QUANTIZATION EVERYWHERE WITH ROTATED FOURIER SURROGATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization is essential for efficient neural network deployment, whereas its optimization overwhelmingly relies on the Straight-Through Estimator (STE). However, STE discards the structure of the quantization operator by replacing its derivative with an identity mapping, resulting in a mismatch between the forward and backward passes that can fundamentally limit the achievable quality of quantized models. To address this limitation, we introduce Rotated Fourier Surrogate (RFS), a unified framework that extends STE with a principled gradient estimator. RFS analytically reflects the structure of quantization while remaining stable and computationally efficient. We further provide rigorous theoretical guarantees for RFS, demonstrating well-conditioned gradients and favorable optimization properties. Moreover, RFS is plug-and-play and incurs negligible computational overhead with our optimized kernel. We extensively evaluate RFS across diverse quantization paradigms, spanning integer and floating-point quantization, weight and activation quantization, as well as quantization-aware training, quantization-aware distillation, and post-training quantization. Our experiments cover a broad spectrum of architectures, including large language models, diffusion language models, vision transformers, and convolutional neural networks, as well as diverse applications including generative recommendation, image classification, pre-training, and reasoning post-training. Across these diverse settings, RFS consistently improves upon strong state-of-the-art quantization baselines, demonstrating its broad applicability across quantization paradigms, model architectures, and applications.

1 INTRODUCTION

Neural network quantization is essential for efficient deployment, spanning integer and floating-point formats, weight and activation quantization, and paradigms including quantization-aware-training (QAT), post-training-quantization (PTQ), and quantization-aware-distillation QAD (Frantar et al., 2023; Xiao et al., 2023; Liu et al., 2024a; Du et al., 2024b). Despite these differences, they share a common challenge: discrete quantization must be optimized with continuous gradients. Since rounding is non-differentiable, the dominant Straight-Through Estimator (STE) (Bengio et al., 2013b; Hubara et al., 2018) replaces its derivative with identity. While simple and efficient, STE ignores the quantization structure, creating a *forward–backward mismatch* that becomes increasingly problematic at aggressive low precision (e.g., sub-4-bit) and ultimately limits quantized model performance.

There exist several alternatives to hard quantization based on different relaxations. Methods using soft rounding, smooth approximations, stochastic, or quantization noise can provide richer gradient information than STE (Gong et al., 2019b; Défossez et al., 2022; Huang et al., 2022; Semenov, 2025). However, these approaches often incur additional computational overhead and implementation complexity, are tailored to specific quantization settings, or suffer from sub-optimal optimization behavior due to ill-conditioned surrogate gradients. This naturally leads to the following question:

> *Can we construct a quantization-aware gradient surrogate that explicitly captures the discrete structure of quantization and resolves the forward–backward mismatch of STE, while preserving the **simplicity**, **generality**, **high performance**, and **computational efficiency**?*

We answer this question with **Rotated Fourier Surrogate (RFS)**. At core, we geometrically observe that a 45° coordinate rotation transforms the non-differential rounding function into a continuous

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

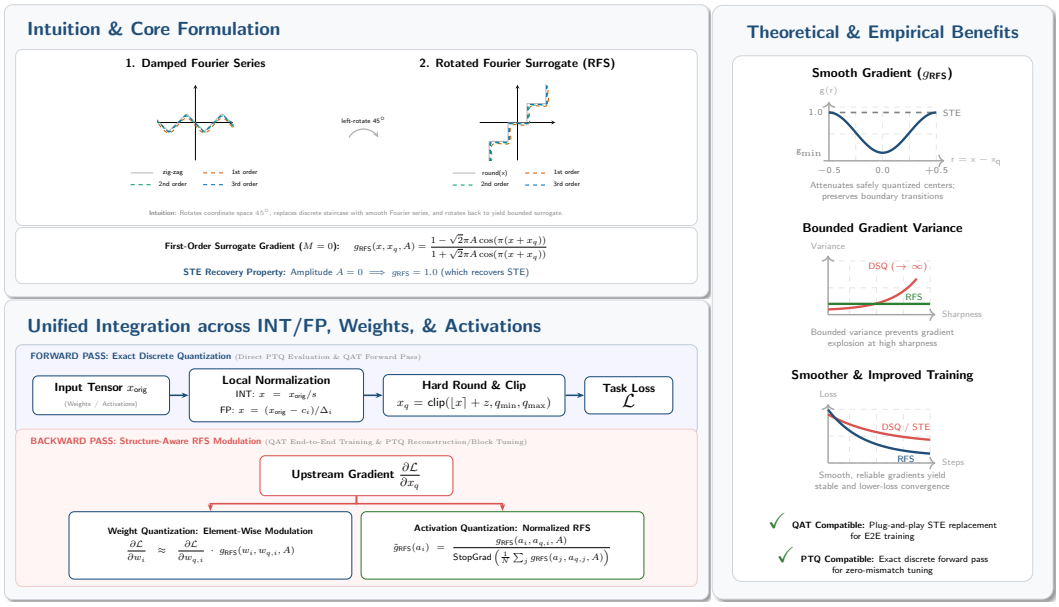


Figure 1: Overview of our proposed Rotated Fourier Surrogate (RFS), methodology, and its features.

periodic waves. This rotated function admits an analytic Fourier representation, allowing the structure of rounding to be encoded into a smooth surrogate gradient. Moreover, when the Fourier amplitude is set to zero, RFS reduces exactly to STE. Thus, RFS can be viewed as a conceptually strict generalization of STE that continuously interpolates with a structured, quantization-aware gradient. RFS is broadly applicable across INT and floating-point quantization, weight and activation quantization, and different optimization paradigms including QAT, QAD, and PTQ.

Our contributions are summarized as follows:

- **A unified gradient surrogate for quantization.** We introduce the Rotated Fourier Surrogate (RFS), which derives a structured surrogate gradient directly from a rotated Fourier representation of the rounding operator. RFS conceptually generalizes STE, recovering it at zero Fourier amplitude.
- **Unified integer/floating-point and weight/activation quantization.** RFS adopts cross integer (INT) and floating-point (FP) quantization through a common locally normalized rounding formulation. For weights, RFS gives heterogeneous element-wise gradient modulation. For activations, a normalized RFS stabilizes gradients of shared quantization parameters.
- **Well-conditioned and theoretically grounded.** We characterize RFS's conditioning as a function of its Fourier amplitude and derive an interpretable well-conditioned operating region, establish its approximation advantage over STE, analyze its alignment with a smoothed reference gradient, and show bounded gradient variance even in the near-hard-rounding limit.
- **Broad applicability at negligible cost.** RFS is a plug-and-play extension of STE. Experiments across QAT, QAD, and PTQ with INT and FP, weight and activation, over multiple model families, and diverse downstream tasks show consistent gains upon strong baselines with retained efficiency.

2 Related Work and Preliminaries

We review preliminaries and related work, and defer discussion on additional works to Appendix A.

Quantization maps high-precision floating values (e.g., FP32 or FP16) into low-precision discrete representations (e.g., INT8 or INT4), thereby reducing both memory footprint and compute costs at inference time. Given a floating-point variable x_{original} (e.g., weights or activations) and a target bit-width b, a quantizer maps x_{original} to a domain $[q_{\min}, q_{\max}]$. Quantization can be parameterized by a scaling factor $s \in \mathbb{R}^+$ and a zero-point $z \in \mathbb{Z}$:

$$x_q = \text{clip}\left(\text{round}(x_{\text{original}}/s) + z, q_{\min}, q_{\max}\right) = \text{clip}\left(\text{round}(x) + z, q_{\min}, q_{\max}\right), \quad (1)$$

where $\mathrm{round}(\cdot)$ denotes the round-to-nearest-integer operator, and $\mathrm{clip}(v, a, b) = \min(\max(v, a), b)$ enforces the feasible range constraint. We denote $x \triangleq x_{\text{original}}/s$ for simplicity. Here, the scaling factor s (also referred to as the quantization step size) determines the quantization resolution, while the zero-point z governs the alignment between floating-point and integer domains.

Straight-Through Estimator (STE). The rounding operation in Equation 2 is non-differentiable, with zero derivative almost everywhere, preventing informative gradients from being propagated through quantization. STE (Bengio et al., 2013a), widely used in QAT and PTQ, circumvents this issue by ignoring the rounding operation and approximating $\partial x_q/\partial x$ with an identity function. Accordingly, the gradient of the objective $\mathcal{L}$ with respect to x is approximated as:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial x_q} \cdot \boxed{\frac{\partial x_q}{\partial x}}_{\text{STE}} \approx \frac{\partial \mathcal{L}}{\partial x_q}, \tag{2}$$

While simple and efficient, STE completely ignores the structure of the rounding operator, creating a substantial forward-backward mismatch that can lead to biased gradient updates, impaired convergence, and sub-optimal performance.

Characteristic	RFS	DSQ	STE	EWGS
Theoretical Bounded	✓	✗	✗	✗
Computation Efficiency	✓	✗	✓	✗
Performance and Stability	✓	✗	✗	✗
Generality (INT/FP, Weight/Act)	✓	✗	✓	✗

Alternatives to STE improve gradient estimation through smooth relaxation, stochastic approximation, or gradient correction. Soft-surrogate methods such as DSQ (Gong et al., 2019b) and related approaches (Semenov, 2025) replace rounding with parameterized `tanh` or `sigmoid` relaxations to capture richer quantization structure. However, increasingly sharp relaxations can lead to ill-conditioned optimization and introduce non-negligible computational overhead. Noise-based methods such as NICE (Baskin et al., 2021) and DiffQ (Défossez et al., 2021) approximate quantization errors with random noise, yielding smoother optimization but introducing train–deployment mismatch by replacing deterministic quantization with a distributional proxy. EWGS (Lee et al., 2021) retains hard quantization and corrects STE gradients using local quantization errors, but relies on heuristic or approximate second-order information rather than an explicit surrogate of the quantization operator. Overall, existing alternatives either compromise optimization stability, incur additional computational cost, or only indirectly capture quantization structure. This motivates RFS, which directly captures the quantization structure while providing a simple, efficient, plug-and-play, and well-conditioned gradient surrogate.

3 Rotated Fourier Surrogate

We propose **Rotated Fourier Surrogate (RFS)**, a principled gradient surrogate to address the forward–backward mismatch of the Straight-Through Estimator (STE). RFS builds on a key observation: a $45°$ rotation turns the staircase rounding function into a continuous periodic function with an analytic Fourier representation, yielding a smooth, well-conditioned surrogate gradient (Section 3.1-3.2). Notably, RFS strictly generalizes the STE surrogate, recovering it as a special case. Unlike prior methods tailored to either integer or floating-point quantization, our RFS formulation covers uniform integer and non-uniform floating-point representations in one framework (Section 3.3). We instantiate it for both weight and activation quantization, and introduce a normalized variant that mitigates the gradient distortion caused by aggregation over heterogeneous activation distributions (Section 3.4).

3.1 Rotated Fourier Surrogate Extends STE

Coordinate Rotation. Consider the rounding relation $x_q = \mathrm{round}(x)$, where x is the normalized pre-quantized variable defined in the preliminary section. The graph of $x_q = \mathrm{round}(x)$ forms a staircase-shaped function and is discontinuous in the original (x, x_q) coordinate system.

Our key observation is that applying a $45°$ counterclockwise rotation transforms this staircase into a continuous and periodic piecewise-linear function. Specifically, we introduce the rotated coordinates $\left(t = (x + x_q)/\sqrt{2},\ f(t) = (x_q - x)/\sqrt{2}\right)$. Under this transformation, the rounding function becomes a periodic triangle wave with fundamental period $T = \sqrt{2}$. The transformed function is expressed as

$$f(t) = \frac{1}{2\sqrt{2}} \left(1 - 4\left|\left\{\frac{t - T/4}{T}\right\} - \frac{1}{2}\right|\right), \tag{3}$$

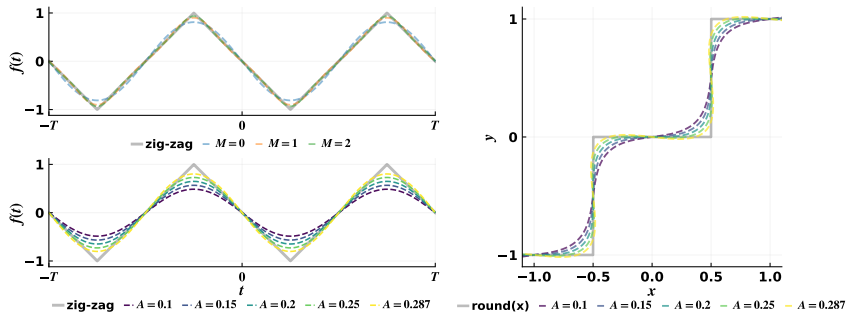


Figure 2: Rotated Fourier Surrogate (RFS) under different orders and amplitudes. Higher orders improve approximation fidelity but increase computational complexity. A 1st order truncation ($M = 0$) can already capture the essential structure while providing stable and efficient optimization.

where $\{\cdot\}$ denotes the fractional-part operator. The rotation therefore converts the discontinuous staircase-shaped quantizer into a continuous periodic signal whose structural information can be characterized analytically through Fourier analysis.

Fourier Representation. Since the rotated function in Equation 3 is periodic and square-integrable, it admits a Fourier series representation:

$$f(t) = -A \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)^2} \sin\Big((2m+1)\sqrt{2}\pi t\Big), \tag{4}$$

where the exact Fourier representation corresponds to $A = \frac{2\sqrt{2}}{\pi^2}$. Rather than fixing A to the exact Fourier amplitude, we treat it as a damping parameter that controls the strength of the rounding-aware modulation. As Section 3.2, this damping directly controls the conditioning of the resulting surrogate gradient. For practical computation, we truncate the Fourier expansion after $M + 1$ components:

$$f_M(t) = -A \sum_{m=0}^{M} \frac{(-1)^m}{(2m+1)^2} \sin\Big((2m+1)\sqrt{2}\pi t\Big). \tag{5}$$

Higher-order terms encode finer structural details of the rounding function, while low-order approximations provide a favorable trade-off between structural fidelity and computational efficiency.

Inverse Rotation and Surrogate Gradient. We next map the Fourier representation back to the original (x, x_q) coordinates. Substituting the truncated Fourier expansion and implicitly differentiating with respect to x yields the RFS surrogate gradient

$$g_{\text{RFS}}(x, x_q, A, M) = \frac{1 - A\sqrt{2}\pi \sum_{m=0}^{M} \frac{(-1)^m}{2m+1} \cos\left((2m+1)\pi(x + x_q)\right)}{1 + A\sqrt{2}\pi \sum_{m=0}^{M} \frac{(-1)^m}{2m+1} \cos\left((2m+1)\pi(x + x_q)\right)}. \tag{6}$$

In practice, we use the first-order approximation with $M = 0$, which already captures the essential structure of the rounding operator while introducing negligible computational overhead. This gives

$$\boxed{g_{\text{RFS}}(x, x_q, A) = \frac{1 - \sqrt{2}\pi A \cos\left(\pi(x + x_q)\right)}{1 + \sqrt{2}\pi A \cos\left(\pi(x + x_q)\right)}} \tag{7}$$

Accordingly, RFS replaces the identity surrogate used by STE in the backward pass:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial x_q} \cdot \boxed{\frac{\partial x_q}{\partial x}}_{\text{RFS}} \approx \frac{\partial \mathcal{L}}{\partial x_q} \cdot g_{\text{RFS}}(x, x_q, A). \tag{8}$$

STE as a Special Case. When $A = 0$, Equation 7 reduces to $g_{\text{RFS}}(x, x_q, A = 0) = 1$, which exactly becomes the STE. Therefore, STE naturally emerges as a special case of RFS, whereas $A > 0$ introduces structured, quantization-aware modulation into the backward gradient.

3.2 Well-Conditioned Surrogate Gradients

Effect of Amplitude. The amplitude A controls the sharpness of RFS and how closely it approximates the true rounding operator. When A is small (approaching STE as $A \to 0$), the surrogate is overly smooth, which weakens its ability to capture the fine-grained structure and leads to biased or ineffective gradient signals. Increasing A improves approximation accuracy by bringing RFS closer to rounding. However, as it sharpens, it progressively inherits the pathological optimization behavior of rounding: gradients vanish over large regions and become unstable near discontinuities. This trade-off motivates the use of a damped amplitude to balance expressiveness and numerical stability.

Well-Conditioned Surrogate Region. With the above discussion, we treat A as a tunable parameter and determine its practical range via conditioning analysis Nocedal & Wright (2006). Considering a similar condition number as the ratio of the largest to the smallest value of g_{RFS}, we obtain

$$\kappa \coloneqq \frac{\max g_{\mathrm{RFS}}(\cdot)}{\min g_{\mathrm{RFS}}(\cdot)} = \frac{1+\sqrt{2}\pi A}{1-\sqrt{2}\pi A} \quad \Longrightarrow \quad A = \frac{1}{\sqrt{2}\pi} \cdot \frac{\kappa - 1}{\kappa + 1},$$

Since A increases monotonically with κ, imposing $\kappa \le 100$ yields $0 \le A \le 0.2206$. By default, we choose the conservative conditioning number κ as 10 associated with $A \approx 0.184$.

3.3 Unified RFS for INT and Floating-Point Quantization

We now show that RFS is applicable for both INT and FP quantization with the same formulation. Considering an arbitrary ordered codebook $\mathcal{C} = \{c_1 < c_2 < \cdots < c_K\}$, for two adjacent c_i and c_{i+1}, denote their local spacing by $\Delta_i = c_{i+1} - c_i$. Within this interval, we define the normalized pre-quantized variable used by RFS as $x \coloneqq (x_{\text{original}} - c_i)/\Delta_i, x_q = \mathrm{round}(x)$, where x_{original} is the value in the original quantization domain. For INT quantization, Δ_i is constant across the codebook. For FP quantization, Δ_i varies across representable values due to the exponent-mantissa structure, and equation 7 is applied locally to each pair of adjacent levels. Nevertheless, after normalization, both formats share the identical rounding relation $x_q = \mathrm{round}(x)$. Thus, the distinction between INT and FP quantization lies only in how the normalized variable x is constructed: INT admits a global uniform normalization, whereas FP requires a local normalization determined by its non-uniform codebook spacing. Once expressed in the (x, x_q) coordinates, RFS formulation remains unchanged.

3.4 RFS for Weight and Activation Quantization

Weight Quantization. In weight quantization, every weight has its own quantization state, so RFS injects an element-wise scaling factor into the backward pass: $\frac{\partial \mathcal{L}}{\partial x_i} \approx \frac{\partial \mathcal{L}}{\partial x_{q,i}} g_{\mathrm{RFS}}(x_i, x_{q,i}, A)$, where g_{RFS} is defined in equation 7. Each weight receives a distinct scalar set by its position relative to its quantization level, making the backward signal quantization-aware. At $A = 0$ this reduces to STE, which assigns unit gradient to all weights irrespective of quantization state.

Activation Quantization. Activation quantization presents a different optimization structure. Instead of treating activation values themselves as learnable variables, quantization typically optimizes shared quantization parameters, such as the activation quantization step size. The gradient of such a parameter aggregates contributions from a large number of activation elements, e.g.,

$$\frac{\partial \mathcal{L}}{\partial s} = \sum_i \frac{\partial \mathcal{L}}{\partial a_{q,i}} \frac{\partial a_{q,i}}{\partial a_i} \frac{\partial a_i}{\partial s} = \sum_i \frac{\partial \mathcal{L}}{\partial a_{q,i}} \frac{\partial a_i}{\partial s} g_{\mathrm{RFS}}(a_i, a_{q,i}, A), \tag{9}$$

where s denotes the quantization step size and a is the activation and a_q is the quantized instance. When RFS is used in place of STE, each activation contribution is additionally modulated by its corresponding RFS scalar. Unlike weights, however, activation distributions may vary substantially across samples, tokens, and batches, and can contain pronounced outliers. Upon the aggregation of these heterogeneous activation signals, a small subset of atypical activations may induce large modulation and dominate the update. To mitigate this effect, we introduce a normalized RFS variant for activation quantization to normalize them by their empirical mean:

$$g^{\mathrm{act}}_{\mathrm{RFS}}(a_i, a_{q,i}, A) = \frac{g_{\mathrm{RFS}}(a_i, a_{q,i}, A)}{\mathrm{StopGrad}\left(\frac{1}{N}\sum_{j=1}^{N} g_{\mathrm{RFS}}(a_j, a_{q,j}, A)\right)}. \tag{10}$$

The normalized scalars satisfy $\frac{1}{N}\sum_{i=1}^{N} g_{\text{RFS}}^{\text{act}}(a_i, a_{q,i}, A) = 1$, thereby preserving the average gradient scale while retaining the relative heterogeneity among individual activation signals.

4 Theoretical and Efficiency Analysis for RFS

4.1 Gradient alignment comparison between STE and RFS

Following the population/reference gradient paradigm of (Yin et al., 2019), we show that RFS provides a more accurate approximation to a smoothed reference gradient than STE. The true quantized loss $\mathbb{E}_{v\sim\mathcal{D}}[\mathcal{L}(Q(x); v)]$ is piecewise constant in x, so its gradient is zero or undefined almost everywhere. To obtain a meaningful reference gradient, we apply Gaussian smoothing Nesterov & Spokoiny (2017) and obtain the following reference gradient $\tilde{\mathcal{L}}(x) = \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I_d),\, v\sim\mathcal{D}}\, \mathcal{L}(Q(x+u); v)$, where v is a data sample from distribution $\mathcal{D}$, u is a Gaussian perturbation, and $\sigma > 0$ controls the smoothing bandwidth. The gradient $\nabla\tilde{\mathcal{L}}(x)$ captures the local landscape of the quantized loss, making it an informative reference gradient, although computing it exactly is generally intractable in practice.

Theorem 4.1 (RFS gradient alignment with reference gradient)**.** *Suppose that $\mathbb{E}[\mathcal{L}(\cdot; v)]$ is L_1-Lipschitz and $\mathbb{E}[\nabla\mathcal{L}(\cdot; v)]$ is L_2-Lipschitz. Set $\sigma = \sqrt{L_1\epsilon_M/L_2}$, where ϵ_M is the upper-bound deviation between RFS's M-order quantizer and the true quantizer.*
(i) ***(Error bound)*** $\|\nabla\mathcal{L}_{\text{RFS}}(x) - \nabla\tilde{\mathcal{L}}(x)\|_2 \le 2\sqrt{L_1 L_2 d\epsilon_M}$.
(ii) ***(Cosine similarity)*** $\cos(\nabla\mathcal{L}_{\text{RFS}}(x), \nabla\tilde{\mathcal{L}}(x)) \ge 1 - 4\sqrt{L_1 L_2 d\epsilon_M}/\|\nabla\tilde{\mathcal{L}}(x)\|_2$.

Remark I. Proof is present in Appendix C.3. As M increases, ϵ_M converges to 0, and the gradient error $\|\nabla\mathcal{L}_{\text{RFS}} - \nabla\tilde{\mathcal{L}}\|_2 \le 2\sqrt{L_1 L_2 d\,\epsilon_M}$ vanishes. Moreover, any $M \ge 1$ yields a strictly smaller ϵ_M than $M = 0$ (STE), so RFS achieves a strictly tighter error bound than STE for every positive Fourier order. In addition, as M increases, the cosine similarity approaches 1, meaning RFS becomes asymptotically collinear with the reference gradient $\nabla\tilde{\mathcal{L}}(x)$. Following (Yin et al., 2019), RFS serves as a better optimization direction than STE.

4.2 Comparison between DSQ and RFS

Next, we compare DSQ and RFS in terms of gradient stability by analyzing the statistical properties of surrogate gradients over the clipping range. Our analysis focuses on the asymptotic regime in which the surrogate function closely approximates the rounding function.

Theorem 4.2. *Let $g_{\text{DSQ}}(\cdot;\alpha)$ and $g_{\text{RFS}}(\cdot; A)$ be the gradient surrogates of DSQ and RFS, parameterized by $\alpha > 0$ and $A \in \left(0, 1/(\sqrt{2}\pi)\right)$, respectively. Let random variable ξ be uniformly distributed on the interval $[l, u]$, i.e., $\xi \sim U(l, u)$. Consider the asymptotic regimes in which the surrogate functions move close to the rounding function. Then, the limits of the expectations satisfy*

$$\lim_{\alpha\to 0^+} \mathbb{E}_{\xi\sim U(l,u)}[g_{\text{DSQ}}(\xi;\alpha)] = 1, \quad \lim_{A\to(1/(\sqrt{2}\pi))^-} \mathbb{E}_{\xi\sim U(l,u)}[g_{\text{RFS}}(\xi; A)] \quad = 4/\pi - 1,$$

and the limits of the variances satisfy

$$\lim_{\alpha\to 0^+} \text{Var}_{\xi\sim U(l,u)}[g_{\text{DSQ}}(\xi;\alpha)] = \infty, \quad \lim_{A\to(1/(\sqrt{2}\pi))^-} \text{Var}_{\xi\sim U(l,u)}[g_{\text{RFS}}(\xi; A)] \quad = 16/(3\pi) - 16/\pi^2.$$

Remark II. Proof is present in Appendix C.4. Theorem 4.2 demonstrates a fundamental limitation of DSQ, i.e., DSQ's gradient variance diverges as it sharpens towards the rounding function, whereas RFS maintains bounded variance (≈ 0.076) at maximum sharpness. This distinction directly impacts training stability. Bounded variance ensures consistent gradient magnitudes, while divergent variance may cause gradient explosions. Therefore, gradient stability is a key advantage of RFS over DSQ as presented in Figure 3 and numerical validation in Appendix B.4.

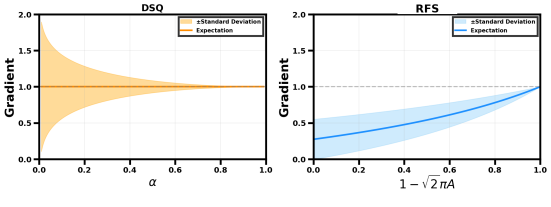


Figure 3: **Gradient spread of DSQ and RFS.** As the surrogate sharpens towards the rounding operator, DSQ exhibits exploding variance, while RFS shows bounded variance.

Table 1: QAT on LLaMA-3.2-1B in continuous pre-training for weight-only INT quantization.

W/A	Method	Setting	Arc-e	Arc-c	Boolq	Hellaswag	Openbookqa	Piqa	SciQ	Winogrande	Best of 5 Avg	Δ	Mean ± Std
16/16	Baseline	Baseline	60.61	36.01	63.70	63.78	36.60	74.43	88.40	60.06	60.45	–	–
4/16	ParetoQ (Liu et al., 2025)	20B Tokens & lr=1e-4	65.49	36.86	63.82	61.22	39.40	74.70	86.70	59.51	60.96	–	60.58 ± 0.36
4/16	DSQ (Gong et al., 2019a)	20B Tokens & lr=1e-4	65.45	37.37	63.73	62.04	39.40	74.16	86.40	59.98	61.07	–	59.93 ± 0.72
4/16	**RFS** (upon ParetoQ)	20B Tokens & lr=1e-4	65.74	37.54	64.01	61.15	40.00	74.59	86.40	60.46	**61.24**	+0.25	60.64 ± 0.35
3/16	ParetoQ (Liu et al., 2025)	20B Tokens & lr=2e-4	64.24	35.69	61.23	58.91	37.40	73.32	87.50	58.27	59.57	–	58.80 ± 0.63
3/16	DSQ (Gong et al., 2019a)	20B Tokens & lr=2e-4	60.77	32.85	59.76	56.21	36.60	72.42	83.10	57.62	57.42	–	52.14 ± 10.71
3/16	**RFS** (upon ParetoQ)	20B Tokens & lr=2e-4	64.06	36.43	63.24	59.49	37.80	73.67	86.7	59.59	**60.12**	+2.70	59.93 ± 0.14
2/16	ParetoQ (Liu et al., 2025)	30B Tokens & lr=1e-4	60.02	34.22	57.65	55.63	35.80	72.91	82.90	59.19	57.29	–	52.14 ± 10.71
2/16	DSQ (Gong et al., 2019a)	30B Tokens& lr=1e-4	59.13	32.59	58.72	55.65	36.20	72.03	80.50	56.27	56.39	–	44.86 ± 13.67
2/16	**RFS** (upon ParetoQ)	30B Tokens& lr=1e-4	61.53	32.94	63.00	56.24	37.40	72.63	83.10	57.77	**58.08**	+1.69	54.73 ± 5.48

4.3 Efficiency Comparison

We benchmark RFS against STE, DSQ, and EWGS on LLaMA-3.2-1B with batch size 4 and sequence length 128. As shown in Figure 4, RFS nearly matches STE regarding forward and backward latency and peak memory. DSQ contains a multi-step gradient estimator, increasing both forward and backward cost. EWGS incurs substantial backward cost from Hessian approximation, resulting in **3.04× higher peak backward memory** than RFS and STE. These costs can limit their use in modern large models. Consequently, RFS provides a plug-and-play extension of STE with negligible computational overhead.

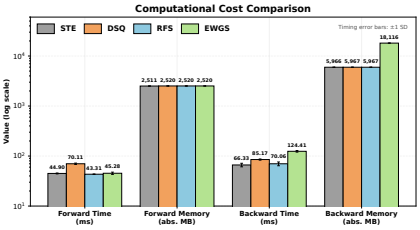


Figure 4: Time and memory comparison on batch size as 4 and sequence length as 128. Lower is better.

5 Experiments

We evaluate RFS across four representative quantization scenarios: *(i)* QAT for generative recommendation, *(ii)* QAT during continuous pre-training of LLMs, *(iii)* QAT for post-training reasoning, and *(iv)* PTQ for image classification. These settings span diverse model architectures, quantization formats, and optimization paradigms, providing a broad evaluation of the effectiveness and generality of RFS. We further examine its optimization behavior through training-stability analysis and ablations on key design choices. Additional results on quantization-aware distillation (QAD) and vision transformers (ViTs) are provided in Appendix B.2 and Appendix B.3, respectively.

Experimental Setup. RFS directly extends the STE and can be seamlessly integrated into existing quantization methods in a plug-and-play manner. For each setting, we adopt a strong state-of-the-art baseline and follow its original experimental implementation or our faithful reproduction. We replace only the gradient estimator with RFS while keeping all other components unchanged. This controlled setup isolates the contribution of RFS and ensures a fair and thorough comparison.

5.1 QAT on Continuous Pre-Training

We first evaluate RFS in the continuous pre-training setting by performing low-bit QAT on LLMs over large-scale pre-training corpora. Following ParetoQ, we conduct weight-only integer QAT on LLaMA-3.2-1B (Meta, 2024) and LLaMA-3.2-3B (Appendix B.1) under 2–4 bits. We compare against two representative surrogate paradigms: ParetoQ (Liu et al., 2025), a state-of-the-art STE-based QAT method, and DSQ (Gong et al., 2019a), a canonical multi-step soft-rounding approach. We exclude EWGS (Lee et al., 2021), another STE alternative, as its Hessian approximation leads to out-of-memory when applied to the experimental setting established by ParetoQ (see Section 4.3). The training corpus mixes SlimPajama (Soboleva et al., 2023) and FineWeb-Edu (Penedo et al., 2024) at a 1:1 ratio, with the token budget varied across quantization settings. This comparison assesses whether a faithful surrogate can improve low-bit quantization beyond STE and existing alternatives.

Results on LLaMA-3.2-1B. As shown in Table 1, RFS consistently achieves strong *Best-of-N* performance while substantially improving training stability across runs. At 4-bit, RFS achieves the best result of 61.24 and a mean of 60.64 ± 0.35, outperforming both ParetoQ and DSQ. The advantage becomes more pronounced at lower precision. At 3-bit, RFS achieves 59.93 ± 0.14, compared with 58.80 ± 0.63 for ParetoQ and 52.14 ± 10.71 for DSQ, while also attaining the highest Best-of-5 result of 60.12. The large variance of DSQ is due to the occurrence of training collapse. At 2-bit,

where optimization becomes substantially more challenging, RFS remains considerably more stable than DSQ (54.73 ± 5.48 vs. 44.86 ± 13.67). These results show that RFS not only improves peak performance, but reduces variance and mitigates training collapse under aggressive quantization.

5.2 QAT ON GENERATIVE RECOMMENDATION

We adopt Diffusion-GR2 (Li et al., 2026; Zhang et al., 2026), a representative generative recommendation framework for the reranking tasks, and perform QAT under both W8A8 MXFP8 and W4A4 MXFP4 quantization. The backbone is a Qwen3-8B diffusion language model (Li et al., 2025), with training and evaluation conducted on the Amazon Beauty dataset (McAuley et al., 2015). As shown in Table 2, RFS consistently outperforms STE across both MXFP8 and MXFP4 settings, yielding clear improvements in ranking quality under different quantization precisions. These results demonstrate that the improved surrogate gradient of RFS transfers effectively to large-scale generative recommendation and remains robust under aggressive low-precision quantization. These results highlight the practical potential of RFS for efficient generative recommendation system.

Table 2: Diffusion-GR2's re-ranking best-of-3 accuracy on Amazon Beauty with Qwen3-8B-DLM as backbone under MXFP8 and MXFP4.

Precision	Surrogate	R@1	R@3	NDCG@3
BF16	–	0.2768	0.5548	0.4358
W8A8 MXFP8	STE	0.2760	0.5531	0.4343
	Weight-RFS	0.2764	0.5544	0.4351
	Weight+Act-RFS	**0.2769**	**0.5550**	**0.4362**
W4A4 MXFP4	STE	0.2731	0.5520	0.4339
	Weight-RFS	0.2756	0.5542	0.4350
	Weight+Act-RFS	**0.2764**	**0.5547**	**0.4352**

5.3 QAT ON POST-TRAINING REASONING

We further evaluate RFS on post-training reasoning using ReQAT (Lee et al., 2026), a recent QAT framework designed for low-precision reasoning models. Following its full-quantization setting, we adopt NVFP4 W4A4KV4 quantization on DeepSeek-R1-14B (Guo et al., 2025) and replace STE with RFS, while keeping the remaining ReQAT training recipe unchanged. We evaluate mathematical reasoning accuracy on AIME-2025 (Zhang & Math-AI, 2025), AIME-90, and MATH-500 (Lightman et al., 2024), and report the mean over three training seeds. As shown in Table 3, RFS consistently improves upon the STE baseline across all three benchmarks, with gains from **1.80%** to **10.00%**, demonstrating its effectiveness for aggressive NVFP4 quantization in reasoning-intensive post-training.

Table 3: Full NVFP4 W4A4KV4 upon ReQAT (Lee et al., 2026) on DeepSeek-R1-14B. Best-of-3 Accuracy (%).

Dataset	Surrogate	Correct	Acc. (%)	Δ
AIME-2025	STE	17 / 30	56.67	–
	Weight-RFS	19 / 30	63.33	**6.66**↑
	Weight+Act-RFS	20 / 30	**66.67**	**10.00**↑
AIME-90	STE	54 / 90	60.00	–
	Weight-RFS	56 / 90	62.22	**2.22**↑
	Weight+Act-RFS	57 / 90	**63.33**	**3.33**↑
MATH-500	STE	470 / 500	94.00	–
	Weight-RFS	479 / 500	95.80	**1.80**↑
	Weight+Act-RFS	481 / 500	**96.20**	**2.20**↑

5.4 PTQ ON IMAGE CLASSIFICATION

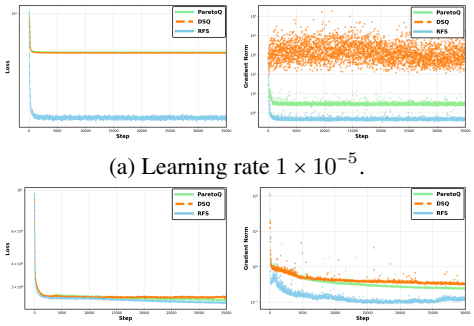
Finally, we evaluate RFS in the PTQ setting for image classification on ImageNet-1K (Deng et al., 2009). We adopt NWQ (Zheng et al., 2022) as the PTQ baseline and consider MobileNetV2 and ResNet18 under W4A4 and W3A3 quantization. Since activation quantization involves shared quantization parameters aggregated over heterogeneous activations, we apply the normalized variant of RFS introduced in section 3. We report the best result over three independent runs. As shown in Table 4, normalized RFS consistently improves over the STE baseline across architectures and bit-widths. Under the aggressive W3A3 setting, RFS improves MobileNetV2 by **1.78%** while also yielding consistent improvements on ResNet18. These results demonstrate that RFS extends effectively beyond QAT and remains beneficial for PTQ with low-bit quantization.

Table 4: Integer PTQ on ImageNet-1K with CNNs. Best-of-3 Top-1 Acc (%).

Method	W/A	ResNet-18	MobileNetV2
Full Precision	32/32	71.06	72.49
ACIQ-Mix (Banner et al., 2018)	4/4	67.00	–
ZeroQ (Cai et al., 2020)	4/4	21.71	26.24
LAPQ (Nahshan et al., 2021)	4/4	60.30	49.70
AdaQuant (Hubara et al., 2020)	4/4	69.60	47.16
Bit-Split (Wang et al., 2020)	4/4	67.56	–
AdaRound (Nagel et al., 2020)	4/4	67.96	61.52
QDROP (Wei et al., 2022)	4/4	69.62	68.84
NICE (Baskin et al., 2021)	4/4	69.79	–
NWQ (Zheng et al., 2022)	4/4	69.85	69.14
Act-RFS (upon NWQ)	4/4	**69.99**	**69.43**
AdaRound (Nagel et al., 2020)	3/3	64.66	15.20
BRECQ (Li et al., 2021)	3/3	65.87	23.41
QDROP (Wei et al., 2022)	3/3	66.75	57.98
NICE (Baskin et al., 2021)	3/3	67.68	–
NWQ (Zheng et al., 2022)	3/3	67.58	61.24
Act-RFS (upon NWQ)	3/3	**68.38**	**63.02**

5.5 Training Stability Validation and Ablations

Gradient Dynamics and Convergence Behavior. Figure 5 empirically supports our theoretical analysis in Section 4. Across learning rates, RFS exhibits smooth and well-behaved optimization, with steadily decreasing training loss and stable gradient norms throughout training. Consistent with Theorem C.1, its smaller approximation error to the rotated rounding function provides more faithful optimization directions, leading to improved convergence and lower training loss. Moreover, RFS maintains substantial gradient signals while avoiding extreme gradient outliers. In contrast, DSQ exhibits sharp gradient spikes and less stable optimization. This observation aligns with Theorem 4.2: the bounded gradient variance of RFS prevents gradient explosion or premature vanishing and enables gradual gradient-norm decay.

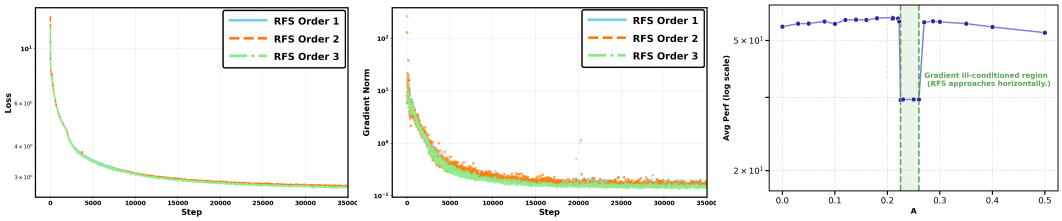


(a) Learning rate 1×10^{-5}.

(b) Learning rate 2×10^{-4}.

Figure 5: Training loss and gradient norm dynamics for LLaMA-3.2-1B under different learning rates.

Normalization for Activation RFS. We then isolate the effect of the normalization in the activation RFS Equation 9 by disabling it, using NWQ as the experimental setting. Without normalization, we observe a regression, as an outlying activation candidate can dominate and distort the overall signal. In particular, MobileNetV2 suffers about 3.5% degradation on W3A3 relative to the NWQ baseline. Once normalization is enabled, RFS resolves this aggregation issue and surpasses the NWQ baseline.



(a) Training Loss (b) Gradient Norm (c) Amplitude Sensitivity

Figure 6: Training dynamics and amplitude ablation of RFS.

Fourier Order of RFS. We next study the effect of the Fourier truncation order M in RFS. Figure 6a and Figure 6b compare training dynamics across $M = 0, 1, 2$. The first-order approximation ($M = 0$) already captures most of the performance and stability gains, yielding smooth loss decay with low, well-controlled gradient norms. Higher-order terms introduce finer structure of the quantization operator, but their empirical benefits are marginal and come with numerical complexity. This validates our choice of the first-order RFS in Equation 7, which balances approximation fidelity, optimization stability, and computational simplicity for quantization.

Amplitude. Finally, we vary the amplitude A to control the sharpness of surrogate gradient, to empirically validate the damped amplitude design predicted by the theoretical analysis in Section 3.2. As shown in Figure 6c, the empirical results closely align with the theoretical characterization and clearly reveal the presence of an ill-conditioned regime, where the surrogate curves become nearly tangential to horizontal plateaus. In contrast, moderate amplitudes provide a favorable balance between approximation fidelity and optimization stability. The selected default setting achieves strong performance, benefiting from an adequate approximation of rounding operator while maintaining stable and well-conditioned training dynamics.

6 Conclusion

We introduce RFS, a general surrogate-gradient framework that goes beyond STE by capturing the discrete structure of quantization while maintaining stable, well-conditioned optimization. Across diverse quantization formats and applications, RFS consistently improves upon STE in QAT and PTQ. These results demonstrate that RFS provides a simple, general, and effective approach to quantization.

AI USE STATEMENT

We used AI tools only to assist manuscript editing and baseline implementation. All AI-assisted written content has been reviewed, revised and gated by the authors. Moreover, all research ideas, algorithmic design, RFS's core code-base, experimental analysis, theoretical claims and analysis, and final claims are authors' own handcrafted efforts without AI-assistance.

ETHICS STATEMENT

Quantization does not itself address biases, privacy concerns, or harmful capabilities of the underlying models and training data. Improved quantization stability and quality should not be interpreted as guarantees of safety or fairness. Consequently, models quantized using our method should undergo application-specific evaluation before deployment.

REPRODUCIBILITY STATEMENT

RFS is a plug-and-play module that can be integrated into existing quantization benchmarks and implementations. We provide the full RFS formulation together with its well-conditioned operating region, default amplitude, and sensitivity analysis to facilitate reproduction.

REFERENCES

Myeonghwan Ahn and Sungjoo Yoo. Differentiable, stable and efficient floating-point quantization. *OpenReview (ICLR 2026 submission; withdrawn)*, 2025.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Runsheng Bai, Bo Liu, and Qiang Liu. Skim: Any-bit quantization pushing the limits of post-training quantization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Aciq: Analytical clipping for integer quantization of neural networks. 2018.

Chaim Baskin, Evgenii Zheltonozhkii, Tal Rozen, Natan Liss, Yoav Chai, Eli Schwartz, Raja Giryes, Alexander M Bronstein, and Avi Mendelson. Nice: Noise injection and clamping estimation for neural network quantization. *Mathematics*, 9(17):2144, 2021.

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

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

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 696–697, 2020.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Qbb: Quantization with binary bases for llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13166–13175. IEEE, 2020.

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

Alexandre Défossez, Yossi Adi, and Gabriel Synnaeve. Differentiable model compression via pseudo quantization noise. *arXiv preprint arXiv:2104.09987*, 2021.

Alexandre Défossez, Yossi Adi, and Gabriel Synnaeve. Differentiable model compression via pseudo quantization noise. *Transactions on Machine Learning Research (TMLR)*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Peiyan Dong, Lei Lu, Chao Wu, Cheng Lyu, Geng Yuan, Hao Tang, and Yanzhi Wang. Packqvit: Faster sub-8-bit vision transformers via full and packed quantization on the mobile. *Advances in Neural Information Processing Systems*, 36:9015–9028, 2023.

Dayou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.

Dayou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4852–4861, 2019a.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019b.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2007.

Tianteng Gu, Bei Liu, Haoyu Wang, and Yanmin Qian. Ultra-low bit post-training quantization of large speech models via k-means clustering and mixed precision allocation. In *Proc. Interspeech 2025*, pp. 1988–1992, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Xianghong Hu, Jeffry Wicaksana, Eric Xing, and Kwang-Ting Cheng. Sdq: Stochastic differentiable quantization with mixed precision. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Xijie Huang, Zhiqiang Shen, Pingcheng Dong, and Kwang-Ting Cheng. Quantization variation: A new perspective on training transformers with low-bit precision. *arXiv preprint arXiv:2307.00331*, 2023.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research (JMLR)*, 18(187):1–30, 2018.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*, 2020.

Minsoo Kim, Kyuhong Shim, Seongmin Park, Wonyong Sung, and Jungwook Choi. Teacher intervention: Improving convergence of quantization aware training for ultra-low precision transformers. *arXiv preprint arXiv:2302.11812*, 2023.

Arnav Kundu, Chungkuk Yoo, Minsik Cho, and Saurabh Adya. Differentiable soft min-max loss to restrict weight range for model quantization. 2nd Differentiable Almost Everything Workshop at ICML 2024, 2024.

Janghwan Lee, Sihwa Lee, Jinseok Kim, Yongjik Kim, Jieun Lim, Jinwook Oh, and Jungwook Choi. Reqat: Achieving full-precision reasoning accuracy with 4-bit floating-point quantization-aware training. *arXiv preprint arXiv:2606.15682*, 2026.

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

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*, 2025.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yufei Li, Zaiwei Zhang, Mingfu Liang, Kavosh Asadi, Jay Xu, Jimmy Kim, Chongyang Bai, Jieyi Zhang, Hongye Xie, Prachi Agrawal, et al. Gr2 technical report. *arXiv preprint arXiv:2606.31984*, 2026.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Zhexin Li, Tong Yang, Peisong Wang, and Jian Cheng. Q-vit: Fully differentiable quantization for vision transformer. *arXiv preprint arXiv:2201.07703*, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

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*, 2024a.

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*, 2024b.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024c.

Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, Lin Xiao, Yuandong Tian, Bilge Soran, Raghuraman Krishnamoorthi, Tijmen Blankevoort, and Vikas Chandra. Paretoq: Improving scaling laws in extremely low-bit llm quantization. *arXiv*, 2025.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110(11):3245–3262, 2021.

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

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate lora-finetuning quantization of llms via information retention. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Xiaoyi Qu, David Aponte, Colby Banbury, Daniel P Robinson, Tianyu Ding, Kazuhito Koishida, Ilya Zharkov, and Tianyi Chen. Automatic joint structured pruning and quantization for efficient neural network training and compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15234–15244, 2025.

Sergey Salishev and Ian Akhremchik. Gdnsq: Gradual differentiable noise scale quantization for low-bit neural networks. *arXiv*, 2025.

Stanislav Semenov. Smooth approximations of the rounding function. *arXiv preprint arXiv:2504.19026*, 2025.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R. Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627b token cleaned and deduplicated version of RedPajama, 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B. Cerebras Systems.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Timothy Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama 2: Open foundation and fine-tuned chat models, 2023.

Boris van Breugel, Yelysei Bondarenko, Paul Whatmough, and Markus Nagel. Fptquant: Function-preserving transforms for llm quantization. *arXiv*, 2025.

Georgios Vlassis, Saleh Ashkboos, Alexandra Volkova, Torsten Hoefler, and Dan Alistarh. Beyond outliers: A study of optimizers under quantization. *arXiv*, 2025.

Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In *International conference on machine learning*, pp. 9847–9856. PMLR, 2020.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

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 (ICML)*, 2023.

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.

Chong Yu, Tao Chen, Zhongxue Gan, and Jiayuan Fan. Boost vision transformer with gpu-friendly sparsity and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22658–22668, 2023.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime). *Mathematical Association of America (MAA)*, 2025.

Zhuoxuan Zhang, Kangqi Ni, Yuhang Chen, Mingfu Liang, Xiaohan Wei, Yunchen Pu, Fei Tian, Chonglin Sun, Frank Shyu, Sandeep Pandey, et al. Diffusion-gr2: Diffusion generative reasoning re-ranker. *arXiv preprint arXiv:2607.01170*, 2026.

DanDan Zheng, Yuanliu Liu, Liang Li, et al. Leveraging inter-layer dependency for post-training quantization. *Advances in Neural Information Processing Systems*, 35:6666–6679, 2022.

A More Related Works: Complementary but Orthogonal to RFS

Recent works on post-training quantization (PTQ) focus on mitigating outliers and distribution skewness, which become increasingly problematic at lower bitwidths. Techniques such as function-preserving transformations and channel-wise clustering aim to reshape weight distributions so that uniform or centroid-based quantizers incur less error without modifying model parameters (van Breugel et al., 2025; Bai et al., 2025; Gu et al., 2025). QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2024c) assume that quantization error can be mitigated via orthogonal rotations that redistribute variance across dimensions. However, at ultra-low bitwidths, the dominant source of error shifts from distributional anisotropy to discretization-induced geometry collapse. The quantized space forms an extremely coarse discrete lattice in which many distinct directions become indistinguishable. Orthogonal rotations can redistribute variance but cannot increase the intrinsic angular resolution of the quantized space, and therefore cannot recover representational geometry lost through discretization. AWQ (Lin et al., 2024) allocates quantization precision according to activation saliency by rescaling weight channels. This strategy better preserves semantically important subspaces and thus improves performance at moderate bitwidths. However, AWQ remains fundamentally limited by the discrete resolution of the quantized space: while it redistributes precision across channels, it cannot increase the number of representable states. Consequently, at ultra-low bitwidths, discretization-induced geometry collapse dominates, and channel-wise rescaling becomes insufficient to preserve fine-grained angular relationships in attention and representation spaces.

Beyond reshaping weight distributions, PTQ methods can directly optimize discretization decisions, for example by learning task-aware weight rounding assignments or by exploiting inter-layer dependencies to reduce accumulated quantization error across the network (Frantar et al., 2023). Although such approaches improve the fidelity of quantized pretrained models, they remain fundamentally constrained by the fixed full-precision solution from which quantization begins.

While most PTQ methods aim to approximate full precision model as accurately as possible, QAT instead tries to learn a new model under low-bit constraints. However, as bitwidth decreases, STE-based training often becomes unstable or biased. Several works focus on improving training efficiency and scalability of QAT rather than altering the quantization operator itself; examples include block-wise or staged training schemes that reduce computational cost, as well as data-free distillation strategies that remove dependence on large calibration datasets (Liu et al., 2024b; Chen et al., 2025). Another thread leverages self-distillation objectives to stabilize sub-4-bit settings (Du et al., 2024a).

A complementary line of work reformulates quantization as an approximation or decomposition problem. For instance, QBB represents weights as linear combinations of binary bases, replacing multiplications with additions while optimizing scales via distillation (Bulat et al., 2024; Kim et al., 2023). LoRA-aware approaches such as LoftQ and IR-QLoRA integrate quantization with low-rank adaptation, either by alternating between quantization and decomposition or by explicitly preserving information through auxiliary connections (Li et al., 2024; Qin et al., 2024). Another body of work treats quantization through stochastic or noise-based perspectives, aiming to avoid hard discrete operators during training such as DiffQ which introduces pseudo quantization noise to approximate quantization effects (Défossez et al., 2022). Similarly, stochastic differentiable quantizations have bitwidths sampled probabilistically and optimized via reparameterization (Huang et al., 2022).

Differentiable and soft quantization methods more directly target the rounding operation by introducing continuous relaxations. DSQ is a representative example, utilizing parameterized smooth function whose sharpness increases during training (Gong et al., 2019b). DSMM (Kundu et al., 2024) introduces a differentiable soft min-max loss to reduce weight ranges by making max/min approximations differentiable and learnable.

Other work treats quantization through stochastic or noise-based perspectives. NICE (Baskin et al., 2021) combines quantization-noise injection with learned clamping to improve low-bit training, while DiffQ replaces quantization during training with independent pseudo-quantization noise, providing an STE-free differentiable surrogate that additionally enables end-to-end optimization of bitwidths (Défossez et al., 2021). While these approaches mitigate the biased gradients associated with hard rounding, their continuous noise surrogates approximate the effect of quantization rather than optimizing directly over the discrete quantized representation.

Recent work pushes differentiable quantization further by explicitly optimizing bitwidths or noise scales as continuous variables. GDNSQ (Salishev & Akhremchik, 2025) frames quantization as

Table 5: Multi-seed results on LLaMA-3.2-3B. RFS achieves strong Best-of-5 performance while substantially improving training stability, particularly at low bit-widths.

W/A	Method	Setting	Arc-e	Arc-c	Boolq	Hellaswag	Openbookqa	Piqa	SciQ	Winogrande	Best Avg of N	Δ	Mean ± Std
16/16	Baseline	Baseline	71.63	45.99	73.39	73.61	43.00	77.48	92.70	69.85	68.46	–	–
4/16	ParetoQ (Liu et al., 2025)	20B Tokens & lr=1e-4	71.83	45.48	70.13	71.20	42.40	76.58	90.60	66.25	66.81	–	66.45 ± 0.42
4/16	DSQ (Gong et al., 2019a)	20B Tokens & lr=1e-4	70.19	41.73	68.39	64.58	39.40	76.51	90.93	64.40	64.48	–	63.38 ± 0.69
4/16	**RFS** (upon ParetoQ)	20B Tokens & lr=1e-4	72.05	44.97	70.21	71.28	42.00	77.58	91.50	67.64	**67.15**	**+2.67**	**67.03 ± 0.13**
3/16	ParetoQ (Liu et al., 2025)	20B Tokens & lr=1e-4	71.47	45.16	70.28	70.00	42.00	76.15	91.00	65.93	66.50	–	65.91 ± 0.56
3/16	DSQ (Gong et al., 2019a)	20B Tokens & lr=1e-4	66.58	41.64	70.24	66.58	40.60	76.28	89.80	64.48	64.52	–	57.29 ± 13.61
3/16	**RFS** (upon ParetoQ)	20B Tokens & lr=1e-4	71.04	44.97	69.02	70.81	44.00	78.29	90.40	66.69	**66.90**	**+2.38**	**66.34 ± 0.44**
2/16	ParetoQ (Liu et al., 2025)	30B Tokens & lr=1e-4	65.73	38.08	65.73	64.13	39.40	74.68	84.30	61.43	61.69	–	56.09 ± 10.70
2/16	DSQ (Gong et al., 2019a)	30B Tokens & lr=1e-4	69.19	40.78	65.54	66.03	41.60	74.92	88.60	63.61	**63.78**	–	50.91 ± 14.59
2/16	**RFS** (upon ParetoQ)	30B Tokens & lr=1e-4	68.83	40.98	65.81	65.58	41.00	75.13	88.20	63.10	63.57	-0.21	**59.39 ± 8.06**

a smooth constrained optimization over effective bitwidths, combining STE-style estimators with gradual noise scheduling and distillation losses. Floating-point oriented frameworks (Ahn & Yoo, 2025) extend pseudo-quantization training to mixed floating-point formats, enabling differentiable optimization of mantissa and exponent precisions. Other works provide scaling laws and diagnostics to assess failure modes at low-bit regimes (Liu et al., 2025; Vlassis et al., 2025). However, a key challenge in low-bit QAT that remains lies in reconciling with stable, informative gradients.

B Additional Experimental Results

B.1 QAT on Continuous-Pretraining for LLaMA-3.2-3B

Results on LLaMA-3.2-3B. Similar trends are observed on the larger LLaMA-3.2-3B model (Table 5). At **4-bit** and **3-bit**, RFS achieves the best Best-of-N performance while also attaining the highest multi-seed mean with low variance, reaching 67.03 ± 0.13 and 66.34 ± 0.44, respectively. The stability advantage becomes more pronounced at 2-bit: ParetoQ and DSQ exhibit large run-to-run variance (10.70 and 14.59), indicating frequent training collapse, whereas RFS achieves a substantially higher mean of 59.39 with a lower standard deviation of 8.06. Notably, although DSQ obtains a slightly higher 2-bit Best-of-N result (63.78 vs. 63.57), its mean drops to only 50.91, highlighting its sensitivity to individual runs. Overall, RFS provides both strong peak performance and substantially more robust optimization, with its stability advantage becoming increasingly evident under aggressive quantization.

B.2 Additional Results on Quantization-Aware Distillation

We further evaluate RFS on quantization-aware distillation (QAD) using BitDistiller (Du et al., 2024b) with LLaMA-2-7B (Touvron et al., 2023). BitDistiller combines AWQ initialization (Lin et al., 2024) with confidence-aware self-distillation for low-bit LLM quantization. We directly replace STE with RFS for gradient estimation through rounding.

We follow the default BitDistiller recipe with an effective batch size of 128, learning rate 8×10^{-6}, and sequence length 1024 on Wikipedia and Alpaca data. As shown in Table 6, RFS consistently improves both perplexity and training stability over STE, reducing PPL from 7.561 to 7.545 with 38M tokens and from 7.406 to 7.363 with 150M tokens. This demonstrates that RFS remains effective when combined with quantization-aware distillation.

Table 6: WikiText-2 perplexity of BitDistiller on LLaMA-2-7B under W2A16 INT QAD.

Bits	Ep.	Groups-of	Data	Grad.	Tok.	PPL↓
2	8	128	8K	STE	38M	7.561±.021
				RFS	38M	**7.545±.007**
2	32	128	8K	STE	150M	7.406±.022
				RFS	150M	**7.363±.006**

B.3 Additional Experimental Results on Vision Transformer

To demonstrate the generality of RFS beyond LLMs, we further evaluate it on representative Vision Transformer architectures. Following prior QAT studies (Huang et al., 2023; Qu et al., 2025), we consider DeiT (Touvron et al., 2021) and Swin (Liu et al., 2021). We compare RFS against several state-of-the-art ViT quantization methods, including Q-ViT (Li et al., 2022), LSQ+ (Bhalgat et al., 2020), PackQViT (Dong et al., 2023) and GPUSQ-ViT (Yu et al., 2023), on the ImageNet-1k benchmark (Deng et al., 2009). The results are present in Table 7. We adopted RFS upon the code-base of (Huang et al., 2023) while exclude knowledge distillation for fair comparison. Numerical results

present that RFS consistently brings performance again across varying bits and ViTs, demonstrating its generality and transferability across different applications and architectures.

Table 7: VIT on ImageNet-1K. "Bit-width (W/A)" denotes the bitwidth for weights and activations.

Network	QAT Method	Bit-width (W/A)	Top-1	Bit-width (W/A)	Top-1	Bit-width (W/A)	Top-1
DeiT-T (FP Top-1: 73.8)	Q-ViT	4/4†	72.79	3/3†	69.62	–	–
	GPUSQ-ViT	4/4	71.70	–	–	–	–
	PackQViT	4/4	72.70	–	–	–	–
	LSQ+	4/4	72.62	3/3	68.22	2/2	54.45
	RFS	4/4	**73.08**	3/3	**69.80**	2/2	**56.31**
Swin-T (FP Top-1: 81.0)	Q-ViT	4/4†	80.59	3/3†	79.45	–	–
	GPUSQ-ViT	4/4	80.70	–	–	–	–
	PackQViT	4/4	81.50	–	–	–	–
	Li et al.	4/4	**82.10**	3/3	80.57	2/2	74.31
	LSQ+	4/4	80.61	3/3	79.07	2/2	70.21
	RFS	4/4	81.98	3/3	**81.02**	2/2	**75.86**

† Average bitwidth for mixed-precision quantization.

B.4 Gradient Variance and Optimization Quality

To further examine the optimization behavior of RFS, we compare its gradient variance with DSQ at their actual operating points during W4 QAT of LLaMA-3.2-1B on FineWeb-Edu. We use the default RFS amplitude for RFS and the scheduled sharpness for DSQ, and track both gradient variance and WikiText-2 perplexity throughout training.

As shown in Table 8, RFS consistently exhibits **1.95–4.60× lower gradient variance** than DSQ throughout training. This reduction is accompanied by substantially better optimization quality: RFS steadily reduces perplexity from 54.4 to 29.2, whereas DSQ exhibits a non-monotonic trajectory and converges to a much higher perplexity of 49.4. At the final checkpoint, RFS maintains a **2.36× lower gradient variance** while reducing perplexity by 20.2 points. These results empirically support the connection between the lower-variance surrogate gradients of RFS and improved optimization quality.

Table 8: Gradient variance and WikiText-2 perplexity during W4A16 QAT.

Step	**RFS PPL ↓**	**RFS Grad-Var ↓**	DSQ PPL ↓	DSQ Grad-Var ↓	Ratio
500	**54.4**	1.23e-9	61.3	5.66e-9	4.60×
1,000	**49.5**	0.99e-9	61.4	1.93e-9	1.95×
2,000	**43.6**	0.86e-9	73.7	1.98e-9	2.30×
3,000	**40.1**	0.82e-9	62.7	2.00e-9	2.44×
4,000	**36.9**	0.78e-9	59.2	1.96e-9	2.51×
5,000	**35.0**	0.74e-9	57.4	1.94e-9	2.62×
10,000	**31.4**	0.71e-9	50.0	1.91e-9	2.69×
15,000	**30.2**	0.76e-9	49.9	1.91e-9	2.51×
20,000	**29.2**	0.86e-9	49.4	2.03e-9	2.36×

C Additional Theorem and Proofs

C.1 STE is a special case of RFS

We first compare STE and RFS using surrogate approximations of the rotated rounding function equation 3. Due to periodicity, we focus on a single period $[0, T]$ without loss of generality. As the rotated function is square-integrable, we conduct our analysis in the space $L^2([0, T])$.

Theorem C.1. *Let $f \in L^2([0,T])$ and let f_n be the nth partial Fourier sum of f. Then, for all $n \in \mathbb{N}$,*
(i) f_n is the unique minimizer of the L^2 approximation error among trigonometric polynomials of degree at most n;
(ii) For $n \geq 1$, $\|f - f_n\|_2 < \|f - f_0\|_2$ if and only if f is non-constant almost everywhere.

Proof. See Appendix C.2. □

Remark I. Theorem C.1(i) implies that the nth partial Fourier sum is the optimal L^2 surrogate of the rotated rounding function among trigonometric polynomials of degree at most n. For STE, the surrogate is restricted to constant functions. In contrast, RFS admits surrogates from trigonometric polynomials of degree at most $n \in \mathbb{N}$, making STE a special case of RFS.

Remark II. Theorem C.1(ii) implies that, unless the function is constant almost everywhere, its surrogate drawn from trigonometric polynomials of degree of at least 1 achieves a strictly smaller L^2 approximation error than constant function surrogate. Consequently, RFS admits surrogate functions that are provably closer to the rotated rounding function than STE in the L^2 sense.

C.2 Proof of Theorem C.1

For clarity in the proof, we restate Theorem C.1 using precise mathematical notation.

Theorem C.2. *Let $f \in L^2([0,T])$ and let f_n be the nth partial Fourier sum of the function f. Define the function space of trigonometric polynomials of degree at most n by*

$$\mathcal{G}_n = \operatorname{span}\left\{1, \cos\left(\frac{2\pi ku}{T}\right), \sin\left(\frac{2\pi ku}{T}\right) : 1 \le k \le n\right\}$$

(i) f_n is the unique minimizer of the L^2 approximation error over the space $\mathcal{G}_n$, i.e.,

$$f_n = _{g\in\mathcal{G}_n} \|f - g\|_{L^2}.$$

(ii) For all $n \in \mathbb{N}\backslash\{0\}$, the strict inequality

$$\|f - f_n\|_{L^2} < \|f - f_0\|_{L^2}$$

holds if and only if f is not equal to constant almost everywhere on $[0,T]$.

Proof. We first prove part (i). The Fourier coefficients of f on the interval $[0,T]$ are defined by

$$\begin{aligned}
\alpha_0 &= \frac{1}{T}\int_0^T f(u)\,du,\\
\alpha_k &= \frac{2}{T}\int_0^T f(u)\cos\left(\frac{2\pi ku}{T}\right)du, \text{for } k \ge 1,\\
\beta_k &= \frac{2}{T}\int_0^T f(u)\sin\left(\frac{2\pi ku}{T}\right)du, \text{for } k \ge 1.
\end{aligned} \tag{11}$$

Note that we have, for $j,k \ge 1$,

$$\begin{gathered}
\int_0^T \cos\left(\frac{2\pi ju}{T}\right)\cos\left(\frac{2\pi ku}{T}\right)du = \frac{T}{2}\delta_{jk}, \quad \int_0^T \sin\left(\frac{2\pi ju}{T}\right)\sin\left(\frac{2\pi ku}{T}\right)du = \frac{T}{2}\delta_{jk},\\
\int_0^T \cos\left(\frac{2\pi ju}{T}\right)\sin\left(\frac{2\pi ku}{T}\right)du = 0,
\end{gathered} \tag{12}$$

where δ_{jk} is the Kronecker delta and it is equal to 1 if $j = k$ and 0 otherwise.

Now let $g \in \mathcal{G}_n$ be an arbitrary trigonometric polynomial of the form

$$g(u) = a_0 + \sum_{k=1}^{n}\left[a_k\cos\left(\frac{2\pi ku}{T}\right) + b_k\sin\left(\frac{2\pi ku}{T}\right)\right].$$

Expanding the squared L^2 norm error $\|f-g\|_{L^2}^2$, we have that

$$\|f - g\|_{L^2}^2 = \|f\|_{L^2}^2 - 2\langle f, g\rangle_{L^2} + \|g\|_{L^2}^2. \tag{13}$$

For the inner product $\langle f, g\rangle_{L^2}$ in equation 13, we have from equation 11 that

$$\begin{aligned}
\langle f, g\rangle_{L^2} &= \int_0^T f(u)\left[a_0 + \sum_{k=1}^{n}\left(a_k\cos\left(\frac{2\pi ku}{T}\right) + b_k\sin\left(\frac{2\pi ku}{T}\right)\right)\right]du\\
&= a_0\int_0^T f(u)\,du + \sum_{k=1}^{n}\left[a_k\int_0^T f(u)\cos\left(\frac{2\pi ku}{T}\right)du + b_k\int_0^T f(u)\sin\left(\frac{2\pi ku}{T}\right)du\right]\\
&= T\alpha_0 a_0 + \frac{T}{2}\sum_{k=1}^{n}(\alpha_k a_k + \beta_k b_k).
\end{aligned} \tag{14}$$

For the squared norm $\|g\|_{L^2}^2$ in equation 13, we have from equation 12 that

$$\begin{aligned}\|g\|_{L^2}^2 &= \int_0^T \left[a_0 + \sum_{k=1}^{n}\left(a_k \cos\left(\frac{2\pi k u}{T}\right) + b_k \sin\left(\frac{2\pi k u}{T}\right)\right)\right]^2 du. \\ &= a_0^2 \cdot T + \sum_{k=1}^{n}\left[a_k^2 \cdot \frac{T}{2} + b_k^2 \cdot \frac{T}{2}\right] = T a_0^2 + \frac{T}{2}\sum_{k=1}^{n}(a_k^2 + b_k^2)\end{aligned} \tag{15}$$

It then follows from equation 13, equation 14 and equation 15 that

$$\begin{aligned}\|f-g\|_{L^2}^2 &= \|f\|_{L^2}^2 - 2\left[T\alpha_0 a_0 + \frac{T}{2}\sum_{k=1}^{n}(\alpha_k a_k + \beta_k b_k)\right] + T a_0^2 + \frac{T}{2}\sum_{k=1}^{n}(a_k^2 + b_k^2) \\ &= \|f\|_{L^2}^2 + T(a_0^2 - 2\alpha_0 a_0) + \frac{T}{2}\sum_{k=1}^{n}\left[(a_k^2 - 2\alpha_k a_k) + (b_k^2 - 2\beta_k b_k)\right] \\ &= \|f\|_{L^2}^2 - T\alpha_0^2 - \frac{T}{2}\sum_{k=1}^{n}(\alpha_k^2 + \beta_k^2) + T(\alpha_0 - a_0)^2 + \frac{T}{2}\sum_{k=1}^{n}\left[(\alpha_k - a_k)^2 + (\beta_k - b_k)^2\right].\end{aligned} \tag{16}$$

Recall that the nth partial Fourier sum is

$$f_n(u) = \alpha_0 + \sum_{k=1}^{n}\left[\alpha_k \cos\left(\frac{2\pi k u}{T}\right) + \beta_k \sin\left(\frac{2\pi k u}{T}\right)\right]. \tag{17}$$

Replacing g in $\|f-g\|_{L^2}^2$ with f_n given in equation 17, we have that

$$\|f-f_n\|_{L^2}^2 = \|f\|_{L^2}^2 - T\alpha_0^2 - \frac{T}{2}\sum_{k=1}^{n}(\alpha_k^2 + \beta_k^2). \tag{18}$$

Substituting equation 18 back into equation 16, we obtain

$$\|f-g\|_{L^2}^2 = \|f-f_n\|_{L^2}^2 + T(\alpha_0 - a_0)^2 + \frac{T}{2}\sum_{k=1}^{n}\left[(\alpha_k - a_k)^2 + (\beta_k - b_k)^2\right]. \tag{19}$$

Since all terms on the right-hand side of equation 19 beyond $\|f-f_n\|_{L^2}^2$ are non-negative, it follows that $\|f-g\|_{L^2}^2 \geq \|f-f_n\|_{L^2}^2$ for all $g \in \mathcal{G}_n$. Moreover, equality holds if and only if $\alpha_0 = a_0$, $\alpha_k = a_k$ and $\beta_k = b_k$ for all $k \in \{1, \cdots, n\}$. This establishes that f_n is the unique minimizer of the L^2 approximation error over $\mathcal{G}_n$, which completes the first part of the proof.

We next prove part (ii). In the first part of the proof, we have shown that

$$\|f-f_n\|_{L^2}^2 = \|f\|_{L^2}^2 - T\alpha_0^2 - \frac{T}{2}\sum_{k=1}^{n}(\alpha_k^2 + \beta_k^2).$$

In particular, for $n = 0$, this gives $\|f-f_0\|_{L^2}^2 = \|f\|_{L^2}^2 - Ta_0^2$. Taking the difference between $\|f-f_0\|_{L^2}^2$ and $\|f-f_n\|_{L^2}^2$, we obtain

$$\|f-f_0\|_{L^2}^2 - \|f-f_N\|_{L^2}^2 = \frac{T}{2}\sum_{k=1}^{N}(\alpha_k^2 + \beta_k^2). \tag{20}$$

Since each term at the right-hand side of equation equation 20 is non-negative, we have $\|f-f_N\|_{L^2} \leq \|f-f_0\|_{L^2}$ for all $N \geq 1$, with strict inequality if and only if at least one of $\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n$ is non-zero.

It then remains to show that at least one of $\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n$ is non-zero is equivalent to the statement that f is not equal to constant almost everywhere on $[0, T]$. We will prove the contrapositive statement. First recall that the constant functions on $[0, T]$ form the one-dimensional subspace

$$\mathrm{span}\{1\} \subset L^2([0, T]).$$

Suppose that f is equal to a constant almost everywhere on $[0, T]$. Since $\cos(2\pi k u/T)$ and $\sin(2\pi k u/T)$ are orthogonal to constant functions in the space $L^2([0, T])$, $\alpha_k = \beta_k = 0$ for all $k \geq 1$. It then follows that $f(u) = \alpha_0$ almost everywhere. Conversely, suppose that for all $k \geq 1$, $\alpha_k = \beta_k = 0$. Then the Fourier series of f reduces to the constant term α_0, and the partial Fourier sums satisfy that for all $n \in \mathbb{N}$,

$$f_n(u) = \alpha_0.$$

Since $f_n \to f \in L^2([0, T])$, it follows that $f = \alpha_0$ almost everywhere on $[0, T]$. Therefore, at least one of $\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n$ being nonzero is equivalent to f not being almost everywhere equal to a constant function on $[0, T]$. This completes the proof of part (ii) and hence the theorem. □

C.3 Proof of Theorem 4.1

Theorem C.3 (RFS gradient alignment with reference gradient)**.** *Suppose that* $\mathbb{E}[L(\cdot; z)]$ *is* L_1*-Lipschitz and* $\mathbb{E}[\nabla L(\cdot; z)]$ *is* L_2*-Lipschitz. Set* $\sigma = \sqrt{L_1\epsilon_M/L_2}$*, where* ϵ_M *is the upper-bound deviation between RFS's* M*-order quantizer and the truth quantizer.*
(i) ***(Error bound)*** $\|\nabla L_{RFS}(w) - \nabla\tilde{L}(w)\|_2 \le 2\sqrt{L_1 L_2 d\epsilon_M}$.
(ii) ***(Cosine similarity)*** $\cos(\nabla L_{RFS}(w), \nabla\tilde{L}(w)) \ge 1 - \dfrac{4\sqrt{L_1 L_2 d\epsilon_M}}{\|\nabla\tilde{L}(w)\|_2}$.

Proof. Define the smoothed surrogate loss function as

$$\tilde{L}_{M,\sigma}(w) \;=\; \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I_d)}\,\mathbb{E}_{v\sim\mathcal{D}}\big[L\big(\tilde{Q}_M(w+u);\, v\big)\big].$$

It then follows from the triangle inequality that

$$\big\|\nabla_w\tilde{L}_M - \nabla_w L_\sigma\big\|_2 \;\le\; \underbrace{\big\|\nabla_w\tilde{L}_M - \nabla_w\tilde{L}_{M,\sigma}\big\|_2}_{\text{Term I}} \;+\; \underbrace{\big\|\nabla_w\tilde{L}_{M,\sigma} - \nabla_w L_\sigma\big\|_2}_{\text{Term II}}. \tag{21}$$

Term I (smoothing bias). Since $\tilde{Q}_M$ is differentiable (it is a smooth trigonometric polynomial, cf. Eq. 9 of the main paper), we may write

$$\nabla_w\tilde{L}_M(w) = \mathbb{E}_v[h_M(w,v)] \qquad \text{and} \qquad \nabla_w\tilde{L}_{M,\sigma}(w) = \mathbb{E}_{u,v}[h_M(w+u,v)],$$

where $h_M(w,v) \coloneqq \nabla_w L\big(\tilde{Q}_M(w);\, v\big)$. By assumption, $\mathbb{E}_v[h_M(\cdot,v)]$ is L_2-Lipschitz. Applying Jensen's inequality gives

$$\big\|\nabla_w\tilde{L}_M - \nabla_w\tilde{L}_{M,\sigma}\big\|_2 \;\le\; \mathbb{E}_u\big[L_2\,\|u\|_2\big] \;=\; L_2\,\sigma\sqrt{d}, \tag{22}$$

where the last equality uses $\mathbb{E}[\|u\|_2] = \sigma\sqrt{d}$ for $u \sim \mathcal{N}(0,\sigma^2 I_d)$.

Term II (RFS approximation error). Define

$$H(w) \coloneqq \mathbb{E}_v\big[L\big(\tilde{Q}_M(w),\, v\big) - L\big(Q(w),\, v\big)\big].$$

Since $\mathbb{E}_z[L(\cdot;\, z)]$ is L_1-Lipschitz and $\|\tilde{Q}_M(w) - Q(w)\|_2 \le \epsilon_M$ uniformly, we have

$$|H(w)| \;\le\; L_1\,\epsilon_M \qquad \text{for all } w.$$

Observe that $\tilde{L}_{M,\sigma}(w) - L_\sigma(w) = \int H(w+u)\,p(u)\,du$ where p is the density of $\mathcal{N}(0,\sigma^2 I_d)$. Applying the change of variables $v = w + u$ and differentiating under the integral moves the derivative onto the Gaussian density:

$$\nabla_w\big(\tilde{L}_{M,\sigma} - L_\sigma\big) \;=\; -\int H(w+u)\,\nabla_u p(u)\,du. \tag{23}$$

For the Gaussian density, $\nabla_u p(u) = -(u/\sigma^2)\,p(u)$, so $\int \|\nabla_u p(u)\|\,du = \mathbb{E}[\|u\|_2]/\sigma^2 = \sqrt{d}/\sigma$. Therefore:

$$\big\|\nabla_w\tilde{L}_{M,\sigma} - \nabla_w L_\sigma\big\|_2 \;\le\; L_1\,\epsilon_M\cdot\frac{\sqrt{d}}{\sigma}. \tag{24}$$

Part 1 (Error bound). Substituting equation 22 and equation 24 into equation 21:

$$\big\|\nabla_w\tilde{L}_M - \nabla_w L_\sigma\big\|_2 \;\le\; L_2\,\sigma\sqrt{d} + L_1\,\epsilon_M\,\frac{\sqrt{d}}{\sigma}.$$

Setting $\sigma = \sigma^* \coloneqq \sqrt{L_1\,\epsilon_M/L_2}$ (which minimizes the right-hand side by setting its derivative with respect to σ to zero) yields

$$\begin{aligned} L_2\,\sigma^*\sqrt{d} + L_1\,\epsilon_M\,\frac{\sqrt{d}}{\sigma^*} &= L_2\sqrt{\frac{L_1\,\epsilon_M}{L_2}}\,\sqrt{d} \;+\; L_1\,\epsilon_M\sqrt{\frac{L_2}{L_1\,\epsilon_M}}\,\sqrt{d} \\ &= \sqrt{L_1\,L_2\,\epsilon_M\,d} + \sqrt{L_1\,L_2\,\epsilon_M\,d} \\ &= 2\sqrt{L_1\,L_2\,d\,\epsilon_M}. \end{aligned}$$

Part 2 (Cosine similarity). Decomposing the inner product gives us

$$\begin{aligned}\langle \nabla_w \tilde{L}_M(w),\, \nabla_w L_\sigma(w)\rangle &= \|\nabla_w L_\sigma(w)\|^2 + \langle \nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w),\, \nabla_w L_\sigma(w)\rangle \\ &\geq \|\nabla_w L_\sigma(w)\|^2 - \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2\, \|\nabla_w L_\sigma(w)\|,\end{aligned} \tag{25}$$

where the inequality follows from the Cauchy–Schwarz inequality. By the triangle inequality,

$$\|\nabla_w \tilde{L}_M(w)\| \leq \|\nabla_w L_\sigma(w)\| + \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2. \tag{26}$$

Combining equation 25 and equation 26, we have that

$$\begin{aligned}\cos(\alpha) &= \frac{\langle \nabla_w \tilde{L}_M(w),\, \nabla_w L_\sigma(w)\rangle}{\|\nabla_w \tilde{L}_M(w)\|\, \|\nabla_w L_\sigma(w)\|} \\ &\geq \frac{\|\nabla_w L_\sigma(w)\|_2^2 - \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2\, \|\nabla_w L_\sigma(w)\|_2}{\big(\|\nabla_w L_\sigma(w)\|_2 + \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2\big)\, \|\nabla_w L_\sigma(w)\|_2} \\ &= \frac{\|\nabla_w L_\sigma(w)\|_2 - \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2}{\|\nabla_w L_\sigma(w)\|_2 + \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2}.\end{aligned}$$

We then have that

$$\begin{aligned}1 - \cos(\alpha) &\leq 1 - \frac{\|\nabla_w L_\sigma(w)\|_2 - \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2}{\|\nabla_w L_\sigma(w)\|_2 + \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2} \\ &= \frac{2\|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2}{\|\nabla_w L_\sigma(w)\|_2 + \|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2} \\ &\leq \frac{2\|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2}{\|\nabla_w L_\sigma(w)\|_2}.\end{aligned}$$

Substituting $\|\nabla_w \tilde{L}_M(w) - \nabla_w L_\sigma(w)\|_2 \leq 2\sqrt{L_1\, L_2\, d\, \epsilon_M}$ from Part 1:

$$\cos\big(\nabla \tilde{L}_M(w),\, \nabla L_\sigma(w)\big) \;\geq\; 1 - \frac{4\sqrt{L_1\, L_2\, d\, \epsilon_M}}{\|\nabla_w L_\sigma(w)\|_2}.$$

□

C.4 Proof of Theorem 4.2

To prove Theorem 4.2, we proceed in three steps. First, we derive the DSQ gradient, as stated in Lemma C.4. Next, we establish the expectation and variance formulas for the DSQ and RFS schemes, respectively, which are given in Lemma C.6. Finally, we complete the proof of Theorem 4.2 by combining these results.

DSQ details. To address the non-differentiability of standard quantization, DSQ introduces a smooth asymptotic function $\phi(\cdot)$ that approximates each step of the uniform staircase quantizer. Specifically, the clipping range $[l, u]$ is uniformly partitioned into $2^b - 1$ intervals with width as $\Delta = \frac{u-l}{2^b-1}$ and the i-th quantization interval defined as

$$P_i = [\, l + i\Delta,\; l + (i+1)\Delta), i \in \{0, \ldots, 2^b - 2\}.$$

For an input $x \in P_i$, the smooth approximation is given by

$$\phi(x) = s \cdot \tanh\big(k(x - m_i)\big), \tag{27}$$

where $m_i = l + (i + \frac{1}{2})\Delta$ denotes the midpoint of i-th quantization interval P_i, and scaling factor $s = \frac{1}{\tanh((k\Delta)/2)}$ ensures continuity of $\phi(\cdot)$ across adjacent intervals. The parameter k controls the sharpness of the approximation, with larger values yielding a closer match to the hard staircase quantizer. To allow learnable control over the quantization sharpness, DSQ introduces a variable $\alpha \in (0, 1)$, which measures the approximation gap between soft and hard quantization and is defined as

$$\alpha = 1 - \tanh\big(\tfrac{1}{2} k\Delta\big). \tag{28}$$

Under this parameterization, the asymptotic function is fully determined by α and Δ with

$$s = \frac{1}{1-\alpha}, \qquad k = \frac{1}{\Delta}\ln\left(\frac{2-\alpha}{\alpha}\right). \tag{29}$$

Combining equation 28 and equation 29 allows us to have the identity

$$\alpha = 1 - \tanh\left(\frac{1}{2}\ln\left(\frac{2-\alpha}{\alpha}\right)\right). \tag{30}$$

Based on the asymptotic function $\phi(\cdot)$, the differentiable soft quantization (DSQ) function is defined as

$$Q_{\mathrm{DSQ}}(x) = \begin{cases} l, & x < l, \\ u, & x > u, \\ l + \Delta\left(i + \dfrac{\phi(x)+1}{2}\right), & x \in P_i. \end{cases} \tag{31}$$

Having established the DSQ formulation, we first derive the gradient of the DSQ function.

Lemma C.4 (DSQ Gradient)**.** *Given the input $x \in P_i$, the gradient of the differentiable soft quantization function $Q_{\mathrm{DSQ}}(x)$ defined in equation 31 is*

$$(x) = \frac{\partial Q_{\mathrm{DSQ}}(x)}{\partial x} = \frac{\ln\left(\frac{2-\alpha}{\alpha}\right)}{2(1-\alpha)} \cdot \mathrm{sech}^2\left(\frac{\ln\left(\frac{2-\alpha}{\alpha}\right)}{\Delta}(x - m_i)\right) \tag{32}$$

Proof. Using the definition of DSQ function (See equation 31) and the fact that $\frac{d}{du}\tanh(u) = \mathrm{sech}^2(u)$, we obtain that for all $x \in P_i$,

$$\frac{\partial Q_{\mathrm{DSQ}}(x)}{\partial x} = \frac{\Delta}{2} \cdot s \cdot k \cdot \mathrm{sech}^2(k(x - m_i)).$$

Substituting $s = \frac{1}{1-\alpha}$ and $k = \frac{1}{\Delta}\ln\left(\frac{2-\alpha}{\alpha}\right)$ provided in equation 29 yields the result equation 32, which completes the proof. □

We first present several existing integral results, which will be used to derive the expectation and variance of the RFS and DSQ schemes, respectively.

Lemma C.5. *The following integral identities hold:*

(i) If $c^2 < 1$, then

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{1 + c\cos x}dx = \frac{4}{\sqrt{1-c^2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right).$$

(ii) If $c^2 < 1$, then

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{(1 + c\cos x)^2}dx = \frac{4}{(1-c^2)^{3/2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - \frac{2c}{1-c^2}.$$

(iii) Given any $c > 0$,

$$\int_{-c}^{c} sech^4(x)dx = 2\left(\tanh(c) - \frac{\tanh^3(c)}{3}\right).$$

Proof. (i) Using the result from Gradshteyn & Ryzhik (2007) and the fact that $c^2 < 1$, we have that

$$\int_{-\pi/2}^{\pi/2} \frac{1}{1 + c\cos x}dx = \frac{2}{\sqrt{1-c^2}}\arctan\left(\frac{(1-c)\tan\left(\frac{x}{2}\right)}{\sqrt{1-c^2}}\right)\Bigg|_{-\pi/2}^{\pi/2} = \frac{4}{\sqrt{1-c^2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right),$$

as claimed.
(ii) It follows from Gradshteyn & Ryzhik (2007) and the fact that $c^2 < 1$ that

$$\int_{-\pi/2}^{\pi/2} \frac{1}{(1+c\cos x)^2}dx = -\frac{1}{1-c^2}\left\{\frac{c\sin(x)}{1+c\cdot\cos(x)} - \frac{2}{\sqrt{1-c^2}}\arctan\left(\frac{(1-c)\tan\left(\frac{x}{2}\right)}{\sqrt{1-c^2}}\right)\right\}\Bigg|_{-\pi/2}^{\pi/2}$$
$$= \frac{4}{(1-c^2)^{3/2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - \frac{2c}{1-c^2},$$

as claimed.
(iii) Let $u = \tanh(x)$ and this gives $du = \operatorname{sech}^2(x)\,dx$. It then follows from the fact that $\operatorname{sech}^4(x) = \operatorname{sech}^2(x)\cdot(1-\tanh^2(x))$ that

$$\int_{-c}^{c}\operatorname{sech}^4(x)\,dx = \int_{\tanh(-c)}^{\tanh(c)}(1-u^2)\,du = \left(u-\frac{u^3}{3}\right)\Bigg|_{\tanh(-c)}^{\tanh(c)} = 2\left(\tanh(c) - \frac{\tanh^3(c)}{3}\right),$$

as claimed. □

With these integral identities, we can now derive closed-form expressions for the statistical properties of each gradient estimator. The following lemma provides the expectation and variance under uniform sampling.

Lemma C.6 (Expectation and variance)**.** *The expectation and variance of the gradient estimators under DSQ and RFS schemes are given as follows.*

1. ***(DSQ)***

$$\mathbb{E}_{\xi\sim U(l,u)}[g_{\mathrm{DSQ}}(x;\alpha)] = 1, \mathrm{Var}_{\xi\sim U(l,u)}[g_{\mathrm{DSQ}}(x;\alpha)] = \frac{\ln\left(\frac{2-\alpha}{\alpha}\right)\left(3-(1-\alpha)^2\right)}{6(1-\alpha)} - 1.$$

2. ***(RFS)*** *Let $c = \sqrt{2}\pi A$ with $A \in \left(0, \frac{1}{\sqrt{2}\pi}\right)$. Then,*

$$\mathbb{E}_{\xi\sim U(l,u)}[g_{\mathrm{RFS}}(x;A)] = \frac{8}{\pi\sqrt{1-c^2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - 1,$$
$$\mathrm{Var}_{\xi\sim U(l,u)}[g_{\mathrm{RFS}}(x;A)] = \frac{16}{\pi(1-c^2)^{3/2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - \frac{8c}{\pi(1-c^2)} +$$
$$1 - \left(\frac{8}{\pi\sqrt{1-c^2}}\arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - 1\right)^2.$$

Proof. For notation brevity, we use $\mathbb{E}[\cdot]$ to denote $\mathbb{E}_{\xi\sim U(l,u)}[\cdot]$ and $\mathrm{Var}[\cdot]$ to denote $\mathrm{Var}_{\xi\sim U(l,u)}[\cdot]$. Let $\beta = \mathrm{In}(\frac{2-\alpha}{\alpha})$.

DSQ Expectation. We have from Lemma C.4, change of variable trick, the fact that range $[l, u]$ has $2^b - 1$ interval with length Δ, $\frac{d}{dx}\tanh(x) = \text{sech}^2(x)$, and equation 30 that

$$\begin{aligned}
\mathbb{E}[(\xi;\alpha)] &= \sum_{i=0}^{2^b-2} \int_{m_i-\frac{\Delta}{2}}^{m_i+\frac{\Delta}{2}} \frac{\beta}{2(1-\alpha)} \cdot \text{sech}^2\left(\frac{\beta}{\Delta}(\xi - m_i)\right) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-2} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{\beta}{2(1-\alpha)} \cdot \text{sech}^2\left(\frac{\beta}{\Delta}\xi\right) \cdot \frac{1}{u-l} d\xi \\
&= \cdot \frac{2^b-1}{u-l} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{\beta}{2(1-\alpha)} \cdot \text{sech}^2\left(\frac{\beta}{\Delta}\xi\right) d\xi \\
&= \frac{1}{\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{\beta}{2(1-\alpha)} \cdot \text{sech}^2\left(\frac{\beta}{\Delta}\xi\right) d\xi \\
&= \frac{1}{\Delta} \cdot \frac{\beta}{2(1-\alpha)} \cdot \frac{\Delta}{\beta} \cdot \tanh\left(\frac{\beta}{\Delta}\xi\right)\Big|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \\
&= \frac{1}{1-\alpha}\tanh\left(\frac{\beta}{2}\right) = \frac{1}{1-\alpha}\tanh\left(\frac{1}{2}\ln\left(\frac{2-\alpha}{\alpha}\right)\right) = 1.
\end{aligned} \tag{33}$$

DSQ Variance. To compute DSQ variance, one first need to compute the second order moment of random variable $(\xi;\alpha)$. We have from Lemma C.4, change of variable trick, the fact that range $[l, u]$ has $2^b - 1$ interval with length Δ, Lemma C.5(iii), and equation 30 that

$$\begin{aligned}
\mathbb{E}[^2(\xi;\alpha)] &= \sum_{i=0}^{2^b-2} \int_{m_i-\frac{\Delta}{2}}^{m_i+\frac{\Delta}{2}} \left(\frac{\ln\left(\frac{2-\alpha}{\alpha}\right)}{2(1-\alpha)}\right)^2 \cdot {}^4\left(\frac{\ln\left(\frac{2-\alpha}{\alpha}\right)}{\Delta}(\xi - m_i)\right) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-2} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \left(\frac{\beta}{2(1-\alpha)}\right)^2 {}^4\left(\frac{\beta}{\Delta}\xi\right) d\xi \\
&= \frac{1}{\Delta} \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} \left(\frac{\beta}{2(1-\alpha)}\right)^2 \cdot \frac{\Delta}{\beta} \cdot {}^4(\xi) d\xi \\
&= \frac{\beta}{4(1-\alpha)^2} \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} {}^4(\xi) d\xi \\
&= \frac{\beta}{4(1-\alpha)^2} \left(2\tanh\left(\frac{\beta}{2}\right) - \frac{2}{3}\tanh^3\left(\frac{\beta}{2}\right)\right) \\
&= \frac{\beta}{4(1-\alpha)^2} \left(2(1-\alpha) - \frac{2}{3}(1-\alpha)^3\right) = \frac{\ln\left(\frac{2-\alpha}{\alpha}\right) \cdot (3 - (1-\alpha)^2)}{6(1-\alpha)}.
\end{aligned} \tag{34}$$

It then follows from the definition of variance, equation 33 and equation 34 that

$$\text{Var}[(\xi;\alpha)] = \mathbb{E}[^2(\xi;\alpha)] - (\mathbb{E}[(\xi;\alpha)])^2 = \frac{\ln\left(\frac{2-\alpha}{\alpha}\right)\left(3 - (1-\alpha)^2\right)}{6(1-\alpha)} - 1.$$

For $N = 1$, the RFS gradient is given as for all $\xi \in [l, u]$,

$$g_{RFS}(\xi) = \frac{1 - \sqrt{2}\pi A \cos\left(\pi\left(\frac{\xi}{\Delta} + \left\lfloor\frac{\xi}{\Delta}\right\rceil\right)\right)}{1 + \sqrt{2}\pi A \cos\left(\pi\left(\frac{\xi}{\Delta} + \left\lfloor\frac{\xi}{\Delta}\right\rceil\right)\right)} = \frac{1 - c \cdot \cos\left(\pi\left(\frac{\xi}{\Delta} - \left\lfloor\frac{\xi}{\Delta}\right\rceil\right)\right)}{1 + c \cdot \cos\left(\pi\left(\frac{\xi}{\Delta} - \left\lfloor\frac{\xi}{\Delta}\right\rceil\right)\right)}. \tag{35}$$

RFS Expectation. Using equation 35, the change of variable trick ($\theta = \frac{x}{\Delta} - \lfloor \frac{x}{\Delta} \rfloor$), the fact that range $[l, u]$ has $2^b - 1$ interval with length Δ, and Lemma C.5(i) that

$$\begin{aligned}
\mathbb{E}[g_{RFS}(\xi; A)] &= \int_l^u g_{RFS}(\xi) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-3} \int_{m_i}^{m_{i+1}} g_{RFS}(\xi) \cdot \frac{1}{u-l} d\xi + \int_l^{l+\frac{\Delta}{2}} g_{RFS}(\xi) \cdot \frac{1}{u-l} d\xi + \\
&\quad \int_{u-\frac{\Delta}{2}}^{u} g_{RFS}(\xi) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-3} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \cdot \frac{\Delta}{\pi(u-l)} d\theta + \int_0^{\frac{\pi}{2}} \frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \cdot \frac{\Delta}{\pi(u-l)} d\theta + \\
&\quad \int_{-\frac{\pi}{2}}^{0} \frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \cdot \frac{\Delta}{\pi(u-l)} d\theta \\
&= \frac{(2^b - 1)\Delta}{u-l} \cdot \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} d\theta \\
&= \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} d\theta = \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{2}{1 + c \cos\theta} - 1 \right) d\theta = \frac{8}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}} \right) - 1.
\end{aligned} \tag{36}$$

RFS Variance. Using the change of variable trick ($\theta = \frac{\xi}{\Delta} - \lfloor \frac{\xi}{\Delta} \rfloor$), the fact that range $[l, u]$ has $2^b - 1$ interval with length Δ, and Lemma C.5(i)-(ii) that

$$\begin{aligned}
&\mathbb{E}[g_{RFS}^2(\xi; A)] \\
&= \int_l^u g_{RFS}^2(\xi) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-3} \int_{m_i}^{m_{i+1}} g_{RFS}^2(\xi) \cdot \frac{1}{u-l} d\xi + \int_l^{l+\frac{\Delta}{2}} g_{RFS}^2(\xi) \cdot \frac{1}{u-l} d\xi + \int_{u-\frac{\Delta}{2}}^{u} g_{RFS}^2(\xi) \cdot \frac{1}{u-l} d\xi \\
&= \sum_{i=0}^{2^b-3} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \right)^2 \cdot \frac{\Delta}{\pi(u-l)} d\theta + \int_0^{\frac{\pi}{2}} \left(\frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \right)^2 \cdot \frac{\Delta}{\pi(u-l)} d\theta + \\
&\quad \int_{-\frac{\pi}{2}}^{0} \left(\frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \right)^2 \cdot \frac{\Delta}{\pi(u-l)} d\theta \\
&= \frac{(2^b - 1)\Delta}{\pi(u-l)} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \right)^2 d\theta \\
&= \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{1 - c \cdot \cos(\theta)}{1 + c \cdot \cos(\theta)} \right)^2 d\theta \\
&= \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{4}{(1 + c \cdot \cos\theta)^2} - \frac{4}{1 + c \cdot \cos\theta} + 1 \, d\theta \\
&= \frac{16}{\pi(1-c^2)^{3/2}} \arctan\left(\sqrt{\frac{1-c}{1+c}} \right) - \frac{8c}{\pi(1-c^2)} + 1 - \frac{16}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}} \right).
\end{aligned} \tag{37}$$

It then follows from equation 36 and equation 37 that

$$\begin{aligned}
&\mathrm{Var}[g_{RFS}(\xi; A)] \\
&= \mathbb{E}[g_{RFS}^2(\xi; A)] - (\mathbb{E}[g_{RFS}(\xi; A)])^2 \\
&= \frac{16c^2}{\pi(1-c^2)^{3/2}} \arctan\left(\sqrt{\frac{1-c}{1+c}} \right) - \frac{8c}{\pi(1-c^2)} + 1 - \left(\frac{8}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}} \right) - 1 \right)^2.
\end{aligned}$$

□

Now, we are ready to prove Theorem 4.2. For ease of reading, we restate the theorem as follows.

Theorem C.7. *Let $g_{\text{DSQ}}(\cdot;\alpha)$ and $g_{\text{RFS}}(\cdot;A)$ be the gradient surrogates of DSQ and RFS, parameterized by $\alpha > 0$ and $A \in \left(0, 1/(\sqrt{2}\pi)\right)$, respectively. Let random variable ξ be uniformly distributed on the interval $[l,u]$, i.e., $\xi \sim U(l,u)$. Consider the asymptotic regimes in which the surrogate functions move close to the rounding function. Then, the limits of the expectations satisfy*

$$\lim_{\alpha\to 0^+} \mathbb{E}_{\xi\sim U(l,u)}[g_{\text{DSQ}}(\xi;\alpha)] = 1, \quad \lim_{A\to\left(1/(\sqrt{2}\pi)\right)^-} \mathbb{E}_{\xi\sim U(l,u)}[g_{\text{RFS}}(\xi;A)] \quad = 4/\pi - 1,$$

and the limits of the variances satisfy

$$\lim_{\alpha\to 0^+} \text{Var}_{\xi\sim U(l,u)}[g_{\text{DSQ}}(\xi;\alpha)] = \infty, \quad \lim_{A\to\left(1/(\sqrt{2}\pi)\right)^-} \text{Var}_{\xi\sim U(l,u)}[g_{\text{RFS}}(\xi;A)] \quad = 16/(3\pi) - 16/\pi^2.$$

Proof of Theorem 4.2. We establish each limit separately, drawing upon the closed-form expressions derived in Lemma C.6. Throughout, we denote $c = \sqrt{2}\pi A$ and note that as $A \to \left(\frac{1}{\sqrt{2}\pi}\right)^-$, we have $c \to 1^-$.

DSQ expectation limit. From Lemma C.6, the expected value of the DSQ gradient satisfies $\mathbb{E}_{\xi\sim U(l,u)}[(\xi)] = 1$ for all $\alpha \in (0,1)$. Since this identity is independent of α, it follows that

$$\lim_{\alpha\to 0^+} \mathbb{E}[(\xi;\alpha)] = 1.$$

RFS expectation limit. From Lemma C.6, the expectation of the RFS gradient is given by

$$\mathbb{E}_{\xi\sim U(l,u)}[g_{RFS}(\xi;A)] = \frac{8}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - 1.$$

It then follows from the fact that $\lim_{x\to 0} \frac{\arctan(x)}{x} = 1$ that

$$\begin{aligned} \lim_{A\to\left(\frac{1}{\sqrt{2}\pi}\right)^-} \mathbb{E}[g_{RFS}(x)] &= \lim_{c\to 1^-} \frac{8}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - 1 \\ &= \lim_{c\to 1^-} \frac{\arctan\left(\sqrt{\frac{1-c}{1+c}}\right)}{\left(\sqrt{\frac{1-c}{1+c}}\right)} \cdot \frac{8}{\pi\sqrt{1-c^2}} - 1 = \frac{4}{\pi} - 1. \end{aligned} \tag{38}$$

DSQ variance limit. From Lemma C.6, the variance of the DSQ gradient is given by

$$\text{Var}_{\xi\sim U(l,u)}[(\xi;\alpha)] = \frac{\ln\left(\frac{2-\alpha}{\alpha}\right)\left(3-(1-\alpha)^2\right)}{6(1-\alpha)} - 1. \tag{39}$$

We first analyze the asymptotic behavior of each factor as $\alpha \to 0^+$ as follows.

$$\lim_{\alpha\to 0^+} \ln\left(\frac{2-\alpha}{\alpha}\right) = +\infty, \ \lim_{\alpha\to 0^+} (1-\alpha) = 1, \ \lim_{\alpha\to 0^+} (3-(1-\alpha)^2) = 2. \tag{40}$$

Combining equation 39 and equation 40 allows us to have that

$$\lim_{\alpha\to 0^+} \text{Var}[(\xi;\alpha)] = \lim_{\alpha\to 0^+} \frac{\ln\left(\frac{2-\alpha}{\alpha}\right)\left(3-(1-\alpha)^2\right)}{6(1-\alpha)} - 1 = +\infty.$$

RFS variance limit. From Lemma C.6, the variance of the RFS gradient is given by

$$\begin{aligned} &\text{Var}_{\xi\sim U(l,u)}[g_{RFS}(\xi;A)] \\ &= \underbrace{\frac{16c^2}{\pi(1-c^2)^{3/2}} \arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - \frac{8c}{\pi(1-c^2)}}_{T_1(c)} + 1 - \underbrace{\left(\frac{8}{\pi\sqrt{1-c^2}} \arctan\left(\sqrt{\frac{1-c}{1+c}}\right) - 1\right)^2}_{T_2(c)}. \end{aligned}$$

We note that $\lim_{c\to 1^-} T_2(c)$ has been derived in equation 38. To this end, it remains to compute $\lim_{c\to 1^-} T_1(c)$. Let $u^2 = \frac{1-c}{1+c}$ and this implies that

$$c = \frac{1-u^2}{1+u^2} \text{ and } 1-c^2 = \frac{4u^2}{(1+u^2)^2}. \tag{41}$$

Then, using equation 41 and Taylor series expansion of function $\arctan(x)$ at $x = 0$, we have that

$$\begin{aligned}
\lim_{c\to 1^-} T_1(c) &= \lim_{u\to 0^+} \frac{16(1-u^2)^2}{\pi 8u^3}\arctan(u) - \frac{8\left(\frac{1-u^2}{1+u^2}\right)}{\pi\left(\frac{4u^2}{(1+u^2)^2}\right)} \\
&= \lim_{u\to 0^+} \frac{16(1-u^2)^2(1+u^2)}{\pi 8u^3}\left(u - \frac{u^3}{3} + O(u^5)\right) - \frac{8(1-u^2)(1+u^2)}{4\pi u^2} \\
&= \lim_{u\to 0^+} \frac{16(1-u^2)^2(1+u^2)u}{24\pi} - \frac{8(1-u^2)(1+u^2)}{4\pi u^2} \\
&= \lim_{u\to 0^+} \frac{2(1-u^2-u^4+u^6)}{3\pi} - \frac{2(1-2u^2+u^4)}{\pi u^2} \\
&= \lim_{u\to 0^+} \left(-\frac{8}{3\pi} + O(u^2)\right) = -\frac{8}{3\pi}.
\end{aligned} \tag{42}$$

Combining equation 38 and equation 42 allows us to have that

$$\lim_{A\to\left(\frac{1}{\sqrt{2}\pi}\right)^-} \mathrm{Var}_{\xi\sim U(l,u)}[g_{RFS}(\xi;A)] = \lim_{c\to 1^-} T_1(c) + 1 - T_2(c) = \frac{16}{3\pi} - \frac{16}{\pi^2}.$$

□

$$\frac{\partial x_q}{\partial x} = \frac{1 - A\sqrt{2}\pi\cos(\pi(x+x_q))}{1 + A\sqrt{2}\pi\cos(\pi(x+x_q))}. \tag{43}$$

$$\frac{\partial x_q}{\partial x} = \frac{1 - A\sqrt{2}\pi\sum_{m=0}^{M}\frac{(-1)^m}{2m+1}\cos((2m+1)\pi(x+x_q))}{1 + A\sqrt{2}\pi\sum_{m=0}^{M}\frac{(-1)^m}{2m+1}\cos((2m+1)\pi(x+x_q))} \tag{44}$$

$$t = (x+x_q)/\sqrt{2}$$