

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

Wenkang Wang, Jia Zhang, Zelin Wei,
Dongxu Liu, Jiabao Li, Wanli Shi, Bin Gu*
School of Artificial Intelligence, Jilin University, China
wangwk699@mails.jlu.edu.cn, gubin@jlu.edu.cn



Key Message:

ZOE-Grad unifies perturbation distributions and boundary-dependent surrogate gradients through a zeroth-order expectation framework, revealing STE as a degenerate discrete-perturbation case.

01 Motivation

Background: Quantization

- LLMs incur substantial memory and computational costs during deployment. **Quantization** reduces numerical precision to lower memory consumption and accelerate inference.
- **Post-Training Quantization (PTQ):** Quantizes pretrained models without full retraining, but may suffer accuracy degradation at low bit-widths.
 - **Quantization-Aware Training (QAT):** Incorporates quantization into training to better preserve model accuracy through gradient-based optimization.

Challenge: Non-Differentiable Rounding

Multi-bit quantization relies on the rounding operation:

$$r(u) = \text{round}(u), \quad r'(u) = 0 \quad \text{almost everywhere.}$$

Rounding has zero gradients almost everywhere and is non-differentiable at quantization boundaries. Consequently, **surrogate gradients are required** for backpropagation in QAT and gradient-based optimization of quantization parameters.

Limitation of Straight-Through Estimator (STE)

The widely adopted **STE** approximates the derivative of rounding as:

$$\frac{\partial \text{round}(u)}{\partial u} \approx 1.$$

However, this constant gradient ignores the distance between a value and nearby quantization boundaries:

$$\mathcal{B} = \left\{ k + \frac{1}{2} \in \mathbb{Z} \right\}.$$

Although existing methods introduce position-dependent gradient adjustments, a unified theoretical framework for systematically understanding and constructing surrogate gradients remains lacking.

Research question

Can we systematically characterize and construct surrogate gradients through perturbation distributions while capturing quantization-boundary information?

Contributions

- **Unified Framework.** ZOE-Grad establishes bidirectional mappings between perturbation distributions and surrogate gradients, revealing STE as a degenerate special case.
- **Theoretical Guarantees.** We establish convergence guarantees with bounded suboptimality for the original quantized objective.
- **Empirical Validation.** Simulations verify the theoretical framework, and LLM experiments demonstrate consistent improvements over STE using continuous perturbations.

02 The ZOE-Grad Framework

Zeroth-order expectation

We interpret quantization surrogate gradients through the expectation of a symmetric two-point zeroth-order estimator:

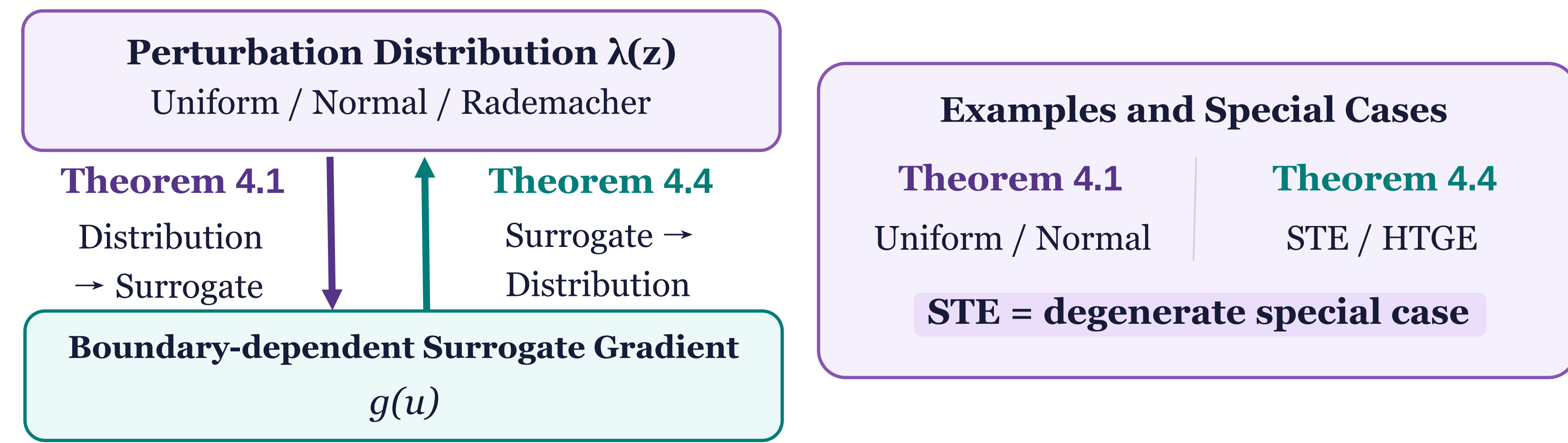
$$g(u) = \mathbb{E}_{z \sim \lambda} \left[\frac{r(u + \delta z) - r(u - \delta z)}{2\delta} z \right].$$

This formulation shows that **surrogate gradients can be understood as expectations induced by perturbation distributions**, providing a zeroth-order perspective on gradient approximation for rounding.

Bidirectional Mapping

ZOE-Grad builds a **bidirectional connection** between perturbation distributions and boundary-dependent surrogate gradients:.

- **Theorem 4.1: Distribution $\rightarrow$ Surrogate**
Given a perturbation distribution, derive the corresponding surrogate gradient. In particular, this yields the surrogate gradients induced by **Uniform** and **Normal** perturbations.
- **Theorem 4.4: Surrogate $\rightarrow$ Distribution**
Given an admissible boundary-dependent surrogate gradient, construct a perturbation distribution that recovers it in expectation. This explains existing surrogates such as **STE** and **HTGE**.



STE is a degenerate special case

Let the perturbation be a fixed-magnitude Rademacher variable:

$$\Pr(z = 1) = \Pr(z = -1) = \frac{1}{2}, \quad \delta = \frac{1}{2}.$$

Then

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

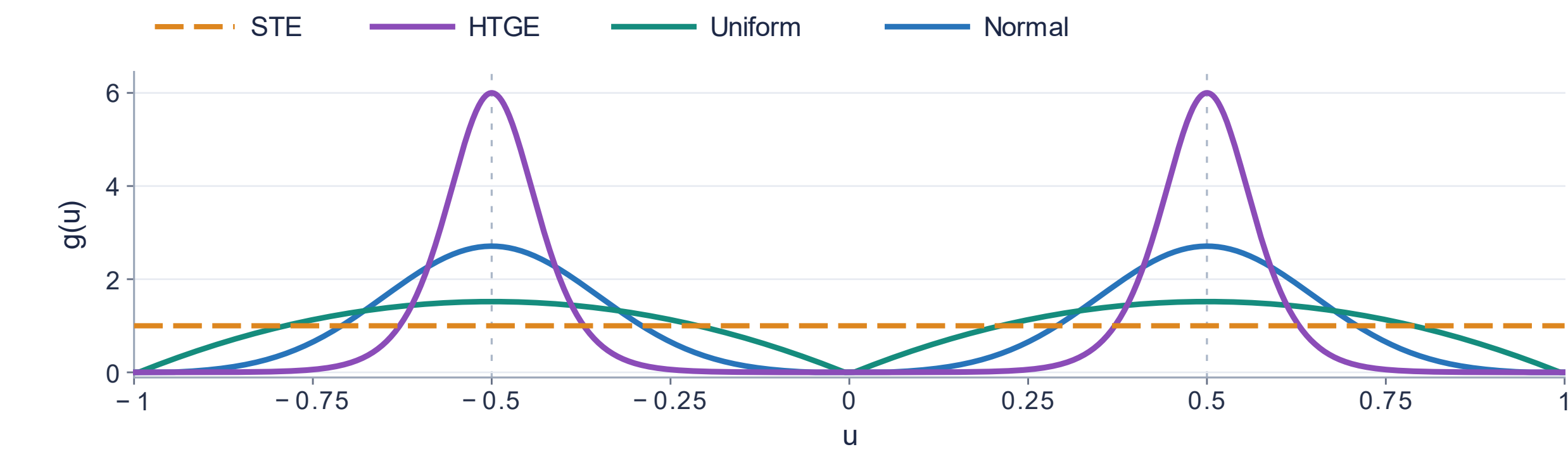
Thus, **STE corresponds to a discrete perturbation that keeps only sign information but discards magnitude information**, so it ignores the distance to nearby quantization boundaries.

Key Takeaway

ZOE-Grad unifies existing and new surrogate gradients through a bidirectional perturbation-surrogate mapping, while identifying STE as a degenerate discrete-perturbation case.

03 Experimental Evidence

Comparison of Four Surrogate Gradients



Unlike the constant STE, HTGE and continuous perturbation-based surrogates exhibit distinct boundary-dependent gradient profiles, capturing the local geometry of quantization boundaries.

W4A16 · SQuAD F1 $\uparrow$

Model	STE	HTGE	Uniform	Normal
OPT-1.3B	16.32	35.84	30.32	30.18
OPT-6.7B	41.55	54.79	54.40	54.38
LLaMA-2-7B	47.33	61.73	66.69	66.15
Qwen3-8B	56.88	65.18	70.29	64.21

Table 1 evaluates four LLMs across eight downstream tasks under W4A16, demonstrating that continuous ZOE-Grad surrogates generally outperform STE.

W4A4KV4 + rotation

Setting	Baseline	Uniform	Gain
Without rotation	20.76	21.99	+1.23
With rotation	21.15	23.28	+2.13

ZOE-Grad improves performance under W4A4KV4 both with and without rotation, demonstrating its compatibility with joint quantization of weights, activations, and KV cache, as well as rotation techniques.

Practical Insights

- **Surrogate Selection.** Pilot fine-tuning with 32 examples and 256 steps selects the best surrogate in 54.17% of cases and avoids the worst in 87.50% across 24 model-task pairs.
- **Inference Efficiency.** Additional computation is confined to backpropagation, introducing no inference-time overhead.

Key Takeaway

Continuous perturbations enable more expressive, boundary-aware surrogate gradients than STE, as supported by both theoretical analysis and LLM fine-tuning evidence.

github.com/wangwk699/ZOE_Grad