)}, & \text{if } -C \leq t \leq C, \\ 0, & \text{otherwise,} \end{cases} \tag{50}$$

where $\phi(t) = \dfrac{1}{\sqrt{2\pi}}e^{-t^2/2}$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function of the standard normal distribution, respectively. Since the standard normal distribution is symmetric and satisfies $\Phi(-C) = 1-\Phi(C)$, we obtain

$$\Phi(C)-\Phi(-C) = \Phi(C)-[1-\Phi(C)] = 2\Phi(C)-1.$$

Thus, for $t \in [-C, C]$, we have:

$$\lambda(t) = \begin{cases} \dfrac{\phi(t)}{2\Phi(C)-1}, & \text{if } -C \leq t \leq C, \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

Given the expectation expression

$$\mathbb{E}_{z\sim\lambda}[G_2^{(\text{round})}(u)] = \frac{1}{\delta}\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^{C} z\,\lambda(z)\,dz, \tag{52}$$

we compute the integral:

$$
\begin{aligned}
I &= \frac{1}{\delta} \sum_{k \in \mathcal{K}_u} \int_{|u-b_k|/\delta}^{C} z\, \lambda(z)\, dz \\
&= \frac{1}{2\Phi(C)-1} \cdot \frac{1}{\delta\sqrt{2\pi}} \sum_{k \in \mathcal{K}_u} \int_{|u-b_k|/\delta}^{C} z e^{-z^2/2}\, dz \\
&= \frac{1}{2\Phi(C)-1} \cdot \frac{1}{\delta\sqrt{2\pi}} \sum_{k \in \mathcal{K}_u} \left[-\, e^{-z^2/2} \Big|_{|u-b_k|/\delta}^{C} \right] \\
&= \begin{cases} \frac{1}{2\Phi(C)-1} \cdot \frac{1}{\delta\sqrt{2\pi}} \sum_{k \in \mathcal{K}_u} \left[\exp\left(-\frac{(u-b_k)^2}{2\delta^2} \right) - \exp\left(-\frac{C^2}{2} \right) \right], & \exists k, s.t. |u-b_k| < \delta C \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
\tag{53}
$$

For the standard normal distribution, the 0.995 quantile is 2.5758, and the 0.999 quantile is 3.0902. Therefore, we set $C = 3$. Similarly, if we enforce $2\delta C \leq 1$, i.e., set $\delta \leq 1/6$, then

$$
I = \begin{cases} \frac{1}{2\Phi(3)-1} \cdot \frac{1}{\delta\sqrt{2\pi}} \left[\exp\left(-\frac{(u-s(u))^2}{2\delta^2} \right) - \exp\left(-\frac{9}{2} \right) \right], & \exists k, s.t. |u-b_k| < 3\delta \\ 0, & \text{otherwise.} \end{cases}
\tag{54}
$$

A.3.3 Truncated Laplace Distribution on a Symmetric Interval $[-C, C]$

Consider a Laplace distribution with parameters $\mu = 0$ and $b = \frac{1}{\sqrt{2}}$, whose probability density function is

$$
f(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|}. \tag{55}
$$

Truncating it to the interval $[-C, C]$ requires computing the normalization constant Z, i.e., the integral of $f(x)$ over this interval:

$$
N = \int_{-C}^{C} f(x) dx = \int_{-C}^{C} \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} dx. \tag{56}
$$

Using symmetry,

$$
\begin{aligned}
N &= \sqrt{2} \int_{0}^{C} e^{-\sqrt{2}x} dx \\
&= -e^{-\sqrt{2}x} \Big|_{0}^{C} \\
&= 1 - e^{-\sqrt{2}C}.
\end{aligned}
\tag{57}
$$

Hence, the probability density function of the truncated Laplace distribution is:

$$
\lambda(x) = \begin{cases} \dfrac{1}{\sqrt{2}\left(1 - e^{-\sqrt{2}C}\right)} e^{-\sqrt{2}|x|}, & \text{if } |x| \leq C, \\ 0, & \text{otherwise.} \end{cases}
\tag{58}
$$

Define $a_k = \frac{|u-b_k|}{\delta} \geq 0$. We compute the integral:

$$
\begin{aligned}
I &= \frac{1}{\delta} \sum_{k \in \mathcal{K}_u} \int_{|u-b_k|/\delta}^{C} z\, \lambda(z)\, dz \\
&= \frac{1}{\delta} \sum_{k \in \mathcal{K}_u} \int_{a_k}^{C} z\, \frac{1}{\sqrt{2}\left(1 - e^{-\sqrt{2}C}\right)} e^{-\sqrt{2}z}\, dz \\
&= \frac{1}{\delta\sqrt{2}\left(1 - e^{-\sqrt{2}C}\right)} \sum_{k \in \mathcal{K}_u} \int_{a_k}^{C} z e^{-\sqrt{2}z}\, dz \\
&= \frac{1}{\delta\sqrt{2}\left(1 - e^{-\sqrt{2}C}\right)} \sum_{k \in \mathcal{K}_u} -\left(\frac{z}{\sqrt{2}} + \frac{1}{2}\right) \exp(-\sqrt{2}z) \Big|_{a_k}^{C} \\
&= \frac{1}{2\delta\left(1 - e^{-\sqrt{2}C}\right)} \sum_{k \in \mathcal{K}_u} \left[\left(a_k + \frac{1}{\sqrt{2}}\right) e^{-\sqrt{2}a_k} - \left(C + \frac{1}{\sqrt{2}}\right) e^{-\sqrt{2}C}\right].
\end{aligned}
\tag{59}
$$

For the standard Laplace distribution with mean 0 and scale parameter $b = \frac{1}{\sqrt{2}}$ (i.e., unit variance), the 0.995 quantile is 3.2563, and the 0.999 quantile is 4.3944. Therefore, we set $C = 4.5$. Similarly, if we enforce $2\delta C \leq 1$, i.e., set $\delta \leq 1/9$, then

$$
I = \frac{1}{2\delta\left(1 - e^{-\frac{9\sqrt{2}}{2}}\right)} \left[\left(a(u) + \frac{1}{\sqrt{2}}\right) e^{-\sqrt{2}a(u)} - \left(C + \frac{1}{\sqrt{2}}\right) e^{-\frac{9\sqrt{2}}{2}}\right], \tag{60}
$$

where $a(u) = \frac{|u-s(u)|}{\delta}$.

B From Surrogate Gradients to Perturbation Distributions

B.1 Proofs of Theorem 4.4

Theorem 4.4. Given a surrogate gradient $g(u)$ that admits the representation

$$
g(u) := \sum_{k \in \mathcal{K}_u} \psi(|u - b_k|),
$$

where $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a continuously differentiable, non-increasing function satisfying $\psi(\delta C) = 0$. We further assume that $c := -2\delta^2 \int_0^C \frac{1}{z}\psi'(z\delta)\, dz < \infty$ and define

$$
\lambda(z) = \begin{cases} -\dfrac{\delta^2}{cz}\psi'(z\delta), & 0 < |z| \leq C, \\ 0, & \text{otherwise.} \end{cases}
$$

So that λ is a probability density function. Then, we have:

$$
c\, \mathbb{E}_{z \sim \lambda}\left[G_r^2(u; z, \delta)\right] = g(u).
$$

Proof. From Eq. (12), we have

$$
\mathbb{E}_{z \sim \lambda}[G_r^2(u; z, \delta)] = \frac{1}{\delta} \sum_{k \in \mathcal{K}_u} \int_{|u-b_k|/\delta}^{C} z\lambda(z)\, dz. \tag{61}
$$

Substituting the expression for λ yields:

$$\begin{aligned} c\,\mathbb{E}_{z\sim\lambda}[G_r^2(u;z,\delta)] &= \frac{c}{\delta}\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^{C} z\left(-\frac{\delta^2}{cz}\psi'(z\delta)\right)dz \\ &= -\delta\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|/\delta}^{C}\psi'(z\delta)\,dz. \\ &= -\sum_{k\in\mathcal{K}_u}\int_{|u-b_k|}^{\delta C}\psi'(x)\,dx. \\ &= -\sum_{k\in\mathcal{K}_u}\left[\psi(\delta C)-\psi(|u-b_k|)\right] \\ &= \sum_{k\in\mathcal{K}_u}\psi(|u-b_k|). \end{aligned} \tag{62}$$

Therefore, we have

$$c\,\mathbb{E}_{z\sim\lambda}[G_r^2(u;z,\delta)] = \sum_{k\in\mathcal{K}_u}\psi(|u-b_k|) = g(u). \tag{63}$$

□

B.2 Proofs of Corollary 4.5

We consider the hyperbolic tangent gradient estimator (HTGE), whose surrogate gradient is given by

$$\begin{aligned} g(u) &= \frac{t}{2}\left[1-\tanh^2\left(t\left(u-\lfloor u\rfloor-\frac{1}{2}\right)\right)\right] \\ &= \frac{t}{2}\left[1-\tanh^2\left(t\left|u-\lfloor u\rfloor-\frac{1}{2}\right|\right)\right]. \end{aligned} \tag{64}$$

Note that $u-\lfloor u\rfloor-\frac{1}{2}$ is mathematically equivalent to $u-s(u)$, where $s(u)$ denotes the boundary point closest to u. Therefore, $g(u)$ can be written as

$$g(u) = \psi(|u-s(u)|),$$

which corresponds to a single-boundary formulation. Hence, we set $\delta C = \frac{1}{2}$. Define $\psi:\mathbb{R}_{\geq 0}\to\mathbb{R}$ as

$$\psi(x) = \frac{t}{2}\left[1-\tanh^2(tx)\right] = \frac{t}{2}\operatorname{sech}^2(tx). \tag{65}$$

When t is sufficiently large, $\psi(\frac{1}{2}) = \frac{t}{2}\operatorname{sech}^2(\frac{t}{2})$ approaches zero. The derivative of $\psi(x)$ is

$$\psi'(x) = -t^2\operatorname{sech}^2(tx)\tanh(tx). \tag{66}$$

Following Theorem 4.4, we define the constant c as

$$\begin{aligned} c :&= -2\delta^2\int_0^C \frac{1}{z}\psi'(z\delta)\,dz \\ &= 2\delta^2t^2\int_0^C \frac{\operatorname{sech}^2(t\delta z)\tanh(t\delta z)}{z}\,dz \\ &= 2\delta^2t^2\int_0^{t\delta C}\frac{\operatorname{sech}^2 u\tanh u}{u}\,du, \end{aligned} \tag{67}$$

Substituting Eqs. (66) and (67) into Eq. (23), we obtain

$$\lambda(z) = \begin{cases} \frac{\delta^2t^2}{c}\cdot\frac{\operatorname{sech}^2(t\delta z)\tanh(t\delta z)}{z}, & 0<|z|\leq C, \\ 0, & \text{otherwise.} \end{cases} \tag{68}$$

Moreover, the integrand has a finite limit at $z=0$, namely

$$\lim_{z\to 0}\frac{\operatorname{sech}^2(t\delta z)\tanh(t\delta z)}{z} = t\delta.$$

Hence, $\lambda(z)$ is continuous at the origin, and we define

$$\lambda(0) = \frac{\delta^3 t^3}{c}.$$

Thus, we obtain the perturbation distribution corresponding to the HTGE surrogate gradient.

B.3 Proofs of Corollary 4.6

For the widely used straight-through estimator (STE), the surrogate can be viewed as a special case of our zeroth-order expectation perspective. While Theorem 4.4 constructs a continuous perturbation distribution such that the expected zeroth-order estimator recovers a given surrogate gradient under the specified perturbation, STE corresponds to the expectation of a two-point zeroth-order estimator under a discrete perturbation, namely

$$\delta = \frac{1}{2}, \quad z \sim \text{Rademacher } (\mathbb{P}(z=1) = \mathbb{P}(z=-1) = 1/2). \tag{69}$$

The proof follows directly from Eq. (4). Specifically,

$$\begin{aligned}
\mathbb{E}_{z\sim\lambda}\left[G_r^2(u;z,\delta)\right] &= \mathbb{E}_{z\sim\lambda}\left[\frac{r(u+\delta z) - r(u-\delta z)}{2\delta}\, z\right] \\
&= \frac{1}{2}\left[r\left(u+\frac{1}{2}\right) - r\left(u-\frac{1}{2}\right)\right] - \frac{1}{2}\left[r\left(u-\frac{1}{2}\right) - r\left(u+\frac{1}{2}\right)\right] \\
&= r\left(u+\frac{1}{2}\right) - r\left(u-\frac{1}{2}\right)
\end{aligned} \tag{70}$$

The resulting function $f(u) = \text{round}\left(u+\frac{1}{2}\right) - \text{round}\left(u-\frac{1}{2}\right)$ is illustrated in Figure 2.

C Proofs of the convergence analysis

C.1 Notation and Setup

Let $\mathbf{W}_f$ denote the fixed full-precision weights, and let $\mathbf{s}$ denote the learnable scale parameters. We define the normalized weights as

$$u(\mathbf{s}) := \mathbf{W} \oslash \mathbf{s}, \tag{71}$$

where $\oslash$ denotes element-wise division. We consider a uniform quantization scheme. The discrete quantizer is defined as

$$q(u) := \text{clamp}(\text{round}(u), -2^{N-1}, 2^{N-1}-1), \tag{72}$$

and the corresponding dequantized weights are

$$\widehat{\mathbf{W}}^{\text{disc}}(\mathbf{s}) := \mathbf{s} \odot q(u(\mathbf{s})). \tag{73}$$

We define the smoothed quantizer as

$$q'(u) := \text{clamp}(\hat{r}(u), -2^{N-1}, 2^{N-1}-1), \tag{74}$$

and the corresponding dequantized weights as

$$\widehat{\mathbf{W}}^{\text{sm}}(\mathbf{s}) := \mathbf{s} \odot q'(u(\mathbf{s})). \tag{75}$$

The original quantized objective and the smoothed objective are defined as

$$\mathcal{L}(\mathbf{s}) := L\left(\widehat{\mathbf{W}}^{\text{disc}}(\mathbf{s})\right), \qquad \mathcal{L}'(\mathbf{s}) := L\left(\widehat{\mathbf{W}}^{\text{sm}}(\mathbf{s})\right). \tag{76}$$

Finally, we define the distance from a scalar u to the nearest boundary as

$$\text{dist}(u, \mathcal{B}) := |u - s(u)|. \tag{77}$$

C.2 Assumption

Assumption C.2.1. Function properties.

(a) (ℓ_∞-Lipschitz continuity of the loss L) We assume that the loss function $L(\cdot)$ is Lipschitz continuous with respect to the ℓ_∞-norm, i.e., there exists a constant $G_L > 0$ such that

$$|L(\mathbf{W}_1) - L(\mathbf{W}_2)| \le G_L \|\mathbf{W}_1 - \mathbf{W}_2\|_\infty, \qquad \forall \mathbf{W}_1, \mathbf{W}_2. \tag{78}$$

(b) (Smoothness of the smoothed objective) The smoothed objective $\mathcal{L}'(\mathbf{s})$ is differentiable and M-smooth, i.e.,

$$\|\nabla \mathcal{L}'(\mathbf{s}_1) - \nabla \mathcal{L}'(\mathbf{s}_2)\| \le M \|\mathbf{s}_1 - \mathbf{s}_2\|, \qquad \forall \mathbf{s}_1, \mathbf{s}_2. \tag{79}$$

(c) (Lower boundedness) The smoothed objective and the original objective are lower bounded:

$$\mathcal{L}'^\star := \inf_{\mathbf{s}} \mathcal{L}'(\mathbf{s}) > -\infty. \qquad \mathcal{L}^\star := \mathcal{L}(\mathbf{s}^\star) = \inf_{\mathbf{s}} \mathcal{L}(\mathbf{s}). \tag{80}$$

(d) (PL condition for $\mathcal{L}'$) The smoothed objective $\mathcal{L}'$ satisfies the Polyak–Łojasiewicz condition with parameter $\mu > 0$:

$$\mathcal{L}'(\mathbf{s}) - \mathcal{L}'^\star \le \frac{1}{2\mu} \|\nabla \mathcal{L}'(\mathbf{s})\|^2, \qquad \forall \mathbf{s}. \tag{81}$$

Assumption C.2.2. Gradient properties.

(a) (Bias decomposition of the update direction) Let $\{\mathcal{F}_k\}$ be the natural filtration. The update direction satisfies

$$\mathbb{E}[\hat{\mathbf{g}}_k \mid \mathcal{F}_k] = \nabla \mathcal{L}'(\mathbf{s}_k) + \mathbf{b}_k. \tag{82}$$

(b) (Bounded conditional variance) There exists $\sigma \ge 0$ such that

$$\mathbb{E}\big[\|\hat{\mathbf{g}}_k - \mathbb{E}[\hat{\mathbf{g}}_k \mid \mathcal{F}_k]\|^2 \,\big|\, \mathcal{F}_k\big] \le \sigma. \tag{83}$$

(c) (Uniformly bounded bias) There exists $B \ge 0$ such that

$$\|\mathbf{b}_k\|^2 \le B^2, \qquad \forall k = 0, 1, \ldots, K-1. \tag{84}$$

C.3 Proofs of Theorem 5.1

Before proving Theorem 5.1, we first give a lemma characterizing the pointwise error between the smoothed rounding function and the rounding function.

Lemma C.1. *Let $g_\delta(\cdot)$ denote the normalized surrogate gradient defined in Eq. 18, and let $\hat{r}(u)$ denote the normalized smoothed rounding function defined in Eq. 19. For any $u \in \mathbb{R}$, let $s(u) \in \mathcal{B}$ be the nearest quantization boundary and define $d := |u - s(u)|$. Under the single-boundary condition $2\delta C \le 1$, the pointwise approximation error satisfies*

$$|\hat{r}(u) - \mathrm{round}(u)| = \Psi_\lambda(d) = \begin{cases} \frac{1}{2} - \int_b^{b+d} g_\delta(t)\, dt, & \textit{surrogate gradient based on ZOE-Grad,} \\ \frac{1}{2} - \frac{\tanh(td)}{2}, & \textit{existing surrogate gradient,} \end{cases} \tag{85}$$

where $b \in \mathcal{B}$ is an arbitrary quantization boundary.

For the ZOE-Grad surrogate gradient, the error function $\Psi_\lambda(d)$ equals $1/2$ at $d = 0$ and vanishes as $d \to \delta C$. HTGE exhibits analogous behavior: $\Psi_\lambda(d)$ equals $1/2$ at $d = 0$ and approaches 0 as $d \to \frac{1}{2}$ when t is sufficiently large.

Proof. **Surrogate gradients derived from the ZOE-Grad framework.**

Under the single-boundary condition $2\delta C \le 1$, we now consider a single normalized weight $u \in \mathbb{R}$. In the unit interval $[k, k+1]$, the boundary $b := s(u) = k + \frac{1}{2}$ is constant. From Proposition 4.2, $\int_k^{k+1} g_\delta(u)du = 1$ and g_δ vanishes outside $[b - \delta C, b + \delta C]$, i.e.,

$$\int_k^{k+1} g_\delta(u)du = \int_{b-\frac{1}{2}}^{b+\frac{1}{2}} g_\delta(u)du = \int_{b-\delta C}^{b+\delta C} g_\delta(u)du = 1. \tag{86}$$

Since g_δ is symmetric about b,

$$\int_{b-\delta C}^{b} g_\delta(u)\,du = \int_{b}^{b+\delta C} g_\delta(u)\,du = \frac{1}{2}. \tag{87}$$

Denote $k \le u \le b - \delta C$ and $b + \delta C \le u \le k + 1$ as the flat regions. Let $u = b + x$ and consider the local offset $x \in [-\delta C, \delta C]$. Then

$$\hat{r}(b + x) = \hat{r}(b - \delta C) + \int_{b-\delta C}^{b+x} g_\delta(u)\,du. \tag{88}$$

Since $b - \delta C$ lies in the flat region, $\hat{r}(b - \delta C) = \mathrm{round}(b - \delta C) = b - \frac{1}{2}$.

- When $x \le 0$, $b + x \le b$, and the integration limits fall in the left half:

$$\hat{r}(b + x) = b - \tfrac{1}{2} + \underbrace{\int_{b-\delta C}^{b} g_\delta(u)\,du}_{=1/2} - \int_{b+x}^{b} g_\delta(u)\,du = b - \int_{b+x}^{b} g_\delta(u)\,du. \tag{89}$$

  Here we used $\int_{b-\delta C}^{b+x} g = \int_{b-\delta C}^{b} g - \int_{b+x}^{b} g$ (since $b + x < b$).

- When $x > 0$, $b + x > b$, the integration limits enter the right half, and we directly write

$$\hat{r}(b + x) = b - \tfrac{1}{2} + \int_{b-\delta C}^{b} g_\delta(u)\,du + \int_{b}^{b+x} g_\delta(u)\,du = b + \int_{b}^{b+x} g_\delta(u)\,du. \tag{90}$$

Thus we obtain a compact piecewise expression:

$$\hat{r}(b + x) = \begin{cases} b - \int_{b+x}^{b} g_\delta(u)\,du, & x \le 0, \\ b + \int_{b}^{b+x} g_\delta(u)\,du, & x > 0. \end{cases} \tag{91}$$

For a half-integer boundary b, the rounding function takes $b - \frac{1}{2}$ on the left side of the center and $b + \frac{1}{2}$ on the right side (the value at the center $x = 0$ can be defined arbitrarily without affecting the integral error). That is,

$$\mathrm{round}(b + x) = \begin{cases} b - \dfrac{1}{2}, & x \le 0, \\ b + \dfrac{1}{2}, & x > 0. \end{cases} \tag{92}$$

We now compute the absolute error.

- Case 1: $x \le 0$ (or $x = 0$). round $= b - \frac{1}{2}$,

$$\hat{r} - \mathrm{round} = \frac{1}{2} - \int_{b+x}^{b} g_\delta(u)\,du. \tag{93}$$

  Using symmetry $g_\delta(b - t) = g_\delta(b + t)$, for $d \ge 0$ we have

$$\int_{b-d}^{b} g_\delta(u)\,du = \int_{b}^{b+d} g_\delta(u)\,du. \tag{94}$$

  Hence, let $-x = d$:

$$\int_{b+x}^{b} g_\delta(u)\,du = \int_{b}^{b-x} g_\delta(u)\,du. \tag{95}$$

  That is,

$$\hat{r} - \mathrm{round} = \frac{1}{2} - \int_{b}^{b-x} g_\delta(u)\,du. \tag{96}$$

- Case 2: $x > 0$. $\text{round} = b + \frac{1}{2}$,

$$\text{round} - \hat{r} = \frac{1}{2} - \int_b^{b+x} g_\delta(u)\, du. \tag{97}$$

Combining both cases, let $d = |x| = |u - s(u)|$ $(= \text{dist}(u, \mathcal{B}))$. Then we always have

$$|\,\text{round}(b + x) - \hat{r}(b + x)| = \frac{1}{2} - \int_b^{b+d} g_\delta(t)\, dt, \quad 0 \le d < \delta C. \tag{98}$$

When $d = 0$, the integral is 0 and the error is $1/2$ (exactly at the midpoint); when $d \to \delta C$, the integral tends to $1/2$ and the error tends to 0. Since the above conclusion holds for any unit interval $[k, k + 1], k \in \mathbb{Z}$, and the surrogate gradient $g_\delta(\cdot)$ is a periodic function with period 1, $b \in \mathcal{B}$ is an arbitrary quantization boundary.

Existing surrogate gradients (e.g., HTGE)

For HTGE, we directly adopt the original function formula from the paper:

$$\hat{r}(u) = \frac{a + b}{2} + \frac{\tanh\left(t\left(u - \frac{a+b}{2}\right)\right)}{2}, \tag{99}$$

where $a = \text{floor}(u)$ and $b = \text{ceil}(u)$.

In the interval $[a, b]$, when $a \le u \le \frac{a+b}{2}$, we have $\text{round}(u) = a$. Thus,

$$\begin{aligned} \hat{r}(u) - \text{round}(u) &= \frac{a + b}{2} + \frac{\tanh\left(t\left(u - \frac{a+b}{2}\right)\right)}{2} - a \\ &= \frac{b - a}{2} + \frac{\tanh\left(t\left(u - s(u)\right)\right)}{2} \\ &= \frac{1}{2} - \frac{\tanh\left(t\,|u - s(u)|\right)}{2}. \end{aligned} \tag{100}$$

When $\frac{a+b}{2} \le u \le b$, we have $\text{round}(u) = b$. Thus,

$$\begin{aligned} \text{round}(u) - \hat{r}(u) &= b - \frac{a + b}{2} - \frac{\tanh\left(t\left(u - \frac{a+b}{2}\right)\right)}{2} \\ &= \frac{b - a}{2} - \frac{\tanh\left(t\left(u - s(u)\right)\right)}{2} \\ &= \frac{1}{2} - \frac{\tanh\left(t\,|u - s(u)|\right)}{2}. \end{aligned} \tag{101}$$

Combining both cases, let $d = |u - s(u)|$ $(= \text{dist}(u, \mathcal{B}))$. Then we always have

$$|\hat{r}(u) - \text{round}(u)| = \frac{1}{2} - \frac{\tanh(td)}{2}, \quad 0 \le d < \frac{1}{2}. \tag{102}$$

When $d = 0$, the integral is 0 and the error is $1/2$ (exactly at the midpoint); when $d \to \frac{1}{2}$ and the value of t is large, the error tends to 0. □

Theorem 5.1. Under Assumption 5.1 (a) and the single-boundary condition $2\delta C \le 1$, for any scale parameter $\mathbf{s}$, the difference between the quantized objective and its smoothed counterpart satisfies

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \le G_L \|\mathbf{s}\|_\infty \cdot \Psi_\lambda(\gamma(\mathbf{s}))\,,$$

where

$$\gamma(\mathbf{s}) := \min_j \text{dist}(u_j(\mathbf{s}), \mathcal{B}), \qquad \mathcal{B} := \left\{k + \tfrac{1}{2} : k \in \mathbb{Z}\right\},$$

and Ψ_λ is the distribution-dependent error function given by

$$\Psi_\lambda(d) = \begin{cases} \frac{1}{2} - \int_b^{b+d} g_\delta(t)\, dt, & \text{surrogate gradient based on ZOE-Grad} \\ \frac{1}{2} - \frac{\tanh(td)}{2}, & \text{existing surrogate gradient} \end{cases}$$

with $b \in \mathcal{B}$ denoting an arbitrary quantization boundary, and $g_\delta(\cdot)$ denoting the normalized surrogate gradient defined in Eq. 18. The function Ψ_λ is continuous and monotonically non-increasing.

Proof. We start from the definitions:

$$\mathcal{L}(\mathbf{s}) = L\left(\widehat{\mathbf{W}}^{\mathrm{disc}}(\mathbf{s})\right), \quad \mathcal{L}'(\mathbf{s}) = L\left(\widehat{\mathbf{W}}^{\mathrm{sm}}(\mathbf{s})\right). \tag{103}$$

By Assumption 5.1 (a), L is ℓ_∞-Lipschitz with constant G_L, hence

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \le G_L \left\|\widehat{\mathbf{W}}^{\mathrm{disc}}(\mathbf{s}) - \widehat{\mathbf{W}}^{\mathrm{sm}}(\mathbf{s})\right\|_\infty. \tag{104}$$

By definition of the dequantized weights,

$$\widehat{\mathbf{W}}^{\mathrm{disc}}(\mathbf{s}) = \mathbf{s} \odot q(u(\mathbf{s})), \quad \widehat{\mathbf{W}}^{\mathrm{sm}}(\mathbf{s}) = \mathbf{s} \odot q'(u(\mathbf{s})). \tag{105}$$

where $u(\mathbf{s}) = \mathbf{W} \oslash \mathbf{s}$. Thus

$$\widehat{\mathbf{W}}^{\mathrm{disc}}(\mathbf{s}) - \widehat{\mathbf{W}}^{\mathrm{sm}}(\mathbf{s}) = \mathbf{s} \odot \big(q(u(\mathbf{s})) - q'(u(\mathbf{s}))\big). \tag{106}$$

Taking the ℓ_∞-norm,

$$\left\|\widehat{\mathbf{W}}^{\mathrm{disc}}(\mathbf{s}) - \widehat{\mathbf{W}}^{\mathrm{sm}}(\mathbf{s})\right\|_\infty \le \|\mathbf{s}\|_\infty \cdot \|q(u(\mathbf{s})) - q'(u(\mathbf{s}))\|_\infty. \tag{107}$$

Since the clipping operator $q(\cdot) = \mathrm{clamp}(\cdot, -2^{N-1}, 2^{N-1} - 1)$ is 1-Lipschitz,

$$\|q(u(\mathbf{s})) - q'(u(\mathbf{s}))\|_\infty \le \|\mathrm{round}(u(\mathbf{s})) - \hat{r}(u(\mathbf{s}))\|_\infty. \tag{108}$$

Thus

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \le G_L \|\mathbf{s}\|_\infty \cdot \|\,\mathrm{round}(u(\mathbf{s})) - \hat{r}(u(\mathbf{s}))\|_\infty. \tag{109}$$

By Lemma C.1, we have

$$|\hat{r}(u) - \mathrm{round}(u)| = \Psi_\lambda(d) \tag{110}$$

where $\Psi_\lambda(d)$ is defined in Eq (31). Now maximizing over all coordinates u_j of $\mathbf{u}(\mathbf{s})$,

$$\|\mathrm{round}(\mathbf{u}(\mathbf{s})) - \hat{r}(\mathbf{u}(\mathbf{s}))\|_\infty = \max_j |\,\mathrm{round}(u_j) - \hat{r}(u_j)|. \tag{111}$$

For each coordinate with distance $d_j = \mathrm{dist}(u_j, \mathcal{B})$, the error is $\Psi_\lambda(d_j)$. Since Ψ_λ is monotonically non-increasing, the maximum is attained at the smallest distance $\gamma(\mathbf{s}) = \min_j d_j$. Therefore

$$\|\mathrm{round}(\mathbf{u}(\mathbf{s})) - \hat{r}(\mathbf{u}(\mathbf{s}))\|_\infty \le \Psi_\lambda(\gamma(\mathbf{s})). \tag{112}$$

Substituting back through the chain of inequalities yields

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \le G_L \|\mathbf{s}\|_\infty \cdot \Psi_\lambda(\gamma(\mathbf{s})), \tag{113}$$

which completes the proof. □

C.4 Explicit error functions for uniform and truncated normal distributions.

Under the single-boundary condition $2\delta C \le 1$, the distribution-dependent error function Ψ_λ in Theorem 5.1 admits the following closed-form expressions.

(a) **Uniform distribution.** Let $\lambda = \mathcal{U}[-\sqrt{3}, \sqrt{3}]$ (variance factor $c = 1$). Then, we have,

$$\Psi_{\mathrm{unif}}(d) = \begin{cases} \dfrac{1}{2} - \dfrac{1}{4\sqrt{3}\,\delta}\left(3d - \dfrac{d^3}{3\delta^2}\right), & 0 \le d \le \sqrt{3}\,\delta, \\ 0, & d \ge \sqrt{3}\delta. \end{cases} \tag{114}$$

(b) **Truncated normal distribution.** Let λ be the distribution obtained by truncating the standard normal $\mathcal{N}(0,1)$ to $[-3,3]$ and renormalizing (i.e., $\lambda(z) = \frac{\phi(z)}{2\Phi(3)-1}$ for $|z| \le 3$, where ϕ and Φ are the standard normal pdf and cdf). Here $C = 3$, and the variance factor is

$$c = 1 - \frac{6\,e^{-9/2}}{(2\Phi(3) - 1)\sqrt{2\pi}}. \tag{115}$$

Then, we have,

$$\Psi_{\mathrm{norm}}(d) = \begin{cases} \dfrac{1}{2} - \dfrac{\Phi(d/\delta) - \frac{1}{2}}{c\,(2\Phi(3) - 1)} + \dfrac{e^{-9/2}\,d}{c\,(2\Phi(3) - 1)\,\delta\sqrt{2\pi}}, & 0 \le d \le 3\delta, \\ 0, & d \ge 3\delta. \end{cases} \tag{116}$$

Proof. **C.4.1 General Setup**

Consider the single-boundary condition $2\delta C \leq 1$. Take an arbitrary quantization boundary $b \in \{k+1/2 \mid k \in \mathbb{Z}\}$. In the interval $[b, b+d]$, there is only the boundary point b; hence for $t \in [b, b+d]$, the nearest-boundary function $s(t) = b$.

The normalized surrogate gradient $g_\delta(u)$ is periodic with period 1 and symmetric about each boundary point. The error function is defined as

$$\Psi_\lambda(d) = \frac{1}{2} - \int_b^{b+d} g_\delta(t)\, dt. \tag{117}$$

Apply the change of variable $x = t - b$; then $dt = dx$ and the integration limits become 0 and d:

$$\int_b^{b+d} g_\delta(t)\, dt = \int_0^d g_\delta(b+x)\, dx. \tag{118}$$

Since for $x \in [0, d]$, $|(b+x) - s(b+x)| = x$ (assuming $x \geq 0$), the surrogate gradient depends only on x.

C.4.2 Uniform distribution $\lambda \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$

Here $C = \sqrt{3}$ and the distribution variance is $c = 1$ (natural normalization). From equation (24) of the paper, the normalized surrogate gradient is

$$g_\delta(u) = \begin{cases} \dfrac{1}{4\sqrt{3}\,\delta}\left(3 - \dfrac{(u - s(u))^2}{\delta^2}\right), & |u - s(u)| < \sqrt{3}\,\delta, \\ 0, & \text{otherwise.} \end{cases} \tag{119}$$

For $x \in [0, d]$ (with $0 \leq d \leq \sqrt{3}\,\delta$), substitute $u = b + x$; then $s(u) = b$, so

$$g_\delta(b+x) = \frac{1}{4\sqrt{3}\,\delta}\left(3 - \frac{x^2}{\delta^2}\right). \tag{120}$$

Thus

$$\begin{aligned} \int_0^d g_\delta(b+x)\, dx &= \frac{1}{4\sqrt{3}\,\delta}\int_0^d \left(3 - \frac{x^2}{\delta^2}\right) dx \\ &= \frac{1}{4\sqrt{3}\,\delta}\left[3x - \frac{x^3}{3\delta^2}\right]_0^d \\ &= \frac{1}{4\sqrt{3}\,\delta}\left(3d - \frac{d^3}{3\delta^2}\right). \end{aligned} \tag{121}$$

When $d \geq \sqrt{3}\,\delta$, the integral no longer increases (the surrogate gradient is zero outside its support), and the integral up to $\sqrt{3}\,\delta$ already equals $\frac{1}{2}$; hence $\Psi_{\text{unif}}(d) = 0$.

Finally, we obtain:

$$\Psi_{\text{unif}}(d) = \begin{cases} \dfrac{1}{2} - \dfrac{1}{4\sqrt{3}\,\delta}\left(3d - \dfrac{d^3}{3\delta^2}\right), & 0 \leq d \leq \sqrt{3}\,\delta, \\ 0, & d \geq \sqrt{3}\delta. \end{cases} \tag{122}$$

C.4.3 Truncated normal distribution ($\mathcal{N}(0,1)$ truncated to $[-3, 3]$)

Here $C = 3$, and the variance factor (normalization constant in Proposition 4.2) is

$$c = \mathbb{E}_\lambda[z^2] = 1 - \frac{6\, e^{-9/2}}{(2\Phi(3) - 1)\sqrt{2\pi}}. \tag{123}$$

The unnormalized surrogate gradient is

$$I(u) = \frac{1}{2\Phi(3)-1} \cdot \frac{1}{\delta\sqrt{2\pi}} \left[\exp\left(-\frac{(u-s(u))^2}{2\delta^2} \right) - e^{-9/2} \right], \tag{124}$$

valid for $|u - s(u)| < 3\delta$, and zero otherwise. After normalization:

$$g_\delta(u) = \frac{1}{c} I(u) = \frac{1}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}} \left[\exp\left(-\frac{(u-s(u))^2}{2\delta^2} \right) - e^{-9/2} \right]. \tag{125}$$

For $x \in [0, d]$ (with $0 \le d \le 3\delta$), set $u = b + x$; then $s(u) = b$, so

$$g_\delta(b+x) = \frac{1}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}} \left(e^{-\frac{x^2}{2\delta^2}} - e^{-9/2} \right). \tag{126}$$

Compute the integral:

$$\int_0^d g_\delta(b+x)\,dx = \frac{1}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}} \left[\int_0^d e^{-\frac{x^2}{2\delta^2}}\,dx \;-\; e^{-9/2} d \right]. \tag{127}$$

For the first term, let $u = x/\delta$; then $dx = \delta\,du$, and the integration limits become 0 to d/δ:

$$\int_0^d e^{-\frac{x^2}{2\delta^2}}\,dx = \delta \int_0^{d/\delta} e^{-u^2/2}\,du = \delta\sqrt{2\pi} \left[\Phi\left(\frac{d}{\delta}\right) - \frac{1}{2} \right]. \tag{128}$$

Substitute back:

$$\begin{aligned} \int_0^d g_\delta(b+x)\,dx &= \frac{1}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}} \left\{ \delta\sqrt{2\pi} \left[\Phi\left(\frac{d}{\delta}\right) - \frac{1}{2} \right] - e^{-9/2} d \right\} \\ &= \frac{1}{c\,(2\Phi(3)-1)} \left[\Phi\left(\frac{d}{\delta}\right) - \frac{1}{2} \right] - \frac{e^{-9/2}\,d}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}}. \end{aligned} \tag{129}$$

Therefore

$$\Psi_{\mathrm{norm}}(d) = \frac{1}{2} - \int_0^d g_\delta(b+x)\,dx = \frac{1}{2} - \frac{\Phi(d/\delta) - \frac{1}{2}}{c\,(2\Phi(3)-1)} + \frac{e^{-9/2}\,d}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}}. \tag{130}$$

When $d \ge 3\delta$, $g_\delta(b+x) = 0$ for all $x > 3\delta$, and the integral up to 3δ already reaches $\frac{1}{2}$; thus, the integral is constantly $\frac{1}{2}$ and $\Psi_{\mathrm{norm}}(d) = 0$.

Finally, we obtain:

$$\Psi_{\mathrm{norm}}(d) = \begin{cases} \dfrac{1}{2} - \dfrac{\Phi(d/\delta) - \frac{1}{2}}{c\,(2\Phi(3)-1)} + \dfrac{e^{-9/2}\,d}{c\,(2\Phi(3)-1)\,\delta\sqrt{2\pi}}, & 0 \le d \le 3\delta, \\ 0, & d \ge 3\delta. \end{cases} \tag{131}$$

□

C.5 Proofs of Theorem 5.3

Theorem 5.3. Suppose Assumptions C.2.1 and C.2.2 hold, and let $\{\mathbf{s}_k\}_{k=0}^K$ denote the sequence of iterates generated by the update rule. If the stepsize satisfies $\eta \le \frac{1}{M}$, then the last iterate $\mathbf{s}_K$ satisfies

$$\mathbb{E}[\mathcal{L}(\mathbf{s}_K) - \mathcal{L}^\star] \le (1-\eta\mu)^K \big(\mathcal{L}'(\mathbf{s}_0) - \mathcal{L}'^\star\big) + \frac{B^2}{2\mu} + \frac{M\eta}{2\mu}\sigma + \mathbb{E}[\Delta_\lambda(\mathbf{s}_K)] + \Delta_\lambda(\mathbf{s}^\star),$$

where the distribution-dependent error term is now defined as

$$\Delta_\lambda(\mathbf{s}) = G_L \|\mathbf{s}\|_\infty \Psi_\lambda(\gamma(\mathbf{s}))\,,$$

with Ψ_λ given by Theorem 5.1.

Proof. We start from the exact decomposition

$$\mathcal{L}(\mathbf{s}_K) - \mathcal{L}^* = \underbrace{\mathcal{L}'(\mathbf{s}_K) - \mathcal{L}'^*}_{T_1} + \underbrace{\mathcal{L}(\mathbf{s}_K) - \mathcal{L}'(\mathbf{s}_K)}_{T_2} + \underbrace{\big(\mathcal{L}'^* - \mathcal{L}^*\big)}_{T_3}. \tag{132}$$

We now bound the three terms on the right-hand side of (132). To bound T_1, we let

$$\Gamma_k := \mathbb{E}[\mathcal{L}'(\mathbf{s}_k) - \mathcal{L}'^\star]. \tag{133}$$

By M-smoothness of $\mathcal{L}'$,

$$\mathcal{L}'(\mathbf{s}_{k+1}) \le \mathcal{L}'(\mathbf{s}_k) + \langle \nabla\mathcal{L}'(\mathbf{s}_k),\, \mathbf{s}_{k+1} - \mathbf{s}_k\rangle + \frac{M}{2}\|\mathbf{s}_{k+1} - \mathbf{s}_k\|^2. \tag{134}$$

Using $\mathbf{s}_{k+1} - \mathbf{s}_k = -\eta\hat{\mathbf{g}}_k$ and taking conditional expectation given $\mathcal{F}_k$, we obtain

$$\begin{aligned}\mathcal{L}'(\mathbf{s}_{k+1}) &\le \mathcal{L}'(\mathbf{s}_k) + \big\langle \nabla\mathcal{L}'(\mathbf{s}_k),\, \mathbf{s}_{k+1} - \mathbf{s}_k\big\rangle + \frac{M}{2}\|\mathbf{s}_{k+1} - \mathbf{s}_k\|^2 \\ &= \mathcal{L}'(\mathbf{s}_k) - \eta\big\langle \nabla\mathcal{L}'(\mathbf{s}_k),\, \hat{\mathbf{g}}_k\big\rangle + \frac{M\eta^2}{2}\|\hat{\mathbf{g}}_k\|^2.\end{aligned} \tag{135}$$

Define the filtration $\mathcal{F}_k$. From Assumption 5.2 (a),

$$\mathbb{E}[\hat{\mathbf{g}}_k \mid \mathcal{F}_k] = \nabla\mathcal{L}'(\mathbf{s}_k) + \mathbf{b}_k. \tag{136}$$

Let

$$\mathbf{g}_k := \nabla\mathcal{L}'(\mathbf{s}_k), \qquad \boldsymbol{\mu}_k := \mathbb{E}[\hat{\mathbf{g}}_k \mid \mathcal{F}_k] = \mathbf{g}_k + \mathbf{b}_k, \qquad \boldsymbol{\xi}_k := \hat{\mathbf{g}}_k - \boldsymbol{\mu}_k. \tag{137}$$

By Assumption 5.2 (b),

$$\mathbb{E}\big[\|\boldsymbol{\xi}_k\|^2 \mid \mathcal{F}_k\big] \le \sigma. \tag{138}$$

Taking conditional expectation in the smoothness inequality:

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}'(\mathbf{s}_{k+1}) \mid \mathcal{F}_k\big] &\le \mathcal{L}'(\mathbf{s}_k) - \eta\langle \mathbf{g}_k,\, \boldsymbol{\mu}_k\rangle + \frac{M\eta^2}{2}\mathbb{E}\big[\|\hat{\mathbf{g}}_k\|^2 \mid \mathcal{F}_k\big] \\ &= \mathcal{L}'(\mathbf{s}_k) - \eta\langle \mathbf{g}_k,\, \boldsymbol{\mu}_k\rangle + \frac{M\eta^2}{2}\mathbb{E}\Big[\|\boldsymbol{\mu}_k\|^2 + \|\boldsymbol{\xi}_k\|^2 + 2\langle \boldsymbol{\mu}_k, \boldsymbol{\xi}_k\rangle \mid \mathcal{F}_k\Big]. \\ &\le \mathcal{L}'(\mathbf{s}_k) - \eta\langle \mathbf{g}_k,\, \boldsymbol{\mu}_k\rangle + \frac{M\eta^2}{2}\mathbb{E}\Big[\|\boldsymbol{\mu}_k\|^2 + \sigma\Big].\end{aligned} \tag{139}$$

We use the identity

$$2\langle \mathbf{g}_k, \boldsymbol{\mu}_k\rangle = \|\mathbf{g}_k\|^2 + \|\boldsymbol{\mu}_k\|^2 - \|\mathbf{g}_k - \boldsymbol{\mu}_k\|^2. \tag{140}$$

Since $\mathbf{g}_k - \boldsymbol{\mu}_k = -\mathbf{b}_k$, this gives

$$-\eta\langle \mathbf{g}_k, \boldsymbol{\mu}_k\rangle = -\frac{\eta}{2}\|\mathbf{g}_k\|^2 - \frac{\eta}{2}\|\boldsymbol{\mu}_k\|^2 + \frac{\eta}{2}\|\mathbf{b}_k\|^2. \tag{141}$$

Substituting into the conditional bound:

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}'(\mathbf{s}_{k+1}) \mid \mathcal{F}_k\big] &\le \mathcal{L}'(\mathbf{s}_k) - \frac{\eta}{2}\|\mathbf{g}_k\|^2 - \frac{\eta}{2}\|\boldsymbol{\mu}_k\|^2 + \frac{\eta}{2}\|\mathbf{b}_k\|^2 + \frac{M\eta^2}{2}\|\boldsymbol{\mu}_k\|^2 + \frac{M\eta^2}{2}\sigma. \\ &= \mathcal{L}'(\mathbf{s}_k) - \frac{\eta}{2}\big(1 - M\eta\big)\|\boldsymbol{\mu}_k\|^2 - \frac{\eta}{2}\|\nabla\mathcal{L}'(\mathbf{s}_k)\|^2 + \frac{\eta}{2}\|\mathbf{b}_k\|^2 + \frac{M\eta^2}{2}\sigma.\end{aligned} \tag{142}$$

Because $\eta \le \frac{1}{M}$ (from the theorem's step-size condition), we have $1 - M\eta \ge 0$. Dropping the non-positive term $-\frac{\eta}{2}(1 - M\eta)\|\boldsymbol{\mu}_k\|^2 \le 0$ yields:

$$\mathbb{E}\big[\mathcal{L}'(\mathbf{s}_{k+1}) \mid \mathcal{F}_k\big] \le \mathcal{L}'(\mathbf{s}_k) - \frac{\eta}{2}\|\nabla\mathcal{L}'(\mathbf{s}_k)\|^2 + \frac{\eta}{2}\|\mathbf{b}_k\|^2 + \frac{M\eta^2}{2}\sigma. \tag{143}$$

From Assumption 5.1 (d) (PL condition for $\mathcal{L}'$):

$$\|\nabla\mathcal{L}'(\mathbf{s}_k)\|^2 \geq 2\mu\big(\mathcal{L}'(\mathbf{s}_k) - \mathcal{L}'^*\big). \tag{144}$$

Plug this into the inequality:

$$\Gamma_{k+1} \leq \Gamma_k - \frac{\eta}{2}\cdot 2\mu\Gamma_k + \frac{\eta}{2}\|\mathbf{b}_k\|^2 + \frac{M\eta^2}{2}\sigma. \tag{145}$$

That is:

$$\Gamma_{k+1} \leq (1-\eta\mu)\,\Gamma_k + \frac{\eta}{2}\|\mathbf{b}_k\|^2 + \frac{M\eta^2}{2}\sigma. \tag{146}$$

By Assumption 5.2 (c): $\|\mathbf{b}_k\|^2 \leq B^2$ for all k. Then:

$$\Gamma_{k+1} \leq (1-\eta\mu)\Gamma_k + \frac{\eta}{2}B^2 + \frac{M\eta^2}{2}\sigma. \tag{147}$$

Iterating from $k=0$ to $K-1$:

$$\begin{aligned}\Gamma_K &\leq (1-\eta\mu)^K\Gamma_0 + \left(\frac{\eta}{2}B^2 + \frac{M\eta^2}{2}\sigma\right)\sum_{j=0}^{K-1}(1-\eta\mu)^j \\ &\leq (1-\eta\mu)^K\big(\mathcal{L}'(\mathbf{s}_0) - \mathcal{L}'^*\big) + \frac{B^2}{2\mu} + \frac{M\eta}{2\mu}\sigma,\end{aligned} \tag{148}$$

where we used $\sum_{j=0}^{\infty}(1-\eta\mu)^j = \frac{1}{\eta\mu}$. Thus we obtain the final bound for T_1:

$$\mathbb{E}\big[\mathcal{L}'(\mathbf{s}_K) - \mathcal{L}'^*\big] \leq (1-\eta\mu)^K\big(\mathcal{L}'(\mathbf{s}_0) - \mathcal{L}'^*\big) + \frac{B^2}{2\mu} + \frac{M\eta}{2\mu}\sigma. \tag{149}$$

Then, we bound T_2. By Theorem 5.1, for any $\mathbf{s}$, we have

$$|\mathcal{L}(\mathbf{s}) - \mathcal{L}'(\mathbf{s})| \leq \Delta_\lambda(\mathbf{s}), \tag{150}$$

where

$$\Delta_\lambda(\mathbf{s}) = G_L\|\mathbf{s}\|_\infty\Psi_\lambda(\gamma(\mathbf{s}))\,. \tag{151}$$

Applying this to $\mathbf{s} = \mathbf{s}_K$ and taking expectation gives

$$\mathbb{E}\big[|\mathcal{L}(\mathbf{s}_K) - \mathcal{L}'(\mathbf{s}_K)|\big] \leq \mathbb{E}\big[\Delta_\lambda(\mathbf{s}_K)\big]. \tag{152}$$

Thus

$$\mathbb{E}\big[\mathcal{L}(\mathbf{s}_K) - \mathcal{L}'(\mathbf{s}_K)\big] \leq \mathbb{E}\big[\,|\mathcal{L}(\mathbf{s}_K) - \mathcal{L}'(\mathbf{s}_K)|\,\big] \leq \mathbb{E}\big[\Delta_\lambda(\mathbf{s}_K)\big] \tag{153}$$

Finally, we bound T_3. By Assumption 5.1 (c), we obtain $\mathcal{L}'^* = \inf_{\mathbf{s}}\mathcal{L}'(\mathbf{s}) \leq \mathcal{L}'(\mathbf{s}^*)$.

$$\mathcal{L}'^* - \mathcal{L}^* \leq \mathcal{L}'(\mathbf{s}^*) - \mathcal{L}(\mathbf{s}^*) \leq \big|\mathcal{L}'(\mathbf{s}^*) - \mathcal{L}(\mathbf{s}^*)\big| \leq \Delta_\lambda(\mathbf{s}^*). \tag{154}$$

Hence

$$\mathcal{L}'^* - \mathcal{L}^* \leq \Delta_\lambda(\mathbf{s}^*). \tag{155}$$

Substituting 149, 153 and 155 into 132, we obtain

$$\mathbb{E}\big[\mathcal{L}(\mathbf{s}_K) - \mathcal{L}^*\big] \leq (1-\eta\mu)^K\big(\mathcal{L}'(\mathbf{s}_0) - \mathcal{L}'^*\big) + \frac{B^2}{2\mu} + \frac{M\eta}{2\mu}\,\sigma + \mathbb{E}\big[\Delta_\lambda(\mathbf{s}_K)\big] + \Delta_\lambda(\mathbf{s}^*). \tag{156}$$

This completes the proof. □

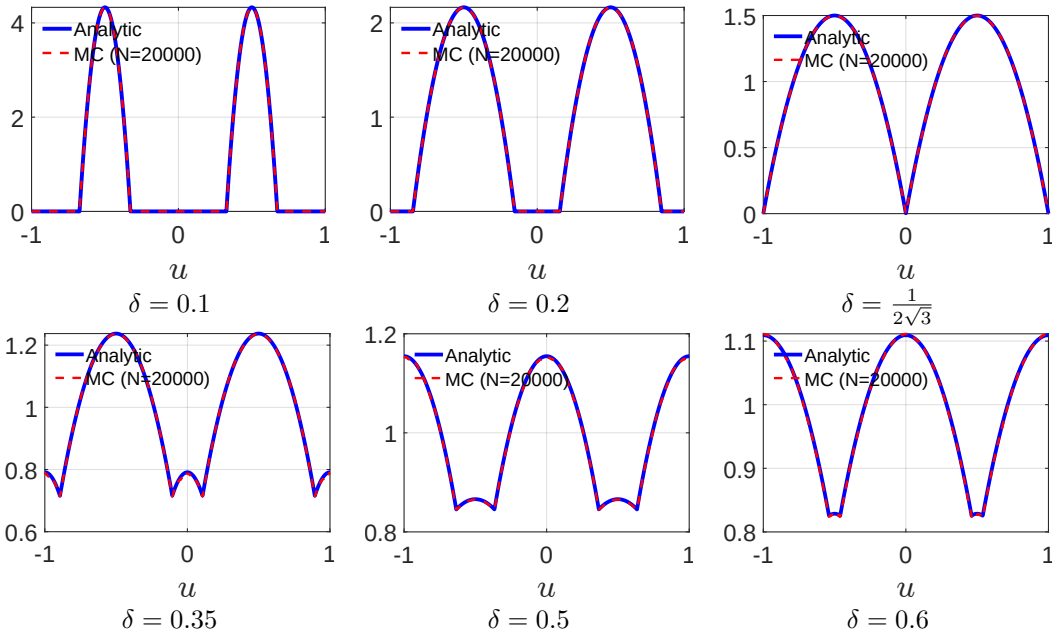


Figure 3: Zeroth-order expectation gradients under uniform perturbation for different values of δ.

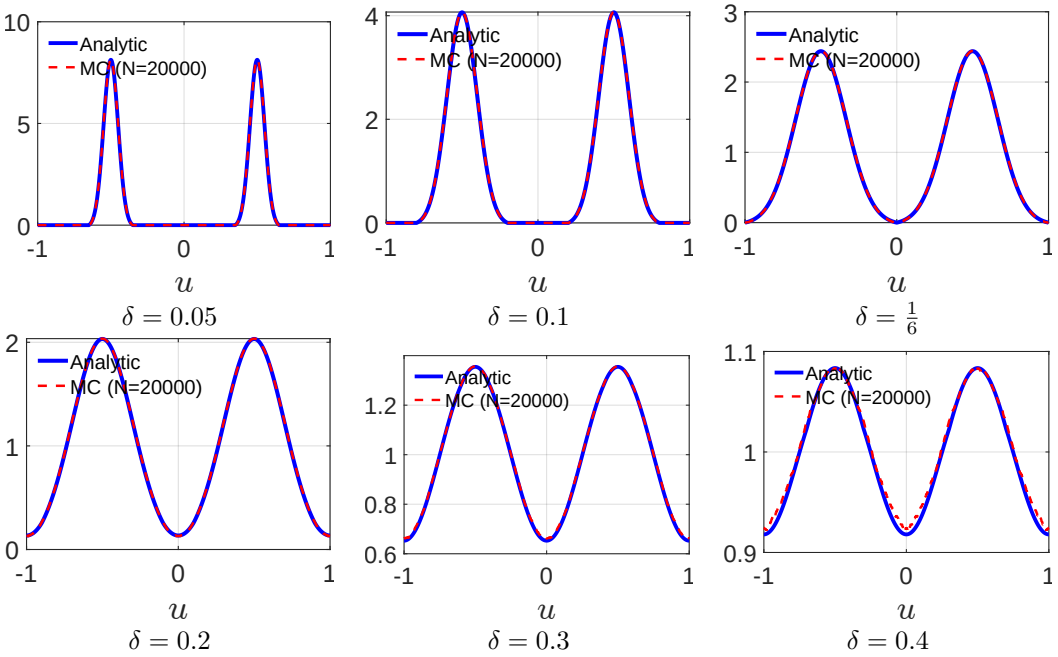


Figure 4: Zeroth-order expectation gradients under truncated Gaussian perturbation $C = 3$ for different values of δ.

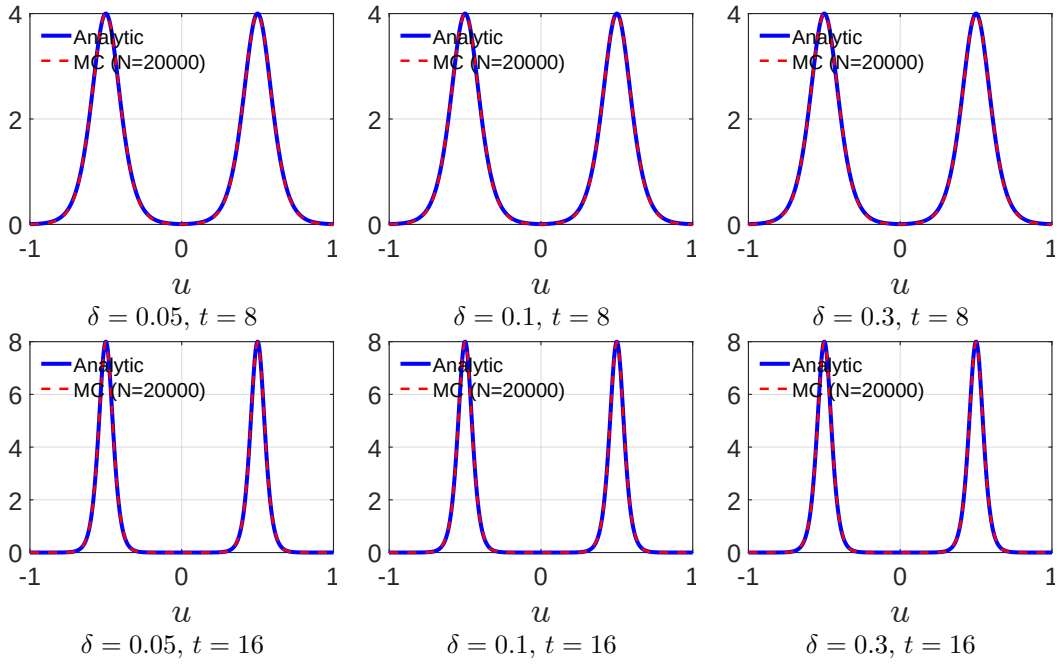


Figure 5: HTGE results for different choices of δ and t.

D Experiments

D.1 Numerical Verification of ZOE-Grad

Figures 3 and 4 provide additional visualizations of zeroth-order expectation gradients under uniform and truncated Gaussian perturbations, respectively, for a wider range of δ values, complementing the numerical verification results presented in Section 6.1.1. In addition, Figure 5 presents further HTGE results over a broader range of δ and t values, extending the numerical verification experiments discussed in Section 6.1.2.

D.2 Fine-tuning Large Language Models

D.2.1 Sensitivity Analysis of δ

We analyze the effect of the perturbation scale δ on OPT-1.3B under both uniform and truncated normal perturbations on the SST-2, RTE, WSC, and WiC tasks. As shown in Table 5, model performance first increases and then decreases as δ grows, reaching a peak at $\delta = 0.285$ for the uniform distribution and at $\delta = 0.15$ for the normal distribution. Both settings correspond to the case where $\mathcal{K}_u$ contains at most one boundary point. In addition, the results indicate that overly small values of δ lead to degenerate gradients due to insufficient boundary coverage, while overly large δ introduce noise from distant boundary points, breaking the local-perturbation assumption of the zeroth-order estimator and resulting in performance degradation. Therefore, values of δ near those satisfying the single-boundary condition of the index set $\mathcal{K}_u$ (i.e., $2\delta C = 1$) tend to yield superior performance.

D.2.2 Additional Results on W3A16 and W2A16 Quantization

We provide additional results under lower-bit quantization settings. Specifically, Table 6 reports the W3A16 results for LLaMA-2-7B, and Table 7 reports the W2A16 results for OPT-6.7B.

D.2.3 LLaMA-2-13B Results

Table 8 reports the complete W4A16 results for LLaMA-2-13B across the eight downstream tasks.

Table 5: Effect of the perturbation scale δ under uniform and truncated normal perturbations on OPT-1.3B

Model	Uniform					Normal				
	δ	SST-2	RTE	WSC	WIC	δ	SST-2	RTE	WSC	WIC
OPT-1.3B	0.07	89.79	49.82	53.85	51.25	0.05	89.51	45.85	54.81	52.82
	0.14	90.48	58.12	58.65	56.89	0.08	91.06	48.38	56.73	57.84
	0.21	90.48	61.37	62.50	56.58	0.10	91.28	52.71	59.62	58.46
	0.285	92.32	61.37	63.46	59.09	0.15	91.40	64.98	63.46	59.25
	0.40	91.63	47.29	61.54	56.89	0.20	90.48	61.01	63.46	58.78
	0.60	87.50	47.29	57.69	52.19	0.40	90.39	55.96	61.54	57.05
	0.80	80.50	46.21	53.85	50.63	0.60	90.94	54.15	59.62	57.52
	1.00	79.36	45.51	51.92	47.17	0.80	89.91	51.62	57.69	52.82

Table 6: Performance comparison of 3-bit weight-only quantization with 16-bit activations using different surrogate gradients on LLaMA-2-7B across downstream tasks.

Model	Method	SST-2	RTE	CB	BoolQ	WSC	WIC	MultiRC	SQuAD
LLaMA-2-7B	Zero-Shot	58.03	62.09	33.93	66.00	36.54	50.16	42.40	58.71
	Zero-Shot-Q	54.47	48.74	41.07	48.50	42.31	51.72	41.80	37.41
	STE	86.47	62.45	64.29	64.60	**65.38**	58.78	66.70	50.93
	HTGE	86.35	52.71	60.71	60.10	52.88	60.97	58.80	**67.39**
	Uniform	88.30	**66.43**	41.07	**73.90**	55.77	61.91	67.00	62.90
	Normal	**88.88**	52.71	**66.07**	69.10	54.81	**63.32**	**70.40**	66.60

Table 7: Performance comparison of 2-bit weight-only quantization with 16-bit activations using different surrogate gradients on OPT-6.7B across downstream tasks.

Model	Method	SST-2	RTE	CB	BoolQ	WSC	WIC	MultiRC	SQuAD
OPT-6.7B	Zero-Shot	61.24	54.87	53.57	57.30	37.50	51.25	41.70	40.37
	Zero-Shot-Q	50.92	47.29	41.07	41.00	50.96	49.69	41.40	0.01
	STE	49.08	50.18	60.71	59.00	56.73	50.47	41.20	0.18
	HTGE	53.78	52.71	**73.21**	**60.10**	**63.46**	50.00	**58.50**	**22.21**
	Uniform	**68.23**	52.71	69.64	**60.10**	56.73	**52.04**	41.20	18.51
	Normal	50.92	**53.07**	67.86	**60.10**	**63.46**	50.78	53.00	21.30

Table 8: Performance comparison of 4-bit weight-only quantization with 16-bit activations using different surrogate gradients on LLaMA-2-13B across downstream tasks.

Model	Method	SST-2	RTE	CB	BoolQ	WSC	WIC	MultiRC	SQuAD
LLaMA-2-13B	Zero-Shot	61.12	50.90	48.21	73.20	39.42	50.47	46.10	64.52
	Zero-Shot-Q	58.03	45.13	53.57	70.90	45.19	50.31	46.10	55.62
	STE	89.45	59.21	71.43	73.60	63.46	63.17	70.40	53.55
	HTGE	**91.74**	77.98	73.21	82.10	71.15	**68.81**	**81.60**	**63.88**
	Uniform	90.14	77.26	**82.14**	**83.20**	70.19	68.34	79.10	60.60
	Normal	90.83	**80.87**	78.57	81.80	**72.12**	68.50	80.10	61.61

D.2.4 Fine Tuning Overhead

Table 9 reports task level fine-tuning times on LLaMA-2-7B under W3A16. The additional computation of the continuous ZOE-based surrogates occurs only during backward computation and does not change inference cost.

Table 9: Task level fine tuning times in minutes and seconds on LLaMA-2-7B under W3A16.

Method	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	SQuAD	Average
STE	44:57	42:41	43:31	49:19	40:26	43:41	60:19	47:12	46:31
HTGE	55:11	54:54	55:56	60:16	52:53	55:09	66:01	56:09	57:04
Uniform	60:08	59:03	52:51	64:29	56:42	59:27	69:24	60:54	60:22
Normal	59:12	58:50	59:54	64:16	57:05	59:17	69:51	60:07	61:04

Table 10: Task specific per device batch sizes and learning rate candidate sets used in the fine tuning experiments.

Task	Batch size per device	Learning rate candidates
SST-2	8	$\{1e{-}6, 5e{-}6, 1e{-}5\}$
RTE	1	$\{5e{-}8, 1e{-}7, 5e{-}7, 5e{-}6\}$
CB	1	$\{8e{-}7, 1e{-}6, 5e{-}6, 1e{-}5\}$
BoolQ	1	$\{1e{-}7, 5e{-}7, 5e{-}5\}$
WSC	2	$\{8e{-}7, 1e{-}6, 1e{-}5, 1e{-}4\}$
WiC	4	$\{5e{-}7, 1e{-}6, 5e{-}6, 1e{-}5\}$
MultiRC	1	$\{5e{-}8, 1e{-}7, 5e{-}7, 5e{-}5, 1e{-}4\}$
SQuAD	1	$\{5e{-}6, 1e{-}5, 5e{-}5\}$

D.2.5 Task Specific Batch Sizes and Learning Rate Candidates

For all models and surrogate gradient methods, the learning rate is selected from the same task specific candidate set. The selected learning rate may differ across model and task pairs, but the candidate grid and selection protocol are shared across methods.

E Limitations

Although ZOE-Grad provides a unified perspective for understanding surrogate gradients, selecting the best surrogate for a specific model and downstream task still relies on heuristic criteria in practice. As shown in Section 6.2.3, different surrogate gradients may perform best across different LLM architectures and tasks. We use lightweight pilot fine-tuning as an empirical strategy for surrogate selection, but a principled theoretical criterion for choosing the optimal perturbation distribution remains an open problem. Developing such a criterion is an important direction for future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix E. In particular, we note that surrogate-gradient selection currently relies on heuristic pilot fine-tuning, and that a theoretical criterion for selecting the optimal perturbation distribution remains open.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper states the assumptions for each theoretical result and provides complete proofs in the appendix. In particular, the proofs of the main theoretical results are deferred to the corresponding appendix sections.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Section 6 and Appendix D.2.5 provide the experimental setup and implementation details, including datasets, models, baselines, training settings, and evaluation metrics. The code is publicly available at the link provided in the abstract.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: The code and reproduction instructions are publicly available at the link provided in the abstract. The experiments are based on publicly available datasets and models.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The training and evaluation details are specified in Section 6 and Appendix D.2.5, including datasets, models, baselines, training steps, sample sizes, evaluation metrics, task-specific batch sizes, and learning rate candidate sets.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests, as repeated fine-tuning of multiple large-scale LLMs under all settings would be computationally expensive. Nevertheless, the experiments cover multiple model architectures, downstream tasks, and perturbation distributions, providing broad empirical evidence for the main conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The experiments are conducted on a machine equipped with a cluster of NVIDIA A100-SXM4-80GB GPUs. Single-GPU task-level fine-tuning times are reported in Table 9 of Appendix D.2.4.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We conform to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

Justification: The paper does not explicitly discuss societal impacts, as its main contribution is a theoretical and empirical study of surrogate gradients for quantized optimization. The work may indirectly support more efficient LLM deployment, but we do not make specific societal impact claims.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: All datasets and models used in this work are publicly available, and we do not release new high-risk data or models.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We properly cite the original papers for the codebases, datasets, and models used in this work. We respect their licenses and terms of use.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We publicly release the implementation code for ZOE-Grad at the link provided in the abstract, together with documentation and instructions for reproducing the main experiments.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were used only for writing, editing, or formatting assistance and did not affect the core methodology, experiments, scientific rigor, or originality of the research.

Guidelines:

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