PiSO: Scale Optimization in Post-Training Quantization

Anonymous authors
Paper under double-blind review

Abstract

Post-training quantization (PTQ) compresses large language models by mapping weights to low-bit representations. While much work has focused on how weights are rounded onto the quantization grid, the scaling factor that defines the grid is typically chosen with simple, data-free heuristics. We observe that, under round-to-nearest (RTN) quantization, the grid assignments are piecewise constant in the scale, making the data-aware layer-output error piecewise quadratic over finitely many intervals, each with a closed-form minimizer. Building on this insight, we present PiSO (Piecewise Scale Optimization), an algorithm that leverages calibration data to efficiently compute the globally optimal channel-wise scales under RTN. We extend PiSO to group-wise quantization via tractable objectives. We also propose strategies for integrating it with error-correction methods such as GPTQ and Qronos, which greedily adjust weight assignments to compensate for the rounding errors of previously quantized weights. Experiments on Llama and Qwen models across multiple sizes and bit-widths show consistent improvements in perplexity and zero-shot accuracy, both standalone and combined with error correction, with gains increasing as the bit-width narrows. For instance, on Llama-3, PiSO reduces the 2-bit perplexity of Qronos by up to 7×, and at 3 bits, RTN with PiSO scales alone outperforms GPTQ with absmax scales.

1 Introduction

Large language models (LLMs) incur significant memory and compute costs that hinder deployment on resource-constrained hardware. Post-training quantization (PTQ) addresses this by mapping weights, and sometimes activations, to low-precision representations without retraining. The target representation is a discrete grid defined by a bit-width, scaling factor, and (optionally) a zero point. Greedy error correction algorithms such as GPTQ (Frantar et al., 2022), and more recently GPTAQ (Li et al., 2025) or Qronos (Zhang et al., 2026), are widely used to minimize accuracy degradation when mapping to such grids. Crucially, these methods assume a fixed grid during error correction and, in practice, select parameters such as scaling factors beforehand.

Common scale-selection rules either map the maximum absolute weight to the extreme grid point, the so-called absmax grid, or minimize weight reconstruction error over a finite set of candidate scales between zero and absmax (Zhang et al., 2026; Egiazarian et al., 2026; Sanjeet et al., 2026). Both approaches are *data-free*: they ignore how input activations amplify different weight errors at the layer output. Previous work on quantization of convolutional neural networks by Idelbayev et al. (2021) introduced an exact, data-free scale optimizer for RTN-quantized weights, which, to the best of our knowledge, has not been adopted in LLM PTQ pipelines. Recent methods (Nair & Suggala, 2024; Zhang et al., 2025a; Zhang & Saab, 2026) address data-aware scale selection in LLM PTQ through grid search or coupled with coordinate descent procedures, without global-optimality guarantees (see Section 2). More broadly, scale selection for LLMs has received little formal study, particularly relative to weight-to-grid assignment. In this work, we show that the optimal channel-wise scales for a data-aware layer-output objective can be computed exactly and efficiently under round-to-nearest (RTN) quantization by exploiting the piecewise structure of the weight-to-grid assignment. We then extend our algorithm to handle group-wise scaling, and propose several levels of integration with error correction algorithms.

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.

Contributions. Our main contributions are: (1) PiSO, an algorithm that efficiently computes *optimal* scales for channel-wise weight quantization with respect to a *data-aware* layer output reconstruction objective under RTN quantization; (2) two tractable objectives for extending PiSO to group-wise quantization, one of which accounts for cross-group interactions; and (3) three principled strategies for integrating PiSO with GPTQ and Qronos that empirically improve the performance of the resulting quantized models. We evaluate PiSO on Llama and Qwen models across multiple sizes, bit-widths, and quantization granularities. It consistently improves perplexity and zero-shot accuracy, both standalone and combined with GPTQ or Qronos, with larger gains at lower bit-widths. At 3 bits, the best results on Llama-3 are obtained by combining PiSO with error correction, and even RTN+PiSO alone outperforms GPTQ with the absmax grid. Furthermore, our calibration size study also points to improved data efficiency: on Llama-3.2-1B with 3-bit channel-wise quantization, RTN + PiSO needed only a few calibration sequences, and GPTQ with PiSO scales outperformed GPTQ with absmax scales using $32\times$ less calibration data.

2 Background and Related Work

Notation. Throughout the paper, $\mathbf{W} \in \mathbb{R}^{D\times M}$ denotes the weight matrix of a single layer with D input dimensions and M output channels, and $\mathbf{w}_j := \mathbf{W}_{:,j} \in \mathbb{R}^D$ is the weight vector of channel j. $\tilde{\mathbf{W}}$ denotes the quantized (and possibly error-corrected) weight matrix. $\mathbf{X} \in \mathbb{R}^{N\times D}$ denotes the input activation matrix of a given layer of the high-precision model, while $\tilde{\mathbf{X}}$ is the input activation matrix produced by the (partially) quantized model.

For a finite grid $\mathbb{G} = \{g_1 < g_2 < \cdots < g_L\} \subset \mathbb{R}$ and *scale* $s > 0$, we define the symmetric quantizer as

$$\mathcal{Q}(x;\, s) \;:=\; s \cdot q(x;\, s), \qquad \text{where} \quad q(x;\, s) := \left\lfloor \tfrac{x}{s} \right\rceil_{\mathbb{G}} \tag{1}$$

is the *unscaled grid assignment*.

Here $\lfloor\cdot\rceil_{\mathbb{G}}$ denotes the round-to-nearest (RTN) operator on the grid: $\lfloor y\rceil_{\mathbb{G}} := \arg\min_{g\in\mathbb{G}} |g - y|$, with a fixed deterministic tie-breaking rule. For signed B-bit integers, $\mathbb{G} = \{-2^{B-1}, -2^{B-1} + 1, \ldots, 2^{B-1} - 1\}$. Channel-wise quantization uses one scale per output channel, whereas group-wise quantization partitions each channel into contiguous groups of weights, each with its own scale. Concretely, given a channel $\mathbf{w} \in \mathbb{R}^D$ and a group size G, $\mathbf{w}$ is partitioned into $K = \lceil D/G \rceil$ groups $\mathbf{w}_{\{1\}}, \ldots, \mathbf{w}_{\{K\}} \in \mathbb{R}^G$. The activation matrices are partitioned column-wise according to the same grouping. We write $\mathbf{X}_{\{k\}}$ for the columns of $\mathbf{X}$ associated with group k. Subscripts $\{\leq k\}$ and $\{< k\}$ denote prefix concatenation over groups, i.e., $\mathbf{w}_{\{\leq k\}} := [\mathbf{w}_{\{1\}}, \ldots, \mathbf{w}_{\{k\}}]^\top$ where, unless stated otherwise, concatenation is along the column dimension.

Post-training quantization. Following Zhang et al. (2026), we distinguish two stages of the PTQ pipeline: (1) *transform*, and (2) *round*. Transform methods modify weights and/or activations to make them more amenable to quantization (e.g., by applying rotations to reduce outliers (Ashkboos et al., 2024; Liu et al., 2025)), whereas round methods map (possibly transformed) weights onto a quantization grid, aiming to minimize a layer-wise reconstruction objective such as:

$$E_{\text{GPTQ}}(\tilde{\mathbf{W}}) = \left\|\tilde{\mathbf{X}}\mathbf{W} - \tilde{\mathbf{X}}\tilde{\mathbf{W}}\right\|_F^2 \quad \text{or} \quad E(\tilde{\mathbf{W}}) = \left\|\mathbf{X}\mathbf{W} - \tilde{\mathbf{X}}\,\tilde{\mathbf{W}}\right\|_F^2. \tag{2}$$

The second objective explicitly accounts for activation drift introduced by quantizing previous layers.[1] *Round* algorithms can be further subdivided based on their optimization approach:

Gradient-based rounding. These algorithms use backpropagation to learn rounding decisions and, in some cases, quantization parameters (Nagel et al., 2020; Li et al., 2021; Hubara et al., 2021; Cheng et al., 2024; Shao et al., 2024). However, their computational cost has limited their applicability to LLMs until recently (Cheng et al., 2024), they require careful hyperparameter tuning, and risk overfitting to the calibration set (Li et al., 2021; Hubara et al., 2021).

[1] The first objective is widely adopted as it is used in GPTQ and many others (Frantar & Alistarh, 2022; Nair & Suggala, 2024; Zhang et al., 2025a; Zhang & Saab, 2026), while methods like GPFQ (Lybrand & Saab, 2021), GPTAQ and Qronos adopt the second.

Greedy rounding. These algorithms typically assume a fixed quantization grid and process weights sequentially in a given layer: each weight is assigned to a grid point, and the resulting quantization error is compensated for in the remaining unquantized weights so as to minimize the layer output error. OBC (Frantar & Alistarh, 2022) introduced this framework by adapting Optimal Brain Surgeon (Hassibi et al., 1993) to layer-wise PTQ, and GPTQ (Frantar et al., 2022) scaled it to LLMs through an efficient implementation. Qronos (Zhang et al., 2026) and GPTAQ (Li et al., 2025) extend GPTQ to account for the quantization error propagated from previously quantized layers.

Scale selection. The choice of scale determines the rounding errors that greedy methods must compensate (Frantar et al., 2022). The dominant choice in PTQ pipelines remains *data-free*: either the absmax scale, or weight-error grid-search over a finite set of candidates in $(0, s_{\text{absmax}}]$ (Zhang et al., 2026; Egiazarian et al., 2026; Sanjeet et al., 2026). Both ignore how input activations amplify weight errors at the layer output, and both implicitly cap s at s_{absmax}, a restriction that is not always benign (as exemplified in Figure 4 in Appendix E). Previous work on CNN quantization by Idelbayev et al. (2021) derived a globally optimal, data-free scale optimizer for weights quantized with RTN. The method minimizes weight reconstruction error, so it ignores how activations amplify and correlate errors at the layer output. EasyQuant (Wu et al., 2020) proposes a data-aware quantization procedure for CNNs that alternates grid searches over weight and activation scales using convolution-output cosine similarity. In the setting most relevant to LLM PTQ, CDQuant (Nair & Suggala, 2024), COMQ (Zhang et al., 2025a), and Beacon (Zhang & Saab, 2026) propose different data-aware methods for weight scale selection together with their own coordinate-descent procedures for optimizing the weight-to-grid assignments. CDQuant selects scales before rounding via grid-search on the layer output error; this inherits the suboptimal $s \leq s_{\text{absmax}}$ constraint and the discrete granularity of the search. COMQ and Beacon jointly refine scales and grid assignments, an NP-hard problem where they do not offer global-optimality guarantees. Moreover, COMQ and Beacon are restricted to channel-wise quantization, while CDQuant's group-wise variant treats groups independently, ignoring cross-group interactions through the activations. A complementary line of work focuses on scale selection in Quantization Aware Training (QAT) (Esser et al., 2020; Liu et al., 2026) or gradient-based PTQ (Shao et al., 2024; Cheng et al., 2024). These approaches require backpropagation, are often compute-intensive and need careful hyperparameter tuning, whereas our method is gradient-free and does not require hyperparameter tuning.

3 METHODS

We first derive PiSO for channel-wise RTN scale optimization. We then extend it to group-wise quantization, via two different tractable objectives; and finally propose three strategies for integration with greedy error correction methods such as GPTQ and Qronos.

We start by restricting ourselves to the RTN quantizer (Equation 1). Therefore, we express the layer output reconstruction objective (Equation 2) as a function of the scaling factors:

$$E(\mathbf{s}) = \left\|\mathbf{X}\mathbf{W} - \tilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{W};\mathbf{s})\right\|_F^2, \quad \mathcal{Q}(\mathbf{W};\mathbf{s}) := \mathbf{s}\cdot\left\lfloor\frac{\mathbf{W}}{\mathbf{s}}\right\rceil_{\mathbb{G}}. \tag{3}$$

This restriction is principled: for weight MSE minimization (data-free), the optimal grid assignments are directly determined by the scales and achieved by RTN (Proposition C.2). It also makes the data-aware problem tractable: jointly optimizing scales and assignments is an NP-hard integer least-squares problem (Remark C.3), whereas under RTN the scales minimizing Equation 3 can be computed *exactly* in polynomial time despite its non-convexity. PiSO rests on three observations: (1) separability across output channels (Zhang & Saab, 2026); (2) a closed-form optimum for a fixed grid assignment (Zhang & Saab, 2026; Zhang et al., 2025a); and (3) **the key insight**: since $q(\mathbf{w}; s)$ is *piecewise constant* in s, the real line (search space of s) splits into finitely many intervals, each with a closed-form minimizer of the error.

3.1 OPTIMAL SCALE ALGORITHM FOR CHANNEL-WISE QUANTIZATION

We now formalize these observations and combine them to propose the scale optimization algorithm PiSO (Algorithm 1 in Appendix A). For intuition, Appendix D works through a two-dimensional example that contrasts data-aware and data-free scale selection.

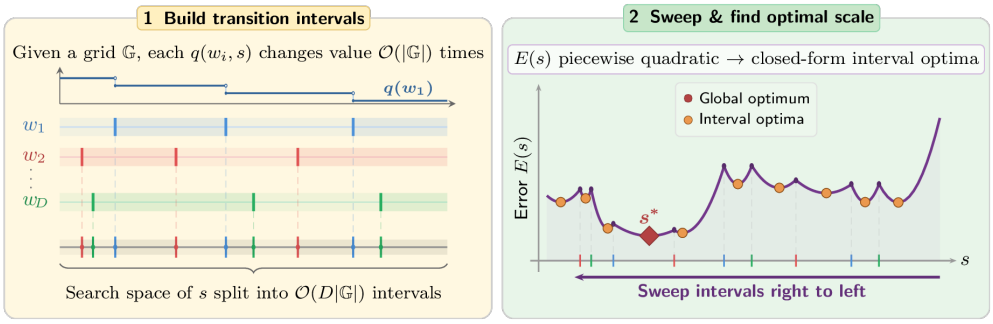


Figure 1: Overview of PiSO. The grid assignment $q(\mathbf{w}; s)$ is piecewise constant in the scale, partitioning the real line into intervals where the objective in Equation 3 simplifies to a quadratic. PiSO sweeps through these intervals, and returns the scale achieving the lowest global error.

Channel-wise separability. Writing $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$ as a collection of column vectors, each with its own scale, the objective in Equation 3 separates into M independent problems (Lemma C.5):

$$E(\mathbf{s}) = \sum_{j=1}^{M} \left\| \mathbf{X}\mathbf{w}_j - \tilde{\mathbf{X}}\, \mathcal{Q}(\mathbf{w}_j;\, s_j) \right\|^2 \quad = \quad \sum_{j=1}^{M} E_j(s_j). \tag{4}$$

Each s_j can therefore be optimized independently. In what follows, we drop the subscript j and focus on finding the optimal scale s for a single output channel with weight vector $\mathbf{w} \in \mathbb{R}^D$.

Closed-form solution for fixed grid assignments. Using $\mathcal{Q}(\mathbf{w};\, s) = s\, q(\mathbf{w};\, s)$ from Equation 1, the channel-wise error reads

$$E(s) = \|\mathbf{X}\mathbf{w} - s\, \tilde{\mathbf{X}}\, q(\mathbf{w};\, s)\|^2, \tag{5}$$

which is non-convex through the rounding operator in $q(\cdot)$. Yet, for any *fixed* grid assignment $\mathbf{q}$, $E(s)$ is convex in s with a closed-form minimizer (Lemma C.6, Appendix C). Defining

$$\mathbf{H} := \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}, \quad \mathbf{G} := \tilde{\mathbf{X}}^\top \mathbf{X} \in \mathbb{R}^{D \times D}, \qquad \alpha(\mathbf{q}) := \mathbf{q}^\top \mathbf{G}\, \mathbf{w}, \quad \beta(\mathbf{q}) := \mathbf{q}^\top \mathbf{H}\, \mathbf{q}, \tag{6}$$

the optimal scale for a fixed $\mathbf{q}$ and its associated error (up to a s-independent constant) are

$$s^*(\mathbf{q}) = \alpha(\mathbf{q})/\beta(\mathbf{q}), \qquad E(s \mid \mathbf{q}) = \text{const} - 2s\, \alpha(\mathbf{q}) + s^2\, \beta(\mathbf{q}) \tag{7}$$

as shown in Appendix C (Remark C.7). Equation 5 accommodates several layer-wise reconstruction objectives: (1) *cross-activation* ($\mathbf{X} \neq \tilde{\mathbf{X}}$) as in the objective of Qronos; (2) *self-activation* (replacing $\mathbf{X}$ by $\tilde{\mathbf{X}}$), as in the GPTQ objective; and (3) *data-free* (setting $\mathbf{X}$ and $\tilde{\mathbf{X}}$ to the identity $\mathbf{I}$). Our algorithm is applicable for any of these choices. The *data-free* objective results in an algorithm equivalent to OptQuant by Idelbayev et al. (2021).

Piecewise-constant structure. Equation 7 assumes a fixed grid assignment $\mathbf{q}$, but $q(\mathbf{w};\, s)$ is not constant in s. However, $q(\mathbf{w};\, s)$ is *piecewise constant* in s, with the breakpoints admitting a closed form. For example, for a positive scalar weight w_i, as s decreases from $+\infty$ toward 0 the ratio w_i/s grows and the assignment $q_i = \lfloor w_i/s \rceil_{\mathbb{G}}$ walks through $\mathbb{G}$, jumping from g_ℓ to $g_{\ell+1}$ exactly when w_i/s crosses their midpoint (Corollary C.10, Appendix C). This yields at most $|\mathbb{G}| - 1$ *transition scales* per weight, and $D(|\mathbb{G}| - 1)$ for the full vector $\mathbf{w} \in \mathbb{R}^D$ that follow the form:

$$t_{i,\ell} = \frac{2\, w_i}{g_\ell + g_{\ell+1}}, \qquad i = 1, \ldots, D, \quad \ell = 1, \ldots, |\mathbb{G}| - 1. \tag{8}$$

Crucially, each $t_{i,\ell}$ is associated with a single weight index i: when s crosses $t_{i,\ell}$, only q_i changes while all other entries stay put (Appendix A covers coinciding transitions). Sorting the positive transitions as $\tau_1 > \cdots > \tau_T$, with $T \leq D(|\mathbb{G}| - 1)$, partitions $(0, +\infty)$ into $T + 1$ intervals on which $\mathbf{q}$ is constant. We focus on positive scales, the signed extension is detailed in Appendix A.

Sweeping with incremental updates. Within each interval (τ_j, τ_{j-1}), $\mathbf{q}$ is constant and $E(s)$ is quadratic in s, with minimizer $s_j^* = \text{clamp}(\alpha/\beta,\, \tau_j,\, \tau_{j-1})$, clamped since α/β may fall outside the interval, and associated error $\phi_j = (s_j^*)^2\,\beta - 2\,s_j^*\,\alpha$ (Equations 6 and 7). The lowest ϕ_j across intervals gives the global infimum of the objective, which PiSO attains whenever it is a minimum and approaches arbitrarily closely otherwise (Proposition C.11). Recomputing $\mathbf{q}$, α, and β from scratch at each interval would cost $\mathcal{O}(D^2)$ per interval and $\mathcal{O}(D^3|\mathbb{G}|)$ in total. Instead, we exploit that at each transition boundary one known entry of $\mathbf{q}$ changes: letting i be that index and $\delta := q_i^{\text{new}} - q_i^{\text{old}}$, we maintain the running scalars α, β and the auxiliary vector $\mathbf{h} := \mathbf{H}\,\mathbf{q}$ via

$$\alpha \leftarrow \alpha + \delta \cdot (\mathbf{G}\mathbf{w})_i, \qquad \beta \leftarrow \beta + 2\delta \cdot h_i + \delta^2 \cdot H_{ii}, \qquad \mathbf{h} \leftarrow \mathbf{h} + \delta \cdot \mathbf{H}_{i,:}. \tag{9}$$

The vector $\mathbf{G}\mathbf{w}$ is precomputed once. Updating α and β costs $\mathcal{O}(1)$, while $\mathbf{h}$ costs $\mathcal{O}(D)$, reducing the overall per-channel complexity of PiSO to $\mathcal{O}(D^2|\mathbb{G}|)$ (Proposition C.12, Appendix C).

Algorithm analysis. When $\mathbf{H}$ is diagonal, $\mathbf{h} = \mathbf{H}\mathbf{q}$ does not need to be maintained explicitly, and the update of β becomes $\beta \leftarrow \beta + H_{ii}(2\,q_i^{\text{old}}\delta + \delta^2)$. This reduces the total complexity of sweeping the intervals to $\mathcal{O}(D|\mathbb{G}|)$ per channel, *linear* in weight dimension. This covers the *data-free* regime and, also, using a diagonal approximation of $\mathbf{H}$ and $\mathbf{G}$, trading fidelity for speed. In both the dense, and diagonal case, the M channel problems are embarrassingly parallel.

3.2 Group-wise Quantization

In group-wise quantization with group size G and K groups the channel-wise objective (Equation 5) becomes

$$E(s_1, \ldots, s_K) = \left\| \mathbf{X}\mathbf{w} - \tilde{\mathbf{X}} \left[s_1 q(\mathbf{w}_{\{1\}}; s_1), \ldots, s_K q(\mathbf{w}_{\{K\}}; s_K) \right]^\top \right\|^2. \tag{10}$$

Minimizing Equation 10 can no longer be solved with a one-dimensional sweep. Instead of solving $\mathcal{O}(D|\mathbb{G}|)$ closed-form subproblems on intervals of the real line, it would require solving $\mathcal{O}((G|\mathbb{G}|)^K)$ subproblems on hyper-intervals of $\mathbb{R}^K$.[2] A joint grid search suffers from the same curse of dimensionality: R candidates per scale require R^K evaluations. We propose two tractable objectives for group-wise quantization, which PiSO can efficiently minimize. We note that they are independent of the 1D solver and could equally be paired with other optimization procedures.

Independent groups. Similarly to Nair & Suggala (2024), we discard all cross-group interactions and minimize the block-diagonal approximation of Equation 10:

$$E_{\text{ind}}(s_1, \ldots, s_K) := \sum_{k=1}^{K} \left\| \mathbf{X}_{\{k\}} \mathbf{w}_{\{k\}} - s_k\, \tilde{\mathbf{X}}_{\{k\}}\, q(\mathbf{w}_{\{k\}};\, s_k) \right\|^2. \tag{11}$$

E_{ind} is separable in the K scales, so each group reduces to a channel-wise sweep on G weights with $\mathbf{H}_{kk} = \tilde{\mathbf{X}}_{\{k\}}^\top \tilde{\mathbf{X}}_{\{k\}}$ and $\mathbf{G}_{kk} = \tilde{\mathbf{X}}_{\{k\}}^\top \mathbf{X}_{\{k\}}$. Each group sweep costs $\mathcal{O}(G^2|\mathbb{G}|)$, for a total per-channel cost of $\mathcal{O}(DG|\mathbb{G}|)$ that parallelizes across both groups and channels. Proposition C.13 (Appendix C) bounds the gap between optimizing the original objective and the independent one by the spectral norms of the discarded cross-group blocks of $\mathbf{H}$ and $\mathbf{G}$.

Sequential groups. To account for cross-group interactions one can process each group at a time. After fixing the scales $s_1^*, \ldots, s_{k-1}^*$ for the first $k-1$ groups, s_k is chosen to minimize the reconstruction error in which the already quantized weights (and their scales) are fixed. More precisely,

$$E_{\text{seq}}(s_k) := \left\| \mathbf{X}_{\{\leq k\}}\, \mathbf{w}_{\{\leq k\}} - \tilde{\mathbf{X}}_{\{<k\}}\, \tilde{\mathbf{w}}_{\{<k\}} - s_k\, \tilde{\mathbf{X}}_{\{k\}}\, q(\mathbf{w}_{\{k\}}; s_k) \right\|^2, \tag{12}$$

where $\tilde{\mathbf{w}}_{\{i\}} := s_i^*\, q(\mathbf{w}_{\{i\}};\, s_i^*)$ for $i < k$. The quantization error of preceding groups shifts the target against which group k is optimized. The procedure provides a middle ground between the independent variant and a full joint optimization. Mirroring the column reordering commonly used in GPTQ implementations (IST-DASLab, 2022; Pappalardo et al., 2025), we process groups by aggregated Hessian importance (more details in Appendix A). Each group sweep costs $\mathcal{O}(G^2|\mathbb{G}|)$, for a total per-channel cost of $\mathcal{O}(DG|\mathbb{G}|)$.

[2] Each scale s_k has its own set of transition points, and the joint assignment $(\mathbf{q}^1, \ldots, \mathbf{q}^K)$ is constant only on the intersection of intervals across all K axes.

3.3 Integration with Existing PTQ Methods

Integration with the *transform* stage of the PTQ pipeline (Section 2) is immediate: PiSO can be applied to the transformed weights produced by methods such as rotation-based equalization (Ashkboos et al., 2024; Liu et al., 2025) before the round stage. Composing with *error correction* methods such as GPTQ is less direct. They keep the scale fixed but deliberately move weights away from their nearest grid point to compensate for earlier rounding errors, interfering with the RTN assumption. Note that jointly optimizing the scales and the grid assignments is NP-hard (Zhang & Saab, 2026), as is the data-aware optimization of the grid assignments alone even for a fixed scale (Hassibi & Vikalo, 2002; Zhang et al., 2026) (Remark C.3). Thus, error correction methods greedily assign weights to a fixed grid and do not offer global optimality guarantees (Zhang et al., 2023; 2025b). Moreover, their performance depends on the choice of the scale, as the quantization grid determines the rounding errors they must compensate.

Conditional optimization. Motivated by this, we approximate the joint problem through two complementary updates. In the first, PiSO computes the *conditionally optimal* scale for the current state of the quantized model, using RTN as a surrogate for the grid assignments that error correction has not made yet. For group-wise quantization, optimality is with respect to the objectives of Section 3.2. In the second, error correction greedily assigns the weights to the resulting grid, diffusing error into the weights that have not yet been quantized. We propose three integration strategies that differ in the *granularity of interleaving* these updates, that is, in how much of the model has already been quantized when PiSO is applied, and hence in how much the surrogate has to cover. We describe them for GPTQ, but they also apply to Qronos and related layer-wise methods. Appendix B elaborates on these integrations, providing their pseudocode.

Decoupled (GPTQ + PiSO). A first pass over the model optimizes the scales of all layers, computing each layer's $\tilde{\mathbf{X}}$ by RTN-quantizing the preceding layers with their PiSO scales. A second pass then applies error correction with these scales held fixed. This fully separates the two stages, at the cost of an extra calibration pass. It also creates an activation mismatch, as error correction changes the inputs to which each layer's scales were fitted.

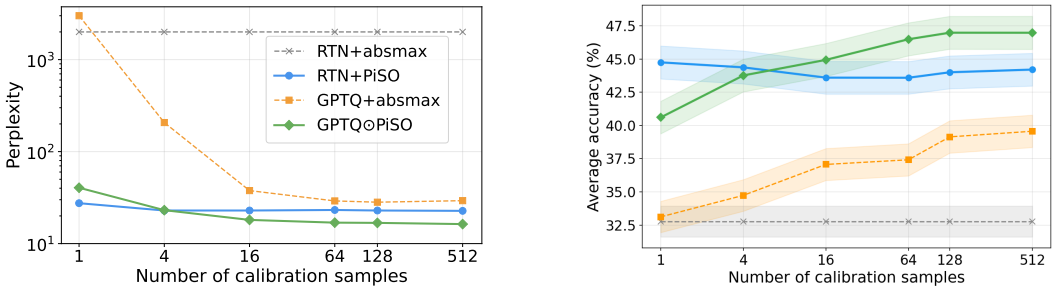
Layer-wise interleaving (GPTQ $\odot$ PiSO). Scale optimization and error correction alternate in a single pass, with each layer's scales optimized immediately before its error correction step. All preceding layers have been both scale-optimized and error-corrected, so the $\tilde{\mathbf{X}}$ used by PiSO matches the activations the layer receives during error correction and at inference. Error correction itself is unchanged and operates on the resulting fixed grid. The surrogate is thus only needed within the layer, for the assignments that error correction makes after the scales are selected.

Group-wise interleaving (GPTQ $\boxdot$ PiSO). In group-wise quantization, the coupling can be tightened further by running PiSO at each group boundary inside the quantization loop of error correction, just before that group is quantized. This works with either objective of Section 3.2, we adjust the sequential one to account for the already diffused error (Appendix B). The scale is thus optimized for weights that already include the error-diffusion updates from all preceding groups, so the surrogate is only needed within the current group. This strategy requires the columns of each group to be processed contiguously, so we use a group-aware ordering: groups are sorted by aggregate Hessian importance, and columns within each group by their own importance (Appendix A).

4 Experiments

We conduct experiments aimed to evaluate: (1) the gains of PiSO's optimal data-aware scales over absmax and data-free scales under RTN; (2) the effectiveness of the independent and sequential objectives for group-wise quantization; and (3) the strategies for integrating PiSO with GPTQ and Qronos. We further compare against data-aware PTQ baselines and study calibration-data efficiency. Throughout the tables, **bold** highlights the best result(s) across each model, bit-width, and metric.

Experimental Setup. We quantize Llama-3 (Grattafiori et al., 2024) (3.2-1B, 3.2-3B, 3.1-8B) and Qwen-2.5 Instruct (Qwen Team et al., 2025) (1.5B, 3B, 7B), using the HuggingFace checkpoints (Wolf et al., 2020). Appendix E extends coverage to Qwen2.5 14B and Gemma models



(a) C4 calibration, WikiText2 perplexity (↓)

(b) WikiText2 calibration, Zero-shot accuracy (↑)

Figure 2: Effect of calibration set size on Llama-3.2-1B for 3-bit channel-wise weight quantization. PiSO is stable with few samples, whereas GPTQ with absmax requires ≥ 64 samples to converge.

Table 1: Channel-wise 2-bit weight-only quantization results on Llama-3 and Qwen-2.5 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). PiSO reduces absmax perplexity by orders of magnitude; the interleaved variants achieve the best results in most configurations.

	2-bit Llama						**2-bit Qwen**					
	WikiText2 (↓)			0-shot (↑)			WikiText2 (↓)			0-shot (↑)		
	1B	3B	8B	1B	3B	8B	1.5B	3B	7B	1.5B	3B	7B
BF16	8.97	7.14	5.89	52.6	62.0	68.6	8.89	7.88	6.92	54.9	56.5	57.7
RTN + absmax	2e5	4e5	4e5	33.0	32.5	33.4	7e6	6e6	1e5	31.8	32.9	32.3
RTN + OptQuant	2e6	8e5	2e6	32.8	32.3	32.1	2e8	3e8	7e6	32.2	31.4	32.2
RTN + PiSO	3e3	3e3	3e3	32.3	31.6	32.0	2e3	2e3	4e3	31.8	32.7	32.2
AWQ	2e6	8e5	1e6	32.7	32.0	32.1	9e7	7e8	9e6	32.4	31.7	32.5
Beacon	52.8	72.4	25.5	33.8	33.3	36.8	1e3	33.2	21.9	32.5	32.4	33.9
GPTQ + absmax	3e3	415	331	31.6	31.9	31.5	675	596	72.7	33.0	32.8	32.1
GPTQ + OptQuant	3e3	9e3	1e6	31.7	32.3	31.8	353	1e5	185	32.1	31.1	34.0
GPTQ + PiSO	92.5	47.5	32.5	33.1	33.2	34.7	**39.2**	32.7	**15.6**	**34.4**	32.9	37.1
GPTQ ⊙ PiSO	54.7	29.5	22.3	35.5	36.3	40.2	44.7	30.6	17.0	33.7	34.4	34.9
Qronos + absmax	197	68.9	49.5	31.6	32.7	33.5	111	35.8	2e4	32.1	32.5	32.1
Qronos + OptQuant	5e3	1e4	3e5	31.2	31.5	31.6	3e3	657	1e4	31.2	32.6	32.6
Qronos + PiSO	216	84.8	76.0	32.1	33.3	33.0	196	27.4	24.5	32.3	32.9	36.1
Qronos ⊙ PiSO	**27.9**	**18.9**	**16.6**	**37.5**	**41.4**	**44.0**	131	**17.2**	16.3	32.7	**34.8**	**38.1**

(Team et al., 2024). For calibration, we sample 512 sequences of 2048 tokens from the WikiText2 training split (Merity et al., 2017). We focus on weight-only quantization, which, not only is a relevant setting by itself, but also allows us to decouple the evaluation from the choice of any specific activation-quantization scheme. We quantize weights to 2- and 3-bits, where quantization remains most challenging and the choice of PTQ algorithm is critical, and also to 4-bits. All scales are stored in BF16.

Metrics. We report perplexity on the WikiText2 (Merity et al., 2017) validation split, and average zero-shot accuracy on ARC-Easy, ARC-Challenge (Clark et al., 2018), and HellaSwag (Zellers et al., 2019) evaluated with LightEval (Habib et al., 2023). Evaluating LLMs remains an open challenge, as no single benchmark fully captures model quality. We therefore prioritize breadth across model families and sizes, bit-widths, quantization granularities, and PTQ pipelines rather than deep benchmarking for a specific configuration. This design allows us to assess whether PiSO provides consistent improvements across diverse settings. For a subset of models, Appendix E extends task coverage to MMLU (Hendrycks et al., 2021) (general knowledge), IFEval (Zhou et al., 2023) (instruction following), and LongBench (Bai et al., 2024) (long-context question answering).

Baselines. We compare PiSO against two scale-selection strategies, each combined with RTN, GPTQ, and Qronos: (1) *absmax*, which maps the largest-magnitude weight to the extreme grid point; and (2) *OptQuant* (Idelbayev et al., 2021), the exact minimizer of the data-free objective

Table 2: Channel-wise 3-bit and 4-bit quantization results on Llama-3 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The gap between absmax and PiSO-optimized scales widens at lower bit-widths, and PiSO consistently achieves the best results.

	3-bit						**4-bit**					
	WikiText2 (↓)			0-shot (↑)			WikiText2 (↓)			0-shot (↑)		
	1B	3B	8B	1B	3B	8B	1B	3B	8B	1B	3B	8B
BF16	8.97	7.14	5.89	52.6	62.0	68.6	8.97	7.14	5.89	52.6	62.0	68.6
RTN + absmax	2e3	436	1e3	32.8	35.7	38.0	22.3	9.61	7.83	46.3	55.6	64.7
RTN + OptQuant	9e3	6e4	5e4	31.9	32.4	32.2	136	8e3	9e4	33.8	31.8	32.2
RTN + PiSO	19.5	10.6	8.99	44.2	52.1	59.9	11.0	7.92	6.64	50.0	60.4	67.3
AWQ	102	50.1	11.8	38.2	52.5	58.7	11.9	8.02	6.76	50.4	59.8	66.6
Beacon	19.1	33.7	9.35	42.4	40.1	53.5	13.6	25.1	7.53	46.1	46.3	59.7
GPTQ + absmax	19.6	11.3	9.23	39.6	49.8	55.1	10.4	7.79	6.49	49.3	59.4	65.6
GPTQ + OptQuant	546	9e3	8e5	33.1	32.0	31.7	255	9e3	4e5	35.1	31.5	31.7
GPTQ + PiSO	12.9	9.19	7.42	47.5	54.6	**63.0**	9.94	7.63	6.32	**52.0**	60.2	**68.3**
GPTQ ⊙ PiSO	12.8	9.08	7.41	47.0	**55.9**	62.6	9.91	7.65	6.32	51.3	60.0	68.1
Qronos + absmax	15.0	9.83	8.27	40.6	51.4	57.9	10.0	7.64	6.36	50.2	60.1	65.2
Qronos + OptQuant	2e3	9e3	8e5	32.8	31.9	32.3	335	9e3	4e5	34.6	31.8	31.1
Qronos + PiSO	**11.6**	8.42	7.01	**48.6**	55.5	61.4	9.61	**7.44**	6.19	51.5	**61.2**	67.4
Qronos ⊙ PiSO	11.6	**8.41**	**6.98**	48.0	55.6	62.7	**9.59**	7.44	**6.18**	51.1	60.5	67.0

Table 3: Group-wise quantization results on Llama-3 3B and Qwen-2.5 3B with group sizes 16 and 32 (G16/G32), reporting WikiText2 perplexity (↓). The subscripts I/S denote the independent/sequential objectives (Section 3.2); entries without subscript report the best of the two.

	2-bit				**3-bit**				**4-bit**			
	Llama		Qwen		Llama		Qwen		Llama		Qwen	
	G16	G32	G16	G32	G16	G32	G16	G32	G16	G32	G16	G32
BF16	7.14	7.14	7.88	7.88	7.14	7.14	7.88	7.88	7.14	7.14	7.88	7.88
RTN + absmax	8e5	8e5	3e8	2e8	4e3	11.7	21.0	102	7.69	7.72	8.78	8.91
RTN + OptQuant	2e6	2e6	33.0	32.9	16.4	23.3	45.7	43.0	9.81	10.2	51.3	52.0
RTN + PiSO_I	2e4	2e4	388	758	8.21	8.62	8.85	9.13	7.33	7.42	8.08	8.18
RTN + PiSO_S	19.4	28.7	17.1	25.9	7.86	8.17	8.58	8.92	7.26	7.33	8.00	8.04
AWQ	8e5	8e5	2e8	2e8	9.61	9.57	11.4	11.4	7.50	7.55	8.37	8.46
GPTQ + absmax	30.3	26.4	34.8	30.4	8.41	8.29	8.86	8.91	7.35	7.38	8.06	8.03
GPTQ + OptQuant	34.8	47.0	35.8	34.5	10.7	11.1	49.4	48.6	9.25	9.34	52.5	52.7
GPTQ + PiSO	21.5	25.2	15.5	18.1	8.12	8.50	8.55	8.79	7.30	7.37	8.01	8.03
GPTQ ⊙ PiSO	20.7	24.0	15.3	18.1	8.11	8.44	8.52	8.72	7.30	7.36	8.01	8.03
GPTQ ⊡ PiSO	17.0	21.3	14.3	16.6	7.92	8.29	8.48	8.68	7.27	7.34	**7.98**	8.02
Qronos + absmax	22.6	24.5	77.3	21.1	8.07	8.08	8.67	8.66	7.32	7.31	8.03	8.03
Qronos + OptQuant	15.1	18.1	17.7	35.9	7.74	**7.91**	8.43	8.64	7.24	**7.27**	7.99	8.02
Qronos + PiSO	14.6	16.3	13.3	14.7	7.75	7.95	8.44	8.62	7.25	7.29	**7.98**	8.02
Qronos ⊙ PiSO	14.8	16.3	13.0	14.4	7.76	7.96	8.44	8.61	7.25	7.29	7.99	8.03
Qronos ⊡ PiSO	**12.9**	**15.2**	**11.9**	**13.2**	**7.72**	**7.91**	**8.37**	**8.53**	**7.23**	7.28	**7.98**	**8.01**

$\|\mathbf{w} - s\, q(\mathbf{w}; s)\|^2$. It is a stronger version of the common grid-search heuristic, being exact and not restricted to $(0, s_{\text{absmax}}]$. We further compare against two standalone PTQ methods. AWQ (Lin et al., 2024) is a widely used weight-only PTQ method that protects salient channels through activation-aware rescaling before quantization. Beacon (Zhang & Saab, 2026) jointly optimizes channel-wise scales and grid assignments via coordinate descent. We select it as none of the data-aware methods (CDQuant, COMQ, Beacon) of Section 2 that optimize scales has publicly available code, and Beacon is the most recent one and methodologically close to COMQ. Appendix E additionally compares against the gradient-based AutoRound (Cheng et al., 2024). Unless stated otherwise, RTN + PiSO minimizes the cross-activation objective, whereas with GPTQ or Qronos, PiSO minimizes the objective of the error correction method.

Channel-wise Quantization Results. At 2 bits (Table 1), RTN collapses regardless of the scale, although PiSO still lowers its perplexity by two orders of magnitude. With error correction, Qronos $\odot$ PiSO achieves the best results in most configurations, reducing the perplexity of Qronos + absmax by 3–7$\times$ on Llama. Interleaving outperforms the decoupled strategy in 10 of 12 perplexity comparisons. At 3 and 4 bits (Table 2, Qwen in Table 7), PiSO yields the best result in most configurations. Notably, at 3 bits, even RTN + PiSO outperforms GPTQ + absmax on all Llama models. At 4 bits the finer grid narrows the gaps, yet PiSO still improves RTN by 2.6–4.8 accuracy points on Llama. RTN + PiSO already achieves lower perplexity than AWQ, with further improvements achieved by the combination of PiSO with error correction. In contrast, the data-free OptQuant scales often perform even worse than absmax.

Group-wise Quantization Results. Table 3 reports group-wise results on Llama 3B and Qwen 3B with group sizes 16 and 32. Under RTN, the sequential objective outperforms the independent one in every configuration, by orders of magnitude at 2 bits, confirming the importance of cross-group interactions. RTN + PiSO_{S} alone outperforms GPTQ + absmax in most configurations, and AWQ in all of them. With error correction, group-wise interleaving is the best integration strategy in every configuration. Qronos$\boxdot$PiSO achieves the best 2-bit perplexities, improving on Qronos + absmax by up to $6.5\times$. Appendix E reports results for Llama 8B and Qwen 7B and the zero-shot accuracies.

Calibration size sensitivity. The amount of calibration data required by PTQ algorithms is often overlooked. We sweep the number of calibration sequences from 1 to 512 on Llama-3.2-1B with 3-bit channel-wise quantization, calibrating on WikiText2, C4 (Raffel et al., 2020), or FineWeb (Penedo et al., 2024) and always evaluating on WikiText2 (Figure 2 for C4, remaining datasets in Appendix E). RTN+PiSO, which only fits one parameter per channel (precisely the scale), is highly data-efficient: four sequences were enough to reach near-asymptotic perplexity on all three datasets. In contrast, GPTQ + absmax, which has to optimize each grid-to-weight assignment, needs about 64 sequences to stabilize. Somewhat surprisingly, PiSO also makes GPTQ more data-efficient: with 16 sequences, GPTQ$\odot$PiSO outperforms GPTQ+absmax calibrated on 512. This suggests that a well-chosen grid can reduce the dependence of error correction on calibration size.

Additional experiments. Appendix E includes ablations over the objectives supported by PiSO and experiments with the non-integer `E2M1` grid. For a subset of configurations, we extend the evaluation to Qwen 14B and Gemma models, broaden the task coverage, and compare against AutoRound, an established gradient-based PTQ method (Tables 10 and 11). Finally, wall-clock runtimes and overhead relative to GPTQ in Table 4 show that PiSO introduces minimal overhead.

5 Conclusion

We revisited an often overlooked component of PTQ: the choice of the scaling factor. We introduced PiSO, which computes the optimal channel-wise scales under RTN for data-aware layer-wise objectives in $\mathcal{O}(D^2|\mathbb{G}|)$ time per channel. Building on this solver, we proposed two tractable objectives for group-wise quantization, one of which accounts for cross-group interactions, and three principled strategies for integrating PiSO with GPTQ and Qronos. PiSO outperformed absmax, OptQuant, AWQ, and Beacon, with gains widening at lower bit-widths. Moreover, *data-free* optimal scales often failed in the channel-wise setting, confirming the need for a *data-aware* objective. The best results are mostly obtained by combining PiSO with error correction, supporting the effectiveness of the proposed integrations. Finally, PiSO made GPTQ markedly more data-efficient in the evaluated settings.

Limitations and future work. Tightening the theoretical analysis of the integration strategies, and further studying the observed data-efficiency gains are natural next steps. We focused on unconstrained BF16 scales. Hardware formats such as MX (Open Compute Project Foundation, 2024) and NVFP4 (Alvarez et al., 2025) impose additional constraints (e.g., low-bit or hierarchical scales). We plan to extend PiSO to such restricted-scale settings, as well as to asymmetric and weight-activation quantization.

AI USE STATEMENT

Generative AI was used for the following tasks: reviewing our proofs and checking for errors in our theory, assisting in the implementation of parts of our code, generating the scripts that produce plots, tables, and figures, editing sentences of the manuscript to improve readability, including correcting grammar and translation errors, rephrasing and shortening sentences, and spotting typos and broken references in the final manuscript. Generative AI was not used to develop the method, formulate its theoretical results, design the experiments, generate or process datasets, write text from scratch or write full functions in the code. We have reviewed all AI-assisted work: in particular, we manually verified every correction proposed and the code suggestions, revised all AI-edited text, checked the generated scripts and every reported number against the underlying raw result files.

ETHICS STATEMENT

This work proposes an algorithm for post-training quantization and evaluates it on publicly available pre-trained models and public benchmarks, in accordance with their corresponding licenses. Our work does not involve human subjects, newly collected data, or training/releasing new models.

REPRODUCIBILITY STATEMENT

We plan to open-source our code together with the camera-ready version. Algorithm 1 gives the pseudocode of PiSO, and Appendix A discusses the details of its implementation. Appendix C states all theoretical results with their assumptions and complete proofs. Section 4 specifies the models, public HuggingFace checkpoints, calibration and evaluation datasets, calibration sampling, bit-widths, group sizes, and evaluation library. Appendix E describes the quantization library used for the baselines. Given the model and the calibration data, PiSO is deterministic.

REFERENCES

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing NVFP4 for efficient and accurate low-precision inference, June 2025. NVIDIA Developer Blog.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Wenhua Cheng, Weiwei Zhang, Haihao Shen, Yiyang Cai, Xin He, Lv Kaokao, and Yi Liu. Optimize weight rounding via signed gradient descent for the quantization of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11332–11350, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Vage Egiazarian, Roberto L. Castro, Denis Kuznedelev, Andrei Panferov, Eldar Kurtic, Shubhra Pandit, Alexandre Noll Marques, Mark Kurtz, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Bridging the gap between promise and performance for microscaling FP4 quantization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. LightEval: A lightweight framework for llm evaluation, 2023. URL https://github.com/huggingface/lighteval.

Babak Hassibi and Haris Vikalo. On the expected complexity of integer least-squares problems. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–1497. IEEE, 2002.

Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *IEEE international conference on neural networks*, pp. 293–299. IEEE, 1993.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *International conference on machine learning*, pp. 4466–4475. PMLR, 2021.

Yerlan Idelbayev, Pavlo Molchanov, Maying Shen, Hongxu Yin, Miguel A Carreira-Perpinán, and Jose M Alvarez. Optimal quantization using scaled codebook. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12090–12099. IEEE, 2021.

IST-DASLab. GPTQ, 2022. URL https://github.com/ist-daslab/gptq.

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. In *International Conference on Learning Representations*, 2021.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *Forty-second International Conference on Machine Learning*, 2025.

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.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Eric Lybrand and Rayan Saab. A greedy algorithm for quantizing neural networks. *Journal of Machine Learning Research*, 22(156):1–38, 2021.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

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.

Pranav Ajit Nair and Arun Sai Suggala. CDQuant: Accurate post-training weight quantization of large pre-trained models using greedy coordinate descent. *CoRR*, abs/2406.17542, 2024.

Open Compute Project Foundation. OCP microscaling formats (MX) specification v1.0. Technical report, Open Compute Project, 2024. Accessed: 2025-07-15.

Alessandro Pappalardo, Giuseppe Franco, Ian Colbert, Fabian Grob, Timothy Costigan, Oscar Savolainen, Andrei Stoian, Anton Gerdelan, Yaman Umuroglu, Tim Paine, et al. Xilinx/brevitas: Release v0. 12.0, 2025.

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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Sai Sanjeet, Ian Colbert, Pablo Monteagudo-Lago, Giuseppe Franco, Yaman Umuroglu, and Nicholas J Fraser. MixQuant: Pushing the limits of block rotations in post-training quantization. *arXiv preprint arXiv:2601.22347*, 2026.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Di Wu, Qi Tang, Yongle Zhao, Ming Zhang, Ying Fu, and Debing Zhang. EasyQuant: Post-training quantization via scale optimization. *arXiv preprint arXiv:2006.16669*, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Aozhong Zhang, Zi Yang, Naigang Wang, Yingyong Qi, Jack Xin, Xin Li, and Penghang Yin. COMQ: A backpropagation-free algorithm for post-training quantization. *IEEE Access*, 2025a.

Haoyu Zhang, Shihao Zhang, Ian Colbert, and Rayan Saab. Provable post-training quantization: Theoretical analysis of OPTQ and Qronos. *arXiv preprint arXiv:2508.04853*, 2025b.

Jinjie Zhang, Yixuan Zhou, and Rayan Saab. Post-training quantization for neural networks with provable guarantees. *SIAM journal on mathematics of data science*, 5(2):373–399, 2023.

Shihao Zhang and Rayan Saab. Beacon: Post-training quantization with integrated grid selection. *IEEE Signal Processing Letters*, 2026.

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Appendix (Supplementary Materials)

A Implementation Details **15**

B Integration with Greedy Error Correction Algorithms **16**

C Proofs **19**

D Low-Dimensional Example **28**

E Complementary Results **32**

Algorithm 1 PiSO: Piecewise Scale Optimization

Require: Weight vector $\mathbf{w} \in \mathbb{R}^D$, matrices $\mathbf{H} \in \mathbb{R}^{D\times D}$, $\mathbf{G} \in \mathbb{R}^{D\times D}$, sorted grid $\mathbb{G} = \{g_1, \ldots, g_L\}$ s.t. $0 \in \mathbb{G}$
Ensure: Optimal scale $\hat{s}$

Compute transition scales $t_{i,\ell} \leftarrow \frac{2w_i}{g_\ell + g_{\ell+1}}$ for all $i \in [D]$, $\ell \in [L-1]$
Include 0 in the boundary list, keeping only one occurrence of 0 {Sign-change boundary}
Sort all boundaries in decreasing order: $\tau_1 > \cdots > \tau_T$, recording the corresponding (i, ℓ) pairs for each nonzero boundary
Set $\tau_{T+1} \leftarrow -\infty$
$\mathbf{p} \leftarrow \mathbf{G}\mathbf{w}$ {Precompute; does not change during sweep}
$\alpha \leftarrow 0, \quad \beta \leftarrow 0, \quad \mathbf{h} \leftarrow \mathbf{0}$ $\{\alpha = \mathbf{w}^\top \mathbf{G}^\top q, \quad \beta = q^\top \mathbf{H} q, \quad \mathbf{h} = \mathbf{H}q\}$
$\hat{s} \leftarrow$ `undef`, $\quad \hat{\phi} \leftarrow +\infty$
Precompute sign-change transition values $(\alpha^{\mathrm{tr}}, \beta^{\mathrm{tr}}, \mathbf{h}^{\mathrm{tr}})$
for $j = 1, \ldots, T$ **do**
  if $\tau_j = 0$ **then**
    $\alpha \leftarrow \alpha^{\mathrm{tr}}, \quad \beta \leftarrow \beta^{\mathrm{tr}}, \quad \mathbf{h} \leftarrow \mathbf{h}^{\mathrm{tr}}$ {When entering the negative scales sweep}
  else
    Let i be the index whose q_i changes at τ_j by $\delta = q_i^{\mathrm{new}} - q_i^{\mathrm{old}}$
    $\alpha \leftarrow \alpha + \delta \cdot p_i$ $\{\mathcal{O}(1)\}$
    $\beta \leftarrow \beta + 2\delta \cdot h_i + \delta^2 \cdot H_{ii}$ $\{\mathcal{O}(1)\}$
    $\mathbf{h} \leftarrow \mathbf{h} + \delta \cdot \mathbf{H}_{:,i}$ $\{\mathcal{O}(D)$ in the general case; $\mathcal{O}(1)$ if $\mathbf{H}$ is diagonal$\}$
  end if
  if $\beta = 0$ **then**
    Choose any finite $s_j^* \in (\tau_{j+1}, \tau_j)$ {The objective is constant on this interval}
  else
    $s_j^* \leftarrow \mathrm{clamp}\big(\alpha/\beta, \tau_{j+1}, \tau_j - \varepsilon\big)$ **if** $\tau_j > 0$, **else** $\mathrm{clamp}\big(\alpha/\beta, \tau_{j+1} + \varepsilon, \tau_j\big)$
  end if
  $\phi_j \leftarrow (s_j^*)^2\, \beta - 2\, s_j^*\, \alpha$
  if $\phi_j < \hat{\phi}$ **and** $|\tau_j - \tau_{j+1}| > \varepsilon$ **then**
    $\hat{s} \leftarrow s_j^*, \quad \hat{\phi} \leftarrow \phi_j$ {Skip degenerate intervals}
  end if
end for
return $\hat{s}$

A Implementation Details

Algorithm 1 summarizes the optimization procedure described in Section 3 for a single channel. In this implementation the sweep proceeds from large to small s (right to left), processing positive scales first and then negative scales, if permitted. We implemented PiSO using PyTorch 2.8 (Ansel et al., 2024), and benefited from batched operations to parallelize it in GPU across channels in a given layer, and simultaneously across groups for the independent approximation (see Section 3.2).

In the following, we provide additional details on how corner cases arising in the scale optimization are handled.

Degenerate Intervals. Two transition scales coincide ($t_{i,\ell} = t_{j,\ell'}$) whenever $w_i/w_j = m_\ell/m_{\ell'}$ for grid midpoints $m_\ell, m_{\ell'}$. When this happens, the interval between them has zero length. At such a point several entries of $\mathbf{q}$ attempt to change simultaneously, but since our sweep processes them one at a time, the intermediate state does not correspond to any realizable rounding. We detect these degenerate intervals by checking $|\tau_{j-1} - \tau_j| < \varepsilon$ and skip the error update on them.

Handling the scale sign change. For quantization grids that are symmetric around zero, i.e. they verify the property that if $q \in \mathbb{G} \Rightarrow -q \in \mathbb{G}$, only positive scales $s > 0$ need to be considered to guarantee optimality. This holds since $s \cdot q(w; s) = (-s) \cdot q(w; -s)$. However, for asymmetric quantization grids, such as those of integer datatypes, the optimal scale may be negative. If one wants to allow for such negative scales, the algorithm needs to handle the sign change at $s = 0$,

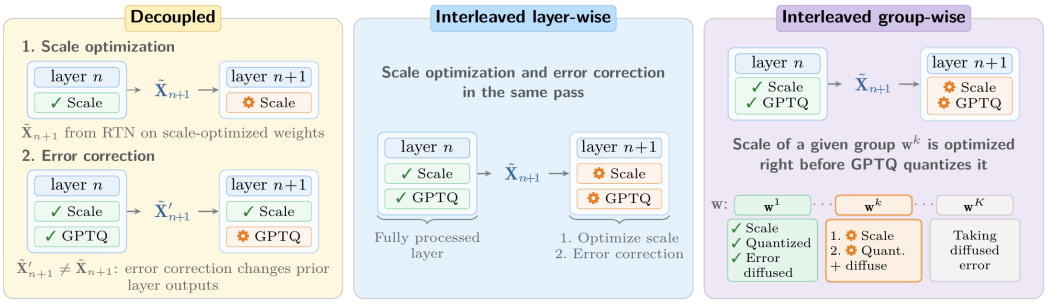


Figure 3: Overview of the three integration strategies of PiSO with error correction algorithms (e.g., GPTQ). The strategies differ in how tightly scale optimization and error correction are interleaved: *decoupled* optimizes all scales in a separate first pass, *interleaved layer-wise* optimizes each layer's scales immediately before its error correction step, and *interleaved group-wise* optimizes each group's scale immediately before error correction quantizes that group.

to account for the simultaneous changes in the unscaled grid assignments. Precisely, when the sweep crosses from positive to negative scales, the ratio w_i/s flips sign for *every* non-zero weight simultaneously. Indeed, apart from degenerate intervals, crossing $s = 0$ is the only point where multiple grid assignments simultaneously change, and incremental updates do not apply. We handle it by precomputing a *transition vector* $\mathbf{q}^{\mathrm{tr}}$ that assigns each weight to its limiting grid value as $s \to 0^-$

$$q_i^{\mathrm{tr}} = \begin{cases} g_L & \text{if } w_i < 0, \\ g_1 & \text{if } w_i > 0, \\ 0 & \text{if } w_i = 0. \end{cases}$$

together with the corresponding α^{tr}, β^{tr}, and $\mathbf{h}^{\mathrm{tr}}$. When the sweep first enters a negative-scale interval, the running state is replaced by these precomputed values, after which incremental updates resume normally for $s < 0$. Under Remark C.1, the assignment is constant on $(a, b]$ between consecutive negative transition scales $a < b < 0$. An excluded left endpoint is approached from inside the interval as $s \to a^+$.

Group processing order. Greedy error correction algorithms typically process the weights of a given layer in descending order of the diagonal entries of $\mathbf{H} := \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ (IST-DASLab, 2022; Zhang et al., 2026; Pappalardo et al., 2025): for a weight vector $\mathbf{w} \in \mathbb{R}^D$, the entry w_i is quantized earlier the larger its associated diagonal element H_{ii}. As mentioned in Section 3.2, in the group-sequential variant of PiSO the order in which the groups are processed matters, since later groups cannot influence earlier ones. We adopt a natural group-wise analogue of the standard weight-level rule: the importance of group k is $\sum_{i \in \mathcal{I}_k^G} H_{ii}$, the sum of the diagonal Hessian entries of its weights, and groups are processed in descending order of this score.

When applying PiSO using the group-wise interleaved integration with error correction described in the *Interleaved group-wise optimization* strategy (Section 3.3), the scale of each group must be optimized *just before* error correction quantizes its weights, which forces all weights of a given group to be processed, before any weight of a different group. The standard $\mathbf{H}$-based ordering used in GPTQ implementations breaks this contiguity. To overcome this we use a *group-aware* reordering: groups are sorted in descending order of their aggregated importance $\sum_{i \in \mathcal{I}_k^G} H_{ii}$ defined above, and within each group columns are sorted in the usual way by H_{ii}. We impose this constraint in the reordering only whenever PiSO is combined with GPTQ at the tightest integration level (GPTQ ⊡ PiSO in the relevant Tables).

B Integration with Greedy Error Correction Algorithms

Greedy error correction algorithms such as GPTQ (Frantar et al., 2022) and Qronos (Zhang et al., 2026) process the D weights sequentially in two alternating steps: (1) error correction, in which the current weight is assigned to the scaled grid value that minimizes the quantization error; (2) and

error diffusion, which adjusts the remaining unquantized to compensate for the induced error. These algorithms assume that the quantization grid has been fixed in advance and, therefore, their weight-to-grid assignments are conditional on the scaling factor that was chosen beforehand. In this regard, the intractable joint optimization of the scaling factor and the weight-to-grid assignments can be approximated in two steps: optimization of the scaling factor for a RTN-surrogate of the weight-to-grid assignment, and error correction to refine the assignments conditionally on the selected grid, as described in Section 3.3.

Following, we provide additional detail into the integration strategies. Let $\tilde{\mathbf{w}}^t$ denote the partially quantized state after step t, with

$$\tilde{\mathbf{w}}^t = \big(\tilde{\mathbf{w}}^t_{\{\leq t\}},\, \mathbf{w}^t_{\{>t\}}\big), \qquad \tilde{\mathbf{w}}^0 = \mathbf{w}, \qquad t \in \{0, \ldots, D\}$$

where $\tilde{\mathbf{w}}^t_{\{\leq t\}} = [s_{\lceil 1/G \rceil} q_1, \ldots, s_{\lceil t/G \rceil} q_t]$ and $\mathbf{w}^t_{\{>t\}} = [w^t_{t+1}, \ldots, w^t_D]$. At step t, error correction selects q_t while holding the other coordinates fixed, and error diffusion updates the remaining weights (Zhang et al., 2026):

$$q_t = \arg\min_{p \in \mathbb{G}} \frac{1}{2} \left\| \mathbf{X}\mathbf{w} - \sum_{j=1}^{t-1} s_{\lceil j/G \rceil} q_j \tilde{\mathbf{X}}_j - s_{\lceil t/G \rceil} p \tilde{\mathbf{X}}_t - \sum_{j=t+1}^{D} w_j^{t-1} \tilde{\mathbf{X}}_j \right\|^2, \tag{13}$$

$$\mathbf{w}^t_{\{>t\}} = \arg\min_{\mathbf{v} \in \mathbb{R}^{D-t}} \frac{1}{2} \left\| \mathbf{X}\mathbf{w} - \sum_{j=1}^{t} s_{\lceil j/G \rceil} q_j \tilde{\mathbf{X}}_j - \sum_{j=t+1}^{D} v_j \tilde{\mathbf{X}}_j \right\|^2. \tag{14}$$

These subproblems admit closed forms: error correction (Equation 13) is a scalar projection onto $\mathbb{G}$, and error diffusion (Equation 14) is a least-squares update over the unquantized weights.

In group-wise quantization, PiSO can be run each time the error correction process reaches a group boundary. Let $t = (k-1)G + 1$ be the first index of group k in the processing order. Before the weights of group k are quantized, PiSO optimizes s_k using the current partially quantized state $\tilde{\mathbf{w}}^{t-1}$. The form of this optimization depends on the group-wise heuristic:

Independent groups interleaved with error correction. Cross-group interactions are discarded and the scale of group k is optimized in isolation, analogously to what is proposed in Equation 11. The key difference is that the weight vector has been modified by error diffusion from preceding groups

$$s_k^* = \arg\min_{s_k} \frac{1}{2} \left\| \mathbf{X}_{\{k\}} \mathbf{w}^{t-1}_{\{k\}} - s_k\, \tilde{\mathbf{X}}_{\{k\}}\, q(\mathbf{w}^{t-1}_{\{k\}};\, s_k) \right\|^2. \tag{15}$$

Sequential groups interleaved with error correction. The scale of group k is optimized under the layer-wise output reconstruction objective, accounting for the quantization error of the preceding groups, analogously to Equation 12. The key difference is that, under interleaving, the preceding groups $i < k$ have been quantized by error correction rather than by round-to-nearest, and the remaining groups $i > k$ have received the error diffusion updates

$$s_k^* = \arg\min_{s_k} \left\| \mathbf{X}\,\mathbf{w} - \tilde{\mathbf{X}} \left[\tilde{\mathbf{w}}^{t-1}_{\{1\}}, \ldots, s_k\, q(\mathbf{w}^{t-1}_{\{k\}}; s_k), \ldots, \mathbf{w}^{t-1}_{\{K\}} \right]^\top \right\|^2. \tag{16}$$

Following, we provide the pseudocode (Algorithm 2) for the three integration strategies of PiSO with error correction algorithms proposed in the main text (Section 3.3). Figure 3 summarizes these graphically.

Algorithm notation. Let $\iota = 1, \ldots, P$ index the layers in processing order. The reference weights $\mathbf{W}_\iota$ and calibration inputs $\mathbf{X}_\iota$ are cached once from the unquantized model, and $\mathcal{M}_\iota$ denotes the model prefix with quantized weights $\tilde{\mathbf{W}}_1, \ldots, \tilde{\mathbf{W}}_\iota$. We write $\tilde{\mathbf{X}}$ for the inputs of the layer currently being processed. Three operators act on this state, all using $\tilde{\mathbf{X}}$ implicitly:

- $\mathsf{A}(\mathcal{M}_{\iota-1})$ returns the calibration inputs reaching layer ι through the prefix $\mathcal{M}_{\iota-1}$.
- $\mathsf{S}(\mathbf{V})$ returns the scales selected by PiSO on the weights $\mathbf{V}$. The group-local call $\mathsf{S}_k(\mathbf{V})$ returns the scale of group k under either the independent objective (Equation 15) or the sequential one (Equation 16).

- $\mathsf{C}(\mathbf{V}; \mathbf{s}_\iota)$ returns the result of GPTQ or Qronos applied to $\mathbf{V}$ at fixed scales. The group-local call $\mathsf{C}_k(\mathbf{V}; \mathbf{s}_{\iota,k})$ quantizes only group k and diffuses its error into the remaining unquantized weights, including later groups; previously quantized weights stay fixed.

Algorithm 2 Three granularities of interleaving scale selection and greedy error correction. S selects scales and C applies error correction for fixed scales.

Shared setup: cache $\mathbf{W}_\iota, \mathbf{X}_\iota$ from the unquantized model. Each panel returns the quantized weights $\tilde{\mathbf{W}}_1, \ldots, \tilde{\mathbf{W}}_P$.

(a) Decoupled: $+$

for $\iota = 1, \ldots, P$ **do**
  $\tilde{\mathbf{X}} \leftarrow \mathsf{A}(\mathcal{M}_{\iota-1})$
  $\mathbf{s}_\iota \leftarrow \mathsf{S}(\mathbf{W}_\iota)$
  $\tilde{\mathbf{W}}_\iota \leftarrow \mathcal{Q}(\mathbf{W}_\iota; \mathbf{s}_\iota)$
end for
for $\iota = 1, \ldots, P$ **do**
  $\tilde{\mathbf{X}} \leftarrow \mathsf{A}(\mathcal{M}_{\iota-1})$
  $\tilde{\mathbf{W}}_\iota \leftarrow \mathsf{C}(\mathbf{W}_\iota; \mathbf{s}_\iota)$
end for

(b) Interleaved layer-wise: $\odot$

for $\iota = 1, \ldots, P$ **do**
  $\tilde{\mathbf{X}} \leftarrow \mathsf{A}(\mathcal{M}_{\iota-1})$
  $\mathbf{s}_\iota \leftarrow \mathsf{S}(\mathbf{W}_\iota)$
  $\tilde{\mathbf{W}}_\iota \leftarrow \mathsf{C}(\mathbf{W}_\iota; \mathbf{s}_\iota)$
end for

(c) Interleaved group-wise: $\boxdot$

for $\iota = 1, \ldots, P$ **do**
  $\tilde{\mathbf{X}} \leftarrow \mathsf{A}(\mathcal{M}_{\iota-1})$
  $\tilde{\mathbf{W}}_\iota \leftarrow \mathbf{W}_\iota$
  for $k = 1, \ldots, K_\iota$ **do**
    $\mathbf{s}_{\iota,k} \leftarrow \mathsf{S}_k(\tilde{\mathbf{W}}_\iota)$
    $\tilde{\mathbf{W}}_\iota \leftarrow \mathsf{C}_k(\tilde{\mathbf{W}}_\iota; \mathbf{s}_{\iota,k})$
  end for
end for

In (a), the RTN-quantized weights from the scale selection loop are discarded prior to error correction. In (c), K_ι is the number of groups per channel in layer ι; groups are processed contiguously in the group-aware order of Appendix A.

C PROOFS

This appendix collects the formal statements and proofs of results referenced in the main text. We begin by introducing the tie-breaking rule for the rounding operator that will be used through out this appendix. Then we derive the optimal functional form of the unscaled assignment q, as a function of the scaling factor $s \in \mathbb{R} \setminus \{0\}$, in the sense of weight mean-squared error minimization:

Remark C.1 (Tie-breaking rule). We assume throughout that $0 \in \mathbb{G}$, as is the case for the standard grids (INT2, ..., INTN, NVFP4, MXFP4, etc.) and for all grids considered in this work. Ties are broken towards zero: writing $\mathcal{N}(y) := \arg\min_{g \in \mathbb{G}} |g - y|$ for the set of grid values nearest to $y \in \mathbb{R}$, we define

$$\lfloor y \rceil_{\mathbb{G}} := \arg\min_{g \in \mathcal{N}(y)} |g|.$$

Note that $|\mathcal{N}(y)| = 2$ if and only if $y = m_\ell := (g_\ell + g_{\ell+1})/2$ for some $\ell \in |\mathbb{G}|$.

Proposition C.2 (Weight mean-squared error minimization is achieved by RTN and scale optimization)**.** *Let $\mathbf{w} \in \mathbb{R}^D$ denote a weight vector, and assume that the scaling factor $s \in \mathbb{R} \setminus \{0\}$ is shared for all the weights in the vector. Then, the* data-free *optimal assignment to the grid $\mathbf{q}$ can be recovered by minimizing*[3]

$$\min_{\substack{s \in \mathbb{R}\setminus\{0\} \\ \mathbf{q} \in \mathbb{G}^D}} \|\mathbf{w} - s\,\mathbf{q}\|^2 = \min_{s \in \mathbb{R}\setminus\{0\}} \|\mathbf{w} - s\,q(\mathbf{w}; s)\|^2, \quad q(\mathbf{w}; s) := \left\lfloor \frac{\mathbf{w}}{s} \right\rceil_{\mathbb{G}}.$$

Proof. First, note that, for fixed $s \in \mathbb{R} \setminus \{0\}$, the objective $\min_{\mathbf{q} \in \mathbb{G}^D} \|\mathbf{w} - s\,\mathbf{q}\|^2$ is separable in each individual weight

$$\|\mathbf{w} - s\,\mathbf{q}\|^2 = \sum_{i=1}^{D} |w_i - s\,q_i|^2.$$

For each variable, since $s \neq 0$, we have

$$|w_i - s q_i|^2 = s^2 |w_i/s - q_i|^2.$$

Since $s^2 > 0$, minimizing this expression over $q_i \in \mathbb{G}$ amounts to choosing a grid value nearest to w_i/s. Thus, a minimizing assignment is

$$q_i^* = \lfloor w_i/s \rceil_{\mathbb{G}},$$

with ties resolved according to Remark C.1. Consequently, the optimal assignment of the weights to the discrete grid can be recovered by optimizing the scaling factor for the objective:

$$\min_{s \in \mathbb{R}\setminus\{0\}} \|\mathbf{w} - s\,q(\mathbf{w}; s)\|^2, \quad q(\mathbf{w}; s) := \left\lfloor \frac{\mathbf{w}}{s} \right\rceil_{\mathbb{G}}.$$

□

The preceding result justifies the use of the round-to-nearest quantizer for fixed scale s. However, once calibration data enters the picture, the situation changes:

Remark C.3. Proposition C.2 does not extend to the data-aware objective

$$\min_{\substack{s \in \mathbb{R}\setminus\{0\} \\ \mathbf{q} \in \mathbb{G}^D}} \left\| \mathbf{X}\mathbf{w} - s\,\tilde{\mathbf{X}}\mathbf{q} \right\|^2,$$

which corresponds to an integer least-squares problem, a class of problems that is NP-hard (Hassibi & Vikalo, 2002). Precisely, when the scaling factor $s \in \mathbb{R} \setminus \{0\}$ is fixed, the optimal solution $\mathbf{q}^*$ is not necessarily $q(\mathbf{w}; s) := \lfloor \mathbf{w}/s \rceil_{\mathbb{G}}$.

[3] We do not claim novelty for this result, although we could not find it explicitly stated as such in prior literature. We include the proof here for completeness.

We next analyze the expected layer-wise reconstruction error. Although Remark C.3 shows that RTN is not optimal for the data-aware objective in general, the following decomposition reveals that the expected error on unseen data is bounded by a term involving the *parameter mismatch* $\|\mathbf{w} - \mathbf{q}_\mathbf{X}\|^2$. Thus, for a given scale, this quantity is minimized by the RTN quantizer (Proposition C.2). Consequently, restricting the grid assignments to those produced by round-to-nearest controls one of the three factors in the generalization bound and provides a principled prior on the solution space of the data-aware reconstruction objective.

Remark C.4 (Layer-wise error decomposition). Let $\mathbf{z} \sim \mathcal{X}$ be a random activation from the test distribution, and let $\tilde{\mathbf{z}}$ denote the activation obtained after passing through previously quantized layers. Let $\mathbf{X}$ be a calibration dataset, and suppose $\mathbf{q}_\mathbf{X}$ denotes the reconstructed quantized weight vector, including its scaling factors, calibrated on $\mathbf{X}$. Then, for an unseen random activation $\mathbf{z}$, the expected layer-wise error can be decomposed as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\mathcal{X}}\left[\left\|\mathbf{z}^\top\mathbf{w} - \tilde{\mathbf{z}}^\top\mathbf{q}_\mathbf{X}\right\|^2\right] = \mathbb{E}_{\mathbf{z}\sim\mathcal{X}}\Big[&\underbrace{\left\|\mathbf{z}^\top(\mathbf{w} - \mathbf{q}_\mathbf{X})\right\|^2}_{(1)} \\ &+ \underbrace{2\left\langle \mathbf{z}^\top(\mathbf{w} - \mathbf{q}_\mathbf{X}), (\mathbf{z} - \tilde{\mathbf{z}})^\top\mathbf{q}_\mathbf{X}\right\rangle}_{(2)} \\ &+ \underbrace{\left\|(\mathbf{z} - \tilde{\mathbf{z}})^\top\mathbf{q}_\mathbf{X}\right\|^2}_{(3)}\Big],\end{aligned}$$

Assuming $\tilde{\mathbf{z}} \approx \mathbf{z}$, terms (2) and (3) can be neglected, and, from (Zhang et al., 2025b, Remark 3.8), term (1) can be rewritten and bounded as

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\mathcal{X}}\left[\left\|\mathbf{z}^\top(\mathbf{w} - \mathbf{q}_\mathbf{X})\right\|^2\right] &= \frac{1}{N}\|\mathbf{X}(\mathbf{w} - \mathbf{q}_\mathbf{X})\|^2 + (\mathbf{w} - \mathbf{q}_\mathbf{X})^\top\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{X}}\left[\mathbf{z}\mathbf{z}^\top\right] - \frac{1}{N}\mathbf{X}^\top\mathbf{X}\right](\mathbf{w} - \mathbf{q}_\mathbf{X}) \\ &\leq \underbrace{\frac{1}{N}\|\mathbf{X}(\mathbf{w} - \mathbf{q}_\mathbf{X})\|^2}_{\text{(1) calibration fit}} + \underbrace{\left\|\mathbb{E}_{\mathbf{z}\sim\mathcal{X}}\left[\mathbf{z}\mathbf{z}^\top\right] - \frac{1}{N}\mathbf{X}^\top\mathbf{X}\right\|_{\text{op}}}_{\text{(2) covariance deviation}} \cdot \underbrace{\left\|\mathbf{w} - \mathbf{q}_\mathbf{X}\right\|^2}_{\text{(3) parameter mismatch}}.\end{aligned}$$

The bound highlights three factors governing generalization error (Zhang et al., 2025b): (1) calibration fit on $\mathbf{X}$, (2) second-moment estimation error; and (3) proximity between $\mathbf{w}$ and $\mathbf{q}_\mathbf{X}$.

The following sequence of results (Lemma C.5, Lemma C.6, Lemma C.8, Corollary C.9 and Proposition C.11) establishes the theoretical foundation for the algorithm presented in Section 3. The resulting sweep identifies the global infimum of the channel-wise objective and returns an admissible scale arbitrarily close to it; in the data-free setting, the global minimum is attained exactly.

Lemma C.5 (Channel-wise separability)**.** *Let* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$*, and assign an independent scale* s_j *to each column* $\mathbf{w}_j$*. Then the objective*

$$E(\mathbf{s}) = \left\|\mathbf{X}\mathbf{W} - \widetilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{W}; \mathbf{s})\right\|_F^2$$

decomposes as

$$E(\mathbf{s}) = \sum_{j=1}^{M}\left\|\mathbf{X}\mathbf{w}_j - \widetilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{w}_j; s_j)\right\|^2.$$

Consequently, each of the channel-wise scales $s_j \in \mathbb{R} \setminus \{0\}$ *can be optimized independently.*

Proof. Writing $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$ and assigning an independent scale s_j to each column, the matrix products decompose column-wise as

$$\mathbf{X}\mathbf{W} = [\mathbf{X}\mathbf{w}_1, \ldots, \mathbf{X}\mathbf{w}_M], \qquad \tilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{W}; \mathbf{s}) = [\tilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{w}_1; s_1), \ldots, \tilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{w}_M; s_M)].$$

Since $\|\mathbf{A}\|_F^2 = \sum_{j=1}^{M}\|\mathbf{a}_j\|^2$, the objective in Equation 2 decomposes as

$$E(\mathbf{s}) = \sum_{j=1}^{M}\left\|\mathbf{X}\mathbf{w}_j - \tilde{\mathbf{X}}\,\mathcal{Q}(\mathbf{w}_j; s_j)\right\|^2. \tag{17}$$

Hence, the objective in Equation 17 separates into M independent channel-wise problems, thus each s_j can be optimized independently. □

Lemma C.6 (Closed-form for grid assignments)**.** *The following result is standard and appears, e.g., as in (Zhang & Saab, 2026, Proposition 2.1); we restate it here for completeness. For fixed $\mathbf{q} \in \mathbb{G}^D$ with $\tilde{\mathbf{X}}\,\mathbf{q} \neq \mathbf{0}$, the objective*

$$E(s) = \|\mathbf{X}\mathbf{w} - s\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2 \tag{18}$$

is a strictly convex quadratic in s, uniquely minimized at

$$s^* = \frac{\langle \tilde{\mathbf{X}}\mathbf{q}, \mathbf{X}\mathbf{w}\rangle}{\|\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2}.$$

Proof. Expanding the squared norm:

$$E(s) = \|\mathbf{X}\mathbf{w}\|^2 - 2s\,\langle \mathbf{X}\mathbf{w}, \tilde{\mathbf{X}}\,\mathbf{q}\rangle + s^2\|\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2.$$

Since $\|\tilde{\mathbf{X}}\,\mathbf{q}\|^2 > 0$, this is a strictly convex quadratic in s. Setting the derivative to zero:

$$\begin{aligned}\frac{dE}{ds}(s^*) = 0 \iff -2\,\langle \mathbf{X}\mathbf{w}, \tilde{\mathbf{X}}\,\mathbf{q}\rangle + 2s^*\|\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2 = 0 \\ \iff s^* = \frac{\langle \tilde{\mathbf{X}}\mathbf{q}, \mathbf{X}\mathbf{w}\rangle}{\|\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2},\end{aligned}$$

which is the unique global minimizer. □

Remark C.7 (Connection to α/β notation). The closed-form minimizer $s^* = \langle \tilde{\mathbf{X}}\mathbf{q}, \mathbf{X}\mathbf{w}\rangle / \|\tilde{\mathbf{X}}\mathbf{q}\|^2$ can be expressed in terms of the matrices $\mathbf{H} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ and $\mathbf{G} = \tilde{\mathbf{X}}^\top\mathbf{X}$ defined in Equation 6. Indeed,

$$\alpha(\mathbf{q}) := \langle \tilde{\mathbf{X}}\mathbf{q}, \mathbf{X}\mathbf{w}\rangle = (\tilde{\mathbf{X}}\mathbf{q})^\top(\mathbf{X}\mathbf{w}) = \mathbf{q}^\top\tilde{\mathbf{X}}^\top\mathbf{X}\,\mathbf{w} = \mathbf{q}^\top\mathbf{G}\,\mathbf{w},$$

and

$$\beta(\mathbf{q}) := \|\tilde{\mathbf{X}}\mathbf{q}\|^2 = (\tilde{\mathbf{X}}\mathbf{q})^\top(\tilde{\mathbf{X}}\mathbf{q}) = \mathbf{q}^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\,\mathbf{q} = \mathbf{q}^\top\mathbf{H}\,\mathbf{q},$$

recovering the definitions in Equation 6 and the closed form $s^*(\mathbf{q}) = \alpha(\mathbf{q})/\beta(\mathbf{q})$ in Equation 7. Similarly, Equation 18 can be expressed in terms of α/β

$$E(s \mid \mathbf{q}) = \|\mathbf{X}\mathbf{w}\|^2 - 2s\,\langle \mathbf{X}\mathbf{w}, \tilde{\mathbf{X}}\,\mathbf{q}\rangle + s^2\|\,\tilde{\mathbf{X}}\,\mathbf{q}\|^2 = \text{const} - 2s\,\alpha(\mathbf{q}) + s^2\,\beta(\mathbf{q}).$$

Lemma C.8 (Decision regions of round-to-nearest)**.** *Let $\mathbb{G} = \{g_1 < g_2 < \cdots < g_L\} \subset \mathbb{R}$ be a finite grid such that $0 \in \mathbb{G}$ and define*

$$m_\ell := \frac{g_\ell + g_{\ell+1}}{2}, \qquad \ell = 1, \ldots, L-1.$$

Then for a value $y \notin \{m_\ell\}$, g_ℓ is the unique nearest grid value whenever

$$y \in (m_{\ell-1}, m_\ell),$$

with $\ell \in \{1, \ldots, L\}$, where $m_0 = -\infty$ and $m_L = +\infty$. At a transition point $y = m_\ell$, the two nearest grid values are g_ℓ and $g_{\ell+1}$, and the tie-breaking rule of Remark C.1 selects

$$\lfloor m_\ell \rceil_{\mathbb{G}} = \begin{cases} g_\ell, & m_\ell > 0, \\ g_{\ell+1}, & m_\ell < 0. \end{cases}$$

Proof. Fix $\ell \in \{2, \ldots, L-1\}$. We first compare the distances from y to the neighboring grid points. Since $g_{\ell-1} < g_\ell$, we have

$$\begin{aligned}|y - g_\ell| < |y - g_{\ell-1}| &\iff y - g_\ell > -(y - g_{\ell-1}) \qquad (\text{since } g_{\ell-1} < g_\ell) \\ &\iff 2y > g_\ell + g_{\ell-1} \\ &\iff y > \frac{g_{\ell-1} + g_\ell}{2} = m_{\ell-1}.\end{aligned}$$

Similarly,

$$\begin{aligned} |y - g_\ell| < |y - g_{\ell+1}| &\iff y - g_\ell < -(y - g_{\ell+1}) \qquad (\text{since } g_\ell < g_{\ell+1}) \\ &\iff 2y < g_\ell + g_{\ell+1} \\ &\iff y < \frac{g_\ell + g_{\ell+1}}{2} = m_\ell. \end{aligned}$$

Therefore,

$$y \in (m_{\ell-1}, m_\ell) \iff \big(|y - g_\ell| < |y - g_{\ell-1}|\big) \wedge \big(|y - g_\ell| < |y - g_{\ell+1}|\big). \qquad (\star)$$

Now let $y \in (m_{\ell-1}, m_\ell)$. For any $j \le \ell - 1$, since $g_j \le g_{\ell-1} < y$, we have

$$|y - g_j| = y - g_j \ge y - g_{\ell-1} > |y - g_\ell|.$$

For any $j \ge \ell + 1$, since $y < g_{\ell+1} \le g_j$, we have

$$|y - g_j| = g_j - y \ge g_{\ell+1} - y > |y - g_\ell|.$$

Hence, g_ℓ uniquely minimizes $|y - g|$ over $g \in \mathbb{G}$, and thus

$$\lfloor y \rceil_{\mathbb{G}} = g_\ell.$$

The endpoint cases $\ell = 1$ and $\ell = L$ follow analogously, using only the existing neighboring grid point.

It remains to consider the transition points. Let $y = m_\ell$. By the definition of the midpoint,

$$|m_\ell - g_\ell| = |m_\ell - g_{\ell+1}|.$$

Moreover, for every $j \le \ell - 1$,

$$|m_\ell - g_j| > |m_\ell - g_\ell|,$$

and, for every $j \ge \ell + 2$,

$$|m_\ell - g_j| > |m_\ell - g_{\ell+1}|.$$

Therefore, the two nearest grid values to m_ℓ are exactly g_ℓ and $g_{\ell+1}$.

Since $0 \in \mathbb{G}$ and g_ℓ and $g_{\ell+1}$ are adjacent grid values, $m_\ell \neq 0$. If $m_\ell > 0$, then $g_\ell + g_{\ell+1} > 0$, and

$$|g_{\ell+1}|^2 - |g_\ell|^2 = (g_{\ell+1} - g_\ell)(g_{\ell+1} + g_\ell) > 0.$$

Thus, $|g_\ell| < |g_{\ell+1}|$, and the tie-breaking rule of Remark C.1 selects g_ℓ. Similarly, if $m_\ell < 0$, then

$$|g_{\ell+1}|^2 - |g_\ell|^2 = (g_{\ell+1} - g_\ell)(g_{\ell+1} + g_\ell) < 0.$$

Thus, $|g_{\ell+1}| < |g_\ell|$, and the tie-breaking rule selects $g_{\ell+1}$. Consequently,

$$\lfloor m_\ell \rceil_{\mathbb{G}} = \begin{cases} g_\ell, & \text{if } m_\ell > 0, \\ g_{\ell+1}, & \text{if } m_\ell < 0. \end{cases}$$

Thus, the rounding assignment is constant between consecutive midpoints, while its value at each midpoint is determined by the tie-breaking rule of Remark C.1. □

Corollary C.9 (Step-wise structure)**.** *Let $\mathbb{G} = \{g_1 < \cdots < g_L\}$ be a finite grid, such that $0 \in \mathbb{G}$, with midpoints $m_\ell = (g_\ell + g_{\ell+1})/2$, with $\ell \in \{1, \ldots, L-1\}$. Set $m_0 = -\infty$ and $m_L = +\infty$. For any $w \in \mathbb{R}$ and $s \in \mathbb{R}^+$, define $q(s) := \lfloor w/s \rceil_{\mathbb{G}}$. Whenever $w/s \notin \{m_1, \ldots, m_{L-1}\}$, for every $\ell \in \{1, \ldots, L\}$,*

$$q(s) = g_\ell \quad \iff \quad \frac{w}{s} \in (m_{\ell-1}, m_\ell).$$

At a midpoint, the assignment is determined by Remark C.1. Consequently, $q(s)$ is piecewise constant on $(0, \infty)$, with jumps occurring only at the positive transition scales of the form

$$t_\ell := \frac{w}{m_\ell}.$$

Moreover, q is right-continuous on $(0, \infty)$: at each such transition,

$$q(t_\ell) = \lim_{\varepsilon \to 0^+} q(t_\ell + \varepsilon).$$

Proof. If $w = 0$, then $q(s) = 0$ for every $s > 0$, so the result holds trivially. For $w \neq 0$, the map $s \mapsto w/s$ is continuous and strictly monotone on $(0, \infty)$. By Lemma C.8, away from the midpoints, $\lfloor y \rceil_{\mathbb{G}} = g_\ell$ if and only if $y \in (m_{\ell-1}, m_\ell)$. Substituting $y = w/s$ yields the stated equivalence.

The assignment can change only when w/s crosses a midpoint, that is, at a positive solution of $w/s = m_\ell$, given by $t_\ell = w/m_\ell$. Thus, $q(s)$ is constant between consecutive positive transition scales. Fix a transition $t_\ell > 0$, so that $w/t_\ell = m_\ell$. For $s > t_\ell$, the ratio w/s has the same sign as m_ℓ (since $s > 0$ and $t_\ell = w/m_\ell > 0$) and satisfies $|w/s| = |w|/s < |w|/t_\ell = |m_\ell|$, hence w/s lies strictly between 0 and m_ℓ. By continuity of $s \mapsto w/s$, there is $\delta > 0$ such that, for all $s \in (t_\ell, t_\ell + \delta)$, $w/s \in (m_{\ell-1}, m_\ell)$ if $m_\ell > 0$, and $w/s \in (m_\ell, m_{\ell+1})$ if $m_\ell < 0$. Therefore, $q(s) = g_\ell$ if $m_\ell > 0$, and $q(s) = g_{\ell+1}$ if $m_\ell < 0$, which by Lemma C.8 equals $q(t_\ell) = \lfloor m_\ell \rceil_{\mathbb{G}}$. Thus q is constant on $[t_\ell, t_\ell + \delta)$, proving the stated limit. □

Any alternative deterministic tie-breaking rule would preserve the piecewise-constant structure of $q(\cdot)$, but would change which of the two adjacent intervals each transition scale t_ℓ belongs to, and hence the endpoint conventions used when sweeping them.

Corollary C.10 (Piecewise-constant structure for the full weight vector)**.** *Let $\mathbb{G} = \{g_1 < \cdots < g_L\}$ be a finite grid such that $0 \in \mathbb{G}$ with midpoints $m_\ell = (g_\ell + g_{\ell+1})/2$, and let $\mathbf{w} \in \mathbb{R}^D$. For each non-zero weight w_i, Corollary C.9 provides at most $L - 1$ transition scales $t_{i,\ell} = w_i/m_\ell$, $\ell = 1, \ldots, L-1$. Collect all T positive transition scales ($T \leq D(L-1)$) and, assuming they are distinct (the degenerate case is treated in Appendix A), sort them in decreasing order*

$$\tau_1 > \tau_2 > \cdots > \tau_T.$$

If $T = 0$ (for example, if all the weights are 0), there are no positive transition scales and the assignment is constant on $(0, \infty)$. Otherwise, define the half-open intervals

$$J_0 := [\tau_1, +\infty), \qquad J_j := [\tau_{j+1}, \tau_j) \ \ (j = 1, \ldots, T-1), \qquad J_T := (0, \tau_T).$$

Then the vector-valued assignment $\mathbf{q}(s) := \lfloor \mathbf{w}/s \rceil_{\mathbb{G}}$ is constant on each J_j, $j = 0, \ldots, T$. Moreover, at each transition τ_j exactly one component of $\mathbf{q}$ changes: if $\tau_j = t_{i,\ell}$, then only q_i jumps (from g_ℓ to $g_{\ell+1}$ or vice versa) while all other entries remain fixed.

Proof. By Corollary C.9, each component $q_i(s) = \lfloor w_i/s \rceil_{\mathbb{G}}$ is piecewise constant on $(0, \infty)$ with changes happening only at the positive transition scales $\{t_{i,\ell}\}_{\ell=1}^{L-1}$. If $w_i = 0$, its assignment is constantly zero. The full vector $\mathbf{q}(s)$ can therefore change only when *some* component q_i changes, which occurs only at one of the identifies transition scales. Between consecutive transition scales, no component crosses a midpoint, so every $q_i(s)$ is constant, and hence $\mathbf{q}(s)$ is constant on J_j. At each positive transition, Corollary C.9 gives

$$\mathbf{q}(\tau_j) = \lim_{\varepsilon \to 0^+} \mathbf{q}(\tau_j + \varepsilon).$$

Thus, the transition belongs to the interval on its right, establishing constancy on the stated intervals, including their closed endpoints.

Since all transition scales are assumed distinct, each τ_j corresponds to a unique pair (i, ℓ). Only q_i crosses a midpoint there, changing between g_ℓ and $g_{\ell+1}$; all other components remain unchanged. If there are no positive transitions, every component is constant throughout $(0, \infty)$. □

Proposition C.11 (Global infimum of the piecewise-quadratic sweep)**.** *Under the setting and notation of Corollary C.10, assume $T \geq 1$ and let $\mathbf{H} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \succeq 0$ and $\mathbf{G} = \tilde{\mathbf{X}}^\top \mathbf{X}$, and consider the channel-wise error*

$$E(s) := \|\mathbf{X}\mathbf{w} - s\,\tilde{\mathbf{X}}\,q(\mathbf{w};\, s)\|^2, \qquad s > 0.$$

Write $\mathbf{q}_j$ for the constant value of $\mathbf{q}(s)$ on J_j, and define

$$\alpha_j := \mathbf{q}_j^\top \mathbf{G}\,\mathbf{w}, \qquad \beta_j := \mathbf{q}_j^\top \mathbf{H}\,\mathbf{q}_j, \qquad \varphi_j(s) := s^2\beta_j - 2s\,\alpha_j + c,$$

where $c := \|\mathbf{X}\mathbf{w}\|^2$ is a constant independent of s. Then:

1. *φ_j is a convex quadratic that coincides with E on J_j, i.e., $E(s) = \varphi_j(s)$ for all $s \in J_j$;*

2. *For each interval J_j,*
$$\inf_{s\in J_j} E(s) = \min_{s\in \overline{J_j}} \varphi_j(s) = \varphi_j(s_j^*).$$
If $\beta_j > 0$, define
$$s_j^* := \operatorname{clamp}\left(\frac{\alpha_j}{\beta_j}, \inf J_j, \sup J_j\right).$$
If $\beta_j = 0$, then $\alpha_j = 0$ and $\varphi_j(s) = c$ for every s; in this case, choose any $s_j^* \in J_j$.

3. $\inf_{s>0} E(s) = \min_{j\in\{0,\dots,T\}} \varphi_j(s_j^*)$. *If there exists an index j^* attaining the outer minimum such that $s_{j^*}^* \in J_{j^*}$, then the infimum is attained exactly. Otherwise, for every $\eta > 0$, there exists an admissible scale $\hat{s} > 0$ such that*
$$E(\hat{s}) - \inf_{s>0} E(s) < \eta.$$
In the data-free setting $\mathbf{X} = \tilde{\mathbf{X}} = \mathbf{I}$, the infimum is always attained.

Proof. Part (i). On J_j the assignment $\mathbf{q}(s) \equiv \mathbf{q}_j$ is constant by Corollary C.10. Expanding the squared norm gives
$$E(s) = \|\mathbf{X}\mathbf{w}\|^2 - 2s\,\mathbf{q}_j^\top \mathbf{G}\,\mathbf{w} + s^2\,\mathbf{q}_j^\top \mathbf{H}\,\mathbf{q}_j = \varphi_j(s).$$
Since $\mathbf{H} \succeq 0$, we have $\beta_j = \mathbf{q}_j^\top \mathbf{H}\,\mathbf{q}_j = \|\tilde{\mathbf{X}}\,\mathbf{q}_j\|^2 \geq 0$, so φ_j is convex.

Part (ii). If $\beta_j = 0$, then $\|\tilde{\mathbf{X}}\mathbf{q}_j\|^2 = 0$, and hence
$$\alpha_j = \langle \tilde{\mathbf{X}}\mathbf{q}_j, \mathbf{X}\mathbf{w}\rangle = 0.$$
Therefore, $\varphi_j(s) = c$ for every s, and any $s_j^* \in J_j$ attains the interval minimum. In particular, this covers the upper interval $J_0 = [\tau_1, +\infty)$, where $\mathbf{q}_0 = \mathbf{0}$.

Now suppose $\beta_j > 0$. The quadratic φ_j has unconstrained minimizer α_j/β_j. If we temporarily include both endpoints of J_j, its minimum is attained by clamping this value:
$$s_j^* = \operatorname{clamp}\left(\frac{\alpha_j}{\beta_j}, \inf J_j, \sup J_j\right).$$
Thus, for every $s \in J_j$,
$$E(s) = \varphi_j(s) \geq \varphi_j(s_j^*).$$
If $s_j^* \in J_j$, equality is attained. Otherwise, s_j^* is an excluded endpoint. We can approach that endpoint using scales inside J_j. Since φ_j is a polynomial, their errors satisfy
$$\lim_{\substack{s\to s_j^* \\ s\in J_j}} E(s) = \lim_{\substack{s\to s_j^* \\ s\in J_j}} \varphi_j(s) = \varphi_j(s_j^*).$$
Consequently,
$$\inf_{s\in J_j} E(s) = \min_{s\in \overline{J_j}} \varphi_j(s) = \varphi_j(s_j^*).$$

For the interval $J_T = (0, \tau_T)$, a candidate scale at zero is interpreted through the limit $s \to 0^+$, as the quantizer is not defined for a scaling factor of 0. Similarly, a candidate scale that lies at an excluded boundary is evaluated through the limit from inside its interval which can be different from the actual objective value at that transition point.

Part (iii). The half-open intervals $\{J_j\}_{j=0}^T$ are pairwise disjoint and their union is $(0, +\infty)$. Therefore
$$\inf_{s>0} E(s) = \min_{j\in\{0,\dots,T\}} \inf_{s\in J_j} E(s) = \min_j \varphi_j(s_j^*),$$
where the second equality uses Part (ii).

Let j^* be an index attaining the outer minimum. If $s_{j^*}^* \in J_{j^*}$, then
$$E(s_{j^*}^*) = \varphi_{j^*}(s_{j^*}^*) = \inf_{s>0} E(s),$$

and the infimum is attained exactly.

In the degenerate case where $s^*_{j^*} = 0$, the global infimum equals $\varphi_{j^*}(0) = c$. This value is also attained at every $s \geq \tau_1$, where all weights round to zero. Thus, a positive scale attains the global infimum in this case as well. It remains to consider the case in which the minimizing candidate is the excluded right endpoint of its interval,

$$s^*_{j^*} = \tau_{j^*} \notin J_{j^*}.$$

For any

$$0 < \varepsilon < \min\{\tau_{j^*} - \inf J_{j^*},\, 1\},$$

define

$$s_\varepsilon := \tau_{j^*} - \varepsilon.$$

Then $s_\varepsilon \in J_{j^*}$ and therefore $E(s_\varepsilon) = \varphi_{j^*}(s_\varepsilon)$. Since

$$\inf_{s>0} E(s) = \varphi_{j^*}(\tau_{j^*}),$$

the sub-optimality gap is

$$\begin{aligned} E(s_\varepsilon) - \inf_{s>0} E(s) &= \varphi_{j^*}(\tau_{j^*} - \varepsilon) - \varphi_{j^*}(\tau_{j^*}) \\ &= 2\varepsilon\big(\alpha_{j^*} - \beta_{j^*}\tau_{j^*}\big) + \beta_{j^*}\varepsilon^2. \end{aligned}$$

Consequently,

$$E(s_\varepsilon) - \inf_{s>0} E(s) = \mathcal{O}(\varepsilon),$$

and the gap converges to zero as $\varepsilon \to 0^+$. Hence, for every $\eta > 0$, one can choose ε sufficiently small so that

$$E(s_\varepsilon) - \inf_{s>0} E(s) < \eta.$$

Finally, consider the data-free setting $\mathbf{X} = \tilde{\mathbf{X}} = \mathbf{I}$. At a transition scale $\tau = t_{i,\ell}$, we have

$$\frac{w_i}{\tau} = m_\ell = \frac{g_\ell + g_{\ell+1}}{2}.$$

It follows that

$$(w_i - \tau g_\ell)^2 = (w_i - \tau g_{\ell+1})^2.$$

All other grid assignments remain unchanged at this transition. Therefore, the data-free error

$$E(s) = \|\mathbf{w} - s\, q(\mathbf{w}; s)\|^2$$

is continuous at every transition scale. In particular, when an interval minimum lies at an otherwise excluded transition point,

$$E(\tau_{j^*}) = \varphi_{j^*}(\tau_{j^*}).$$

Thus, in the data-free setting, the global infimum is attained exactly. □

Proposition C.11 guarantees that the global infimum is identified by enumerating all $T + 1 \leq D(|\mathbb{G}|-1) + 1$ intervals, but a naive implementation that recomputes α, β from scratch at each transition would cost $\mathcal{O}(D^2)$ per interval, yielding $\mathcal{O}(D^3|\mathbb{G}|)$ per channel. The following result shows that, because exactly one component of $\mathbf{q}$ changes at each transition (Corollary C.10), the quantities α, β, and the auxiliary vector $\mathbf{h} = \mathbf{H}\mathbf{q}$ can be maintained incrementally: α and β update in $\mathcal{O}(1)$, while $\mathbf{h}$ updates in $\mathcal{O}(D)$, reducing the channel-wise cost of the sweep to $\mathcal{O}(D^2|\mathbb{G}|)$ with a dense Hessian and to $\mathcal{O}(D|\mathbb{G}|)$ when $\mathbf{H}$ is diagonal.

Proposition C.12 (Correctness of incremental updates)**.** *Define*

$$\begin{aligned} \alpha &:= \alpha(\mathbf{q}) = \mathbf{q}^\top \mathbf{G}\, \mathbf{w}, \\ \mathbf{h} &:= \mathbf{h}(\mathbf{q}) = \mathbf{H}\, \mathbf{q}, \\ \beta &:= \beta(\mathbf{q}) = \mathbf{q}^\top \mathbf{h}(\mathbf{q}), \end{aligned}$$

where $\mathbf{q} \in \mathbb{G}^D$ *is the current grid assignment. Suppose entry* i *of* $\mathbf{q}$ *changes from* q_i^{old} *to* $q_i^{\text{new}} = q_i^{\text{old}} + \delta$*, while all other entries remain unchanged. Writing* $\mathbf{q}' = \mathbf{q} + \delta\,\mathbf{e}_i$ *for the updated assignment (where* $\mathbf{e}_i$ *is the* i*-th standard basis vector), the updated quantities*

$$\begin{aligned}\alpha' &= \alpha + \delta \cdot (\mathbf{G}\mathbf{w})_i,\\ \mathbf{h}' &= \mathbf{h} + \delta \cdot \mathbf{H}_{:,i}\\ \beta' &= \beta + 2\delta \cdot h_i + \delta^2 \cdot H_{ii},\end{aligned}$$

satisfy $\alpha' = \alpha(\mathbf{q}')$*,* $\beta' = \beta(\mathbf{q}')$*, and* $\mathbf{h}' = \mathbf{h}(\mathbf{q}')$*.*

Proof. Since $\mathbf{q}' = \mathbf{q} + \delta\,\mathbf{e}_i$, we have:

Update for α*:* $\alpha(\mathbf{q}') = (\mathbf{q}')^\top \mathbf{G}\,\mathbf{w} = (\mathbf{q}+\delta\,\mathbf{e}_i)^\top \mathbf{G}\,\mathbf{w} = \mathbf{q}^\top \mathbf{G}\,\mathbf{w} + \delta\,\mathbf{e}_i^\top \mathbf{G}\,\mathbf{w} = \alpha + \delta\cdot(\mathbf{G}\mathbf{w})_i = \alpha'$.

Update for $\mathbf{h}$*:* $\mathbf{h}' = \mathbf{H}\,\mathbf{q}' = \mathbf{H}(\mathbf{q} + \delta\,\mathbf{e}_i) = \mathbf{H}\mathbf{q} + \delta\,\mathbf{H}\mathbf{e}_i = \mathbf{h} + \delta \cdot \mathbf{H}_{:,i}$.

Update for β*:*

$$\begin{aligned}\beta' &= (\mathbf{q}')^\top \mathbf{H}\,\mathbf{q}' = (\mathbf{q} + \delta\,\mathbf{e}_i)^\top \mathbf{H}\,(\mathbf{q} + \delta\,\mathbf{e}_i)\\ &= \mathbf{q}^\top \mathbf{H}\mathbf{q} + 2\delta\,\mathbf{e}_i^\top \mathbf{H}\mathbf{q} + \delta^2\,\mathbf{e}_i^\top \mathbf{H}\mathbf{e}_i\\ &= \beta + 2\delta\,(\mathbf{H}\mathbf{q})_i + \delta^2\,H_{ii} = \beta + 2\delta \cdot h_i + \delta^2 \cdot H_{ii}.\end{aligned}$$

□

The remaining results concern the group-wise adaptations (Section 3.2). Proposition C.13 bounds the error introduced by the block-diagonal (independent groups) objective, while Proposition C.14 shows that, under the self-activation assumption ($\mathbf{X} = \tilde{\mathbf{X}}$), the sequential group-wise objective coincides with the layer-wise reconstruction objective.

Proposition C.13 (Block-diagonal approximation error)**.** *Let* $\mathbf{s} = (s_1, \ldots, s_K)$ *and define the stacked vectors*

$$\mathbf{w} := \big[\mathbf{w}_{\{1\}}, \ldots, \mathbf{w}_{\{K\}}\big]^\top \qquad \mathbf{q}(\mathbf{s}) := q(\mathbf{w}; \mathbf{s}) = \big[\mathbf{q}_{\{1\}}, \ldots, \mathbf{q}_{\{K\}}\big]^\top \qquad \mathbf{q}_{\{k\}} := q(\mathbf{w}_{\{k\}}; s_k)\,,$$

and the block-wise matrices $\mathbf{G}_{ij} := (\tilde{\mathbf{X}}_{\{i\}})^\top \mathbf{X}_{\{j\}}$ *and* $\mathbf{H}_{ij} := (\tilde{\mathbf{X}}_{\{i\}})^\top \tilde{\mathbf{X}}_{\{j\}}$*. Let* $\mathbf{S}(\mathbf{s}) := \mathrm{blkdiag}(s_1\mathbf{I}, \ldots, s_K\mathbf{I})$ *(with* $\mathbf{I}$ *matching the group dimension), and define the block matrices*

$$\mathbf{G} := \big[\mathbf{G}_{ij}\big]_{i,j=1}^{K}, \qquad \mathbf{H} := \big[\mathbf{H}_{ij}\big]_{i,j=1}^{K},$$

together with their block-diagonal parts

$$\mathbf{G}_{\text{ind}} := \mathrm{blkdiag}(\mathbf{G}_{11}, \ldots, \mathbf{G}_{KK}), \qquad \mathbf{H}_{\text{ind}} := \mathrm{blkdiag}(\mathbf{H}_{11}, \ldots, \mathbf{H}_{KK}),$$

and off-diagonal parts $\mathbf{G}_{\text{off}} := \mathbf{G} - \mathbf{G}_{\text{ind}}$*,* $\mathbf{H}_{\text{off}} := \mathbf{H} - \mathbf{H}_{\text{ind}}$*. For this proposition, we use* E_{ind} *to denote the objective in Equation 11 shifted by the scale-independent constant*

$$\|\mathbf{X}\mathbf{w}\|^2 - \sum_{k=1}^{K} \|\mathbf{X}_{\{k\}}\mathbf{w}_{\{k\}}\|^2.$$

This leaves its minimizers unchanged and gives E *and* E_{ind} *the same constant term in their quadratic expansions. With these definitions, the objective*

$$E(\mathbf{s}) = \|\mathbf{X}\mathbf{w} - \tilde{\mathbf{X}}\mathcal{Q}(\mathbf{w}; \mathbf{s})\|^2$$

can be expressed as

$$E(\mathbf{s}) = \text{const.} - 2\,\mathbf{w}^\top \mathbf{G}^\top \mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s}) + \big(\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\big)^\top \mathbf{H}\big(\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\big).$$

and its block-diagonal approximation

$$E_{\text{ind}}(\mathbf{s}) = \text{const.} - 2\,\mathbf{w}^\top \mathbf{G}_{\text{ind}}^\top \mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s}) + \big(\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\big)^\top \mathbf{H}_{\text{ind}}\big(\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\big).$$

Then, defining

$$\Delta(\mathbf{s}) := 2\,\|\mathbf{w}\|_2\,\|\mathbf{G}_{\text{off}}\|_2\,\|\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\|_2 \;+\; \|\mathbf{H}_{\text{off}}\|_2\,\|\mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})\|_2^2,$$

Assuming both minima are attained, let $\mathbf{s}^*_{\text{ind}} \in \arg\min_{\mathbf{s}} E_{\text{ind}}(\mathbf{s})$ *and* $\mathbf{s}^* \in \arg\min_{\mathbf{s}} E(\mathbf{s})$*. Then,*

$$|E(\mathbf{s}^*_{\text{ind}}) - E(\mathbf{s}^*)| \le \Delta(\mathbf{s}^*_{\text{ind}}) + \Delta(\mathbf{s}^*).$$

Proof. Write $\mathbf{z}(\mathbf{s}) := \mathbf{S}(\mathbf{s})\mathbf{q}(\mathbf{s})$. Then

$$E(\mathbf{s}) - E_{\text{ind}}(\mathbf{s}) = -2\,\mathbf{w}^\top \mathbf{G}_{\text{off}}^\top \mathbf{z}(\mathbf{s}) + \mathbf{z}(\mathbf{s})^\top \mathbf{H}_{\text{off}} \mathbf{z}(\mathbf{s}).$$

Taking absolute values and applying Cauchy–Schwarz and submultiplicativity of the operator norm gives

$$\left|E(\mathbf{s}) - E_{\text{ind}}(\mathbf{s})\right| \le 2\,\|\mathbf{w}\|_2\,\|\mathbf{G}_{\text{off}}\|_2\,\|\mathbf{z}(\mathbf{s})\|_2 + \|\mathbf{H}_{\text{off}}\|_2\,\|\mathbf{z}(\mathbf{s})\|_2^2 =: \Delta(\mathbf{s}). \tag{19}$$

Therefore, the following chain of inequalities is verified:

$$\begin{aligned} E(\mathbf{s}^*_{\text{ind}}) &\le E_{\text{ind}}(\mathbf{s}^*_{\text{ind}}) + \Delta(\mathbf{s}^*_{\text{ind}}) && \text{from Equation 19} \\ &\le E_{\text{ind}}(\mathbf{s}^*) + \Delta(\mathbf{s}^*_{\text{ind}}) && \text{optimality of } \mathbf{s}^*_{\text{ind}} \text{ for } E_{\text{ind}} \\ &\le E(\mathbf{s}^*) + \Delta(\mathbf{s}^*) + \Delta(\mathbf{s}^*_{\text{ind}}) && \text{from Equation 19.} \end{aligned}$$

Therefore

$$|E(\mathbf{s}^*_{\text{ind}}) - E(\mathbf{s}^*)| \le \Delta(\mathbf{s}^*_{\text{ind}}) + \Delta(\mathbf{s}^*),$$

concluding the proof. □

Proposition C.14 (Equivalence of the sequential objective)**.** *Let* $\mathbf{w} = [\mathbf{w}_{\{1\}}, \dots, \mathbf{w}_{\{K\}}]^\top$ *be partitioned into* K *groups. Define the partially quantized vector*

$$\tilde{\mathbf{w}}(s_k) := [\tilde{\mathbf{w}}_{\{1\}}, \dots, \tilde{\mathbf{w}}_{\{k-1\}}, s_k\, q(\mathbf{w}_{\{k\}}, s_k), \mathbf{w}_{\{k+1\}}, \dots, \mathbf{w}_{\{K\}}]^\top,$$

where $\tilde{\mathbf{w}}_{\{i\}} = s_i^* q(\mathbf{w}_{\{i\}}, s_i^*)$, $i \in [k-1]$*, with the scales* $s_1^*, \dots, s_{k-1}^*$ *fixed for the first* $k-1$ *groups. Consider the objectives*

$$\overline{E}_{\text{seq}}(s_k) := \left\|\mathbf{X}\,\mathbf{w} - \tilde{\mathbf{X}}\,\tilde{\mathbf{w}}(s_k)\right\|^2, \tag{20}$$

$$E_{\text{seq}}(s_k) := \left\|\mathbf{X}_{\{\le k\}}\,\mathbf{w}_{\{\le k\}} - \tilde{\mathbf{X}}_{\{<k\}}\,\tilde{\mathbf{w}}_{\{<k\}} - s_k\,\tilde{\mathbf{X}}_{\{k\}}\,q(\mathbf{w}_{\{k\}}, s_k)\right\|^2. \tag{21}$$

In the self-activation case ($\mathbf{X} = \tilde{\mathbf{X}}$*), the two objectives coincide for every admissible* s_k*. Consequently, they have the same infimum and, whenever it is attained, the same set of minimizers:* $\arg\min_{s_k} \overline{E}_{\text{seq}}(s_k) = \arg\min_{s_k} E_{\text{seq}}(s_k)$.

Proof. Write $\mathbf{q}_{\{k\}} := q(\mathbf{w}_{\{k\}}, s_k)$ for brevity. Decomposing column-wise:

$$\mathbf{X}\,\mathbf{w} = \mathbf{X}_{\{\le k\}}\,\mathbf{w}_{\{\le k\}} + \mathbf{X}_{\{>k\}}\,\mathbf{w}_{\{>k\}},$$

$$\tilde{\mathbf{X}}\,\tilde{\mathbf{w}}(s_k) = \tilde{\mathbf{X}}_{\{<k\}}\,\tilde{\mathbf{w}}_{\{<k\}} + s_k\,\tilde{\mathbf{X}}_{\{k\}}\,\mathbf{q}_{\{k\}} + \tilde{\mathbf{X}}_{\{>k\}}\,\mathbf{w}_{\{>k\}}.$$

Subtracting and defining

$$\begin{aligned} A(s_k) &:= \mathbf{X}_{\{\le k\}}\,\mathbf{w}_{\{\le k\}} - \tilde{\mathbf{X}}_{\{<k\}}\,\tilde{\mathbf{w}}_{\{<k\}} - s_k\,\tilde{\mathbf{X}}_{\{k\}}\,\mathbf{q}_{\{k\}}, \\ B &:= \left(\mathbf{X}_{\{>k\}} - \tilde{\mathbf{X}}_{\{>k\}}\right)\mathbf{w}_{\{>k\}}, \end{aligned}$$

we obtain

$$\mathbf{X}\,\mathbf{w} - \tilde{\mathbf{X}}\,\tilde{\mathbf{w}}(s_k) = A(s_k) + B.$$

Note that $E_{\text{seq}}(s_k) = \|A(s_k)\|^2$. Expanding $\overline{E}_{\text{seq}}$:

$$\overline{E}_{\text{seq}}(s_k) = \|A(s_k) + B\|^2 = \underbrace{\|A(s_k)\|^2}_{=E_{\text{seq}}(s_k)} + 2\langle A(s_k),\, B\rangle + \underbrace{\|B\|^2}_{\text{const. in } s_k}. \tag{22}$$

Then, if $\mathbf{X} = \tilde{\mathbf{X}}$, then $\mathbf{X}_{\{>k\}} - \tilde{\mathbf{X}}_{\{>k\}} = \mathbf{0}$, so $B = \mathbf{0}$. Hence, Equation 22 reduces to $\overline{E}_{\text{seq}}(s_k) = E_{\text{seq}}(s_k) + 2\langle A(s_k),\, \mathbf{0}\rangle + \|\mathbf{0}\|^2 = E_{\text{seq}}(s_k)$, so the two objectives coincide, as claimed. □

D LOW-DIMENSIONAL EXAMPLE

We provide a two-dimensional example illustrating the piecewise-quadratic structure exploited by PiSO and why a scale that is optimal for data-free weight reconstruction can be substantially suboptimal for layer-output reconstruction. Consider a single output channel with

$$\mathbf{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \mathbb{G} = \{-1, 0, 1\}, \qquad \mathbf{X} = \tilde{\mathbf{X}} = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}.$$

The corresponding matrices in Equation 6 are

$$\mathbf{H} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{G} = \tilde{\mathbf{X}}^\top \mathbf{X} = \begin{bmatrix} 9 & 3 \\ 3 & 2 \end{bmatrix}.$$

The unequal diagonal entries show that errors in the two weights are amplified differently, while the nonzero off-diagonal entries capture their interaction at the layer output. We consider positive scales and RTN quantization,

Step 1: Computing the transition scales. Since the unscaled grid $\mathbb{G}$ has three values, we get two midpoints which are

$$m_1 = -\frac{1}{2}, \qquad m_2 = \frac{1}{2}.$$

Since we are focusing on positive scales ($s > 0$) we only have to consider the positive midpoint above ($m_2 = 1/2$). Using Equation 8, we obtain the following transition scales

$$t_{1,2} = \frac{w_1}{m_2} = \frac{1}{1/2} = 2, \qquad t_{2,2} = \frac{w_2}{m_2} = \frac{2}{1/2} = 4.$$

Sorting the transition scales in decreasing order gives

$$\tau_1 = 4, \qquad \tau_2 = 2.$$

No component of $q(\mathbf{w}; s)$ changes between two consecutive transition scales. The transition scales partition the positive scale domain into intervals on which the grid assignment is constant:

$$q(\mathbf{w}; s) = \begin{cases} (0,0)^\top, & s \in [4, +\infty), \\ (0,1)^\top, & s \in [2, 4), \\ (1,1)^\top, & s \in (0, 2). \end{cases}$$

Since the assignment is constant within each interval, it is sufficient to evaluate $q(\mathbf{w}; s)$ at any one scale in that interval to determine the assignment for the whole interval. For example, choosing $s = 8 \in (4, +\infty)$ gives

$$q(\mathbf{w}; 8) = \lfloor \mathbf{w}/8 \rceil_{\mathbb{G}} = \lfloor (1/8,\ 1/4)^\top \rceil_{\mathbb{G}} = (0,0)^\top.$$

The assignments in the other intervals can be obtained analogously.

Following the quadratic decomposition in Equation 7, the term independent of s is

$$c := \|\mathbf{X}\mathbf{w}\|^2 = \|(5,2)^\top\|^2 = 29.$$

Substituting the three assignments into

$$E(s) = c - 2s\alpha + s^2\beta$$

gives the piecewise-quadratic function

$$E(s) = \begin{cases} 29, & s \in [4, +\infty), \\ 29 - 14s + 2s^2, & s \in [2, 4), \\ 29 - 44s + 17s^2, & s \in (0, 2). \end{cases}$$

We next recover these pieces by explicitly following the incremental PiSO sweep.

Step 2: Initializing the sweep. The sweep proceeds from large to small scales. We first precompute

$$\mathbf{p} := \mathbf{G}\mathbf{w} = \begin{bmatrix} 9 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 15 \\ 7 \end{bmatrix}.$$

For $s > 4$, the assignment is

$$\mathbf{q}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Therefore, the running quantities are initialized as

$$\alpha_0 = \mathbf{q}_0^\top \mathbf{G}\mathbf{w} = 0, \qquad \beta_0 = \mathbf{q}_0^\top \mathbf{H}\mathbf{q}_0 = 0, \qquad \mathbf{h}_0 = \mathbf{H}\mathbf{q}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The objective is consequently constant on this interval:

$$\varphi_0(s) = c = 29.$$

For each subsequent interval, PiSO maintains

$$\alpha_j = \mathbf{q}_j^\top \mathbf{G}\mathbf{w}, \qquad \beta_j = \mathbf{q}_j^\top \mathbf{H}\mathbf{q}_j, \qquad \mathbf{h}_j = \mathbf{H}\mathbf{q}_j.$$

If one assignment changes by δ at coordinate i, these quantities are updated as

$$\alpha \leftarrow \alpha + \delta p_i, \qquad \beta \leftarrow \beta + 2\delta h_i + \delta^2 H_{ii}, \qquad \mathbf{h} \leftarrow \mathbf{h} + \delta \mathbf{H}_{:,i}.$$

Step 3: Crossing the transition $\tau_1 = 4$. At $\tau_1 = 4$, the second grid assignment changes from 0 to 1. Thus,

$$i = 2, \qquad \delta = 1, \qquad \mathbf{q}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Starting from $\alpha_0, \beta_0, \mathbf{h}_0$, the incremental updates give

$$\begin{aligned} \alpha_1 &= \alpha_0 + \delta p_2 = 0 + 1 \cdot 7 = 7, \\ \beta_1 &= \beta_0 + 2\delta(\mathbf{h}_0)_2 + \delta^2 H_{22} = 0 + 0 + 2 = 2, \\ \mathbf{h}_1 &= \mathbf{h}_0 + \delta \mathbf{H}_{:,2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}. \end{aligned}$$

These values can also be verified directly:

$$\alpha_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 7 \end{bmatrix} = 7, \qquad \beta_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 9 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 2.$$

The quadratic associated with this interval is

$$\varphi_1(s) = c - 2\alpha_1 s + \beta_1 s^2 = 29 - 14s + 2s^2.$$

Its unconstrained minimizer is

$$s_1^{\mathrm{unc}} = \frac{\alpha_1}{\beta_1} = \frac{7}{2}.$$

Since

$$2 < \frac{7}{2} < 4,$$

the minimizer lies strictly inside the interval and no clamping is needed:

$$s_1^* = \mathrm{clamp}\left(\frac{\alpha_1}{\beta_1},\, 2,\, 4\right) = \frac{7}{2}.$$

Using the reduced score maintained by the algorithm,

$$f_j := (s_j^*)^2 \beta_j - 2 s_j^* \alpha_j,$$

we obtain

$$f_1 = \left(\frac{7}{2}\right)^2 2 - 2\left(\frac{7}{2}\right) 7 = -\frac{49}{2}.$$

The corresponding full layer-output error is

$$\varphi_1(s_1^*) = c + f_1 = 29 - \frac{49}{2} = \frac{9}{2}.$$

Step 4: Crossing the transition $\tau_2 = 2$. At $\tau_2 = 2$, the first grid assignment changes from 0 to 1. Thus,

$$i = 1, \qquad \delta = 1, \qquad \mathbf{q}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Starting from the quantities for the preceding interval, the incremental updates give

$$\begin{aligned} \alpha_2 &= \alpha_1 + \delta p_1 = 7 + 1 \cdot 15 = 22, \\ \beta_2 &= \beta_1 + 2\delta(\mathbf{h}_1)_1 + \delta^2 H_{11} = 2 + 2 \cdot 3 + 9 = 17, \\ \mathbf{h}_2 &= \mathbf{h}_1 + \delta \mathbf{H}_{:,1} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} + \begin{bmatrix} 9 \\ 3 \end{bmatrix} = \begin{bmatrix} 12 \\ 5 \end{bmatrix}. \end{aligned}$$

Again, the resulting quantities can be checked directly:

$$\alpha_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 7 \end{bmatrix} = 22,$$

and

$$\beta_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 17.$$

The quadratic associated with this interval is

$$\varphi_2(s) = c - 2\alpha_2 s + \beta_2 s^2 = 29 - 44s + 17s^2.$$

Its unconstrained minimizer is

$$s_2^{\mathrm{unc}} = \frac{\alpha_2}{\beta_2} = \frac{22}{17}.$$

Since

$$0 < \frac{22}{17} < 2,$$

this minimizer also lies strictly inside its interval:

$$s_2^* = \mathrm{clamp}\left(\frac{\alpha_2}{\beta_2},\, 0,\, 2\right) = \frac{22}{17}.$$

Its reduced score is

$$f_2 = -\frac{\alpha_2^2}{\beta_2} = -\frac{484}{17},$$

and its full layer-output error is

$$\varphi_2(s_2^*) = c + f_2 = 29 - \frac{484}{17} = \frac{9}{17}.$$

Step 5: Selecting the best interval. The smallest layer-output error found in each interval is

$$\varphi_0 = 29, \qquad \varphi_1(s_1^*) = \frac{9}{2}, \qquad \varphi_2(s_2^*) = \frac{9}{17}.$$

Since

$$\frac{9}{17} < \frac{9}{2} < 29,$$

PiSO selects the data-aware scale

$$s_{\mathrm{out}}^* = \frac{22}{17}.$$

Step 6: Comparing with the data-free optimum. Data-free scale optimization instead minimizes

$$E_{\mathrm{df}}(s) = \|\mathbf{w} - s\, q(\mathbf{w}; s)\|^2.$$

The transition scales and grid assignments are unchanged because they depend only on $\mathbf{w}$ and $\mathbb{G}$. For $s \in (0, 2)$, where

$$q(\mathbf{w}; s) = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

the data-free objective is

$$E_{\mathrm{df}}(s) = (1 - s)^2 + (2 - s)^2 = 5 - 6s + 2s^2.$$

Its minimizer is

$$s^*_{\mathrm{df}} = \frac{3}{2}, \qquad E_{\mathrm{df}}(s^*_{\mathrm{df}}) = \frac{1}{2}.$$

For $s \in [2, 4)$, the smallest possible data-free error is 1, while for $s \in [4, +\infty)$ it is 5. Hence $s^*_{\mathrm{df}} = 3/2$ is the globally optimal data-free scale.

Although the data-free and data-aware scales both induce the same RTN assignment,

$$q(\mathbf{w}; s^*_{\mathrm{df}}) = q(\mathbf{w}; s^*_{\mathrm{out}}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

their layer-output errors differ substantially. Evaluating the data-free optimal scale under the actual layer-output objective gives

$$\begin{aligned} E(s^*_{\mathrm{df}}) &= \left\| \mathbf{X} \left(\mathbf{w} - \frac{3}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \right\|^2 \\ &= \left\| \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/2 \\ 1/2 \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} -1 \\ 1/2 \end{bmatrix} \right\|^2 = \frac{5}{4}. \end{aligned}$$

By comparison, the data-aware scale selected by PiSO gives

$$E(s^*_{\mathrm{out}}) = \frac{9}{17}.$$

Consequently,

$$\frac{E(s^*_{\mathrm{df}})}{E(s^*_{\mathrm{out}})} = \frac{5/4}{9/17} = \frac{85}{36} \approx 2.36.$$

Thus, even in this two-weight example, the scale that is globally optimal for weight reconstruction produces more than twice the layer-output error of the data-aware scale. Note that both scales induce exactly the same grid assignment, but the layer output error differs. The improvement comes from selecting the scale according to how the calibration activations amplify and correlate the weight errors. More broadly, this example shows how PiSO enumerates the finitely many intervals on which the RTN assignments are constant, obtains the corresponding quadratic through inexpensive incremental updates, and selects the scale achieving the smallest layer-output reconstruction error.

E COMPLEMENTARY RESULTS

This appendix presents the complete quantization results summarized in the main text. The results are organized by quantization granularity: channel-wise first, then group-wise.

Baseline implementations. Baseline PTQ algorithms were run using the Brevitas quantization library (Pappalardo et al., 2025), with the exception of Beacon, whose implementation was not publicly available and was therefore implemented by us.

Runtime analysis. All runtimes were measured on the same GPU model, with each model quantized and evaluated on a single GPU. Table 4 reports the wall-clock runtimes of GPTQ with PiSO on two Llama 3 models (1B and 3B) at 2-bit and 3-bit weight quantization. GPTQ $\odot$ PiSO introduces negligible overhead (between $0.99\times$ and $1.13\times$ the baseline GPTQ runtime), while GPTQ $+$ PiSO incurs a 1.25–$1.37\times$ overhead due to the additional forward passes over the calibration data. Qronos itself runs at $1.09\times$–$1.13\times$ the GPTQ runtime, and Qronos $\odot$ PiSO remains in the same regime at $1.12\times$–$1.22\times$, i.e. PiSO adds only $1.02\times$–$1.12\times$ on top of Qronos. In contrast, the gradient-based AutoRound (Cheng et al., 2024) requires $2.93\times$–$4.67\times$, so the gains reported in Tables 10 and 11 are obtained at a fraction of its quantization cost.

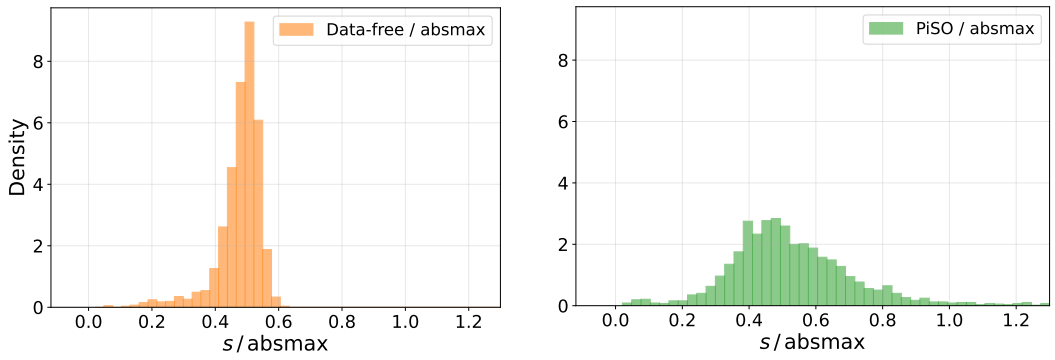
Table 4: Wall-clock absolute (in minutes) and relative (compared to GPTQ) runtimes of GPTQ and Qronos with PiSO scale optimization, and of AutoRound (Cheng et al., 2024). $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3). All relative runtimes are expressed with respect to GPTQ.

		GPTQ	GPTQ + PiSO		GPTQ ⊙ PiSO		Qronos		Qronos ⊙ PiSO		AutoRound	
Model	Bits	min	min	×	min	×	min	×	min	×	min	×
Llama-3.2-1B	2	35.3	48.3	1.37	36.4	1.03	38.9	1.10	39.5	1.12	164.7	4.67
Llama-3.2-1B	3	35.9	44.7	1.25	35.5	0.99	40.4	1.13	41.4	1.15	165.4	4.61
Llama-3.2-3B	2	77.0	104.8	1.36	85.9	1.11	83.8	1.09	88.8	1.15	283.6	3.68
Llama-3.2-3B	3	78.1	105.1	1.34	88.3	1.13	84.9	1.09	95.5	1.22	228.7	2.93

Scale Distribution. Figure 4 shows the optimal channel-wise scale distributions, normalized by the absmax scale, for the data-free ($\mathbf{X} = \tilde{\mathbf{X}} = \mathbf{I}$) and cross-activation ($\mathbf{X} \neq \tilde{\mathbf{X}}$) objectives in third down-projection layer of Llama-3.2-1B. One can clearly observe how the distribution of the scales varies depending on the objective with respect to which they are optimized. Notably, the distribution of optimal scales under the cross-activation objective exhibits a larger spread, with a non-negligible fraction exceeding the absmax value (ratios greater than 1). This behavior highlights one of the limitations of grid-search–based methods such as CDQuant (Nair & Suggala, 2024), which restrict the search space to the interval $(0, s_{\text{absmax}}]$.

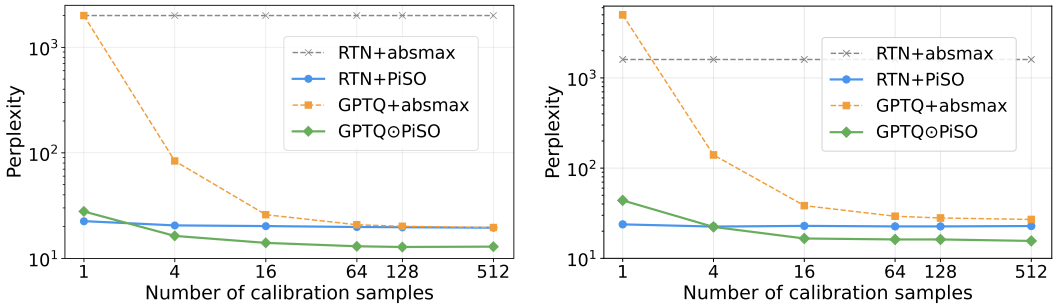
Calibration size sensitivity Figure 5 extends the calibration size sensitivity experiments from the Results section in the main text (Figure 2). The results show that with either matched or mismatched calibration/evaluation datasets, PiSO continues to deliver large gains over the absmax baseline. When PiSO is not used, the gap between the matched/mismatched settings increases. PiSO is the most data-efficient strategy across domains. Under matched WikiText2 calibration, one sequence already gives strong RTN + PiSO perplexity (22.5). Under domain shift, a single sequence is noisier but performance stabilizes quickly (with as few as 4–16 sequences, out-of-domain calibration essentially reaches its asymptotic quality).

Channel-wise Quantization. Tables 5–7 report channel-wise weight-only quantization results at 2, 3, and 4 bits on Llama-3 and Qwen-2.5 Instruct models using the standard integer grid, unpacking the scale optimization objectives and integration strategies with error correction methods that are aggregated in Tables 1 and 2. Table 8 demonstrates the catastrophic quality degradation of data-free scale selection strategies in channel-wise low-bit quantization. Table 9 evaluates PiSO on the non-uniform `E2M1` (4-bit floating-point) grid, demonstrating that the algorithm applies to arbitrary quantization grids beyond integer formats. Finally, Tables 10 and 11 compare PiSO against Au-



$s^* = \arg\min_s \|\mathbf{w} - s\, q(\mathbf{w}; s)\|^2$ $\quad$ $s^* = \arg\min_s \|\mathbf{X}\mathbf{w} - s\, \tilde{\mathbf{X}}\, q(\mathbf{w}; s)\|^2$

Figure 4: Optimal scale distributions for the data-free (left) and cross-activation (right) objectives, normalized by the absmax scales, for the 2048 output channels of the third down-projection layer of Llama-3.2-1B.



(a) WikiText2 calibration, WikiText2 perplexity (↓) (b) FineWeb calibration, WikiText2 perplexity (↓)

Figure 5: Effect of calibration set size on Llama-3.2-1B under 3-bit channel-wise weight quantization. The trend displayed in the main text for C4 calibration (Figure 2) holds: PiSO is stable with few samples, whereas GPTQ with absmax requires ≥ 64 samples to converge. GPTQ⊙PiSO achieves the best perplexity and improves the data efficiency of GPTQ+absmax .

toRound (Cheng et al., 2024), a gradient-based PTQ method that learns the rounding decisions and the scales jointly.

Table 5: Channel-wise 2-bit quantization results on Llama-3 and Qwen-2.5 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, Beacon, AWQ). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

		2-bit Llama						**2-bit Qwen**					
		WikiText2 (↓)			0-shot (↑)			WikiText2 (↓)			0-shot (↑)		
	Obj.	1B	3B	8B	1B	3B	8B	1.5B	3B	7B	1.5B	3B	7B
BF16		8.97	7.14	5.89	52.6	62.0	68.6	8.89	7.88	6.92	54.9	56.5	57.7
RTN + absmax		2e5	4e5	4e5	33.0	32.5	33.4	7e6	6e6	1e5	31.8	32.9	32.3
RTN + OptQuant	**I-I**	2e6	8e5	2e6	32.8	32.3	32.1	2e8	3e8	7e6	32.2	31.4	32.2
RTN + PiSO	**X-X**	1e5	4e4	2e5	32.0	31.6	32.5	1e4	2e4	9e3	31.8	32.2	32.3
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	2e4	2e4	8e4	31.9	32.2	32.7	3e3	3e3	7e3	32.4	31.6	31.8
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	3e3	3e3	3e3	32.3	31.6	32.0	2e3	2e3	4e3	31.8	32.7	32.2
AWQ		2e6	8e5	1e6	32.7	32.0	32.1	9e7	7e8	9e6	32.4	31.7	32.5
Beacon		52.8	72.4	25.5	33.8	33.3	36.8	1e3	33.2	21.9	32.5	32.4	33.9
GPTQ + absmax		3e3	415	331	31.6	31.9	31.5	675	596	72.7	33.0	32.8	32.1
GPTQ + OptQuant	**I-I**	3e3	9e3	1e6	31.7	32.3	31.8	353	1e5	185	32.1	31.1	34.0
GPTQ + PiSO	**X-X**	116	69.7	59.1	32.6	33.1	33.3	45.3	32.7	**15.6**	33.3	32.9	37.1
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	92.5	47.5	32.5	33.1	33.2	34.7	**39.2**	38.5	15.9	**34.4**	33.6	34.9
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	54.7	29.5	22.3	35.5	36.3	40.2	44.7	30.6	17.0	33.7	**34.4**	34.9
Qronos + absmax		197	68.9	49.5	31.6	32.7	33.5	111	35.8	2e4	32.1	32.5	32.1
Qronos + OptQuant	**I-I**	5e3	1e4	3e5	31.2	31.5	31.6	3e3	657	1e4	31.2	32.6	32.6
Qronos + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	216	84.8	76.0	32.1	33.3	33.0	196	27.4	24.5	32.3	32.9	36.1
Qronos $\odot$ PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	**27.9**	**18.9**	**16.6**	**37.5**	**41.4**	**44.0**	131	**17.2**	16.3	32.7	**34.8**	**38.1**

Table 6: Channel-wise 3-bit and 4-bit quantization results on Llama-3 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, Beacon, AWQ). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

		3-bit						**4-bit**					
		WikiText2 (↓)			0-shot (↑)			WikiText2 (↓)			0-shot (↑)		
	Obj.	1B	3B	8B	1B	3B	8B	1B	3B	8B	1B	3B	8B
BF16		8.97	7.14	5.89	52.6	62.0	68.6	8.97	7.14	5.89	52.6	62.0	68.6
RTN + absmax		2e3	436	1e3	32.8	35.7	38.0	22.3	9.61	7.83	46.3	55.6	64.7
RTN + OptQuant	**I-I**	9e3	6e4	5e4	31.9	32.4	32.2	136	8e3	9e4	33.8	31.8	32.2
RTN + PiSO	**X-X**	27.7	11.4	9.69	42.5	50.4	59.3	11.3	8.03	6.72	50.7	59.9	67.2
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	29.7	11.5	9.96	41.3	49.2	59.1	11.5	8.04	6.74	49.2	60.2	66.7
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	19.5	10.6	8.99	44.2	52.1	59.9	11.0	7.92	6.64	50.0	60.4	67.3
AWQ		102	50.1	11.8	38.2	52.5	58.7	11.9	8.02	6.76	50.4	59.8	66.6
Beacon		19.1	33.7	9.35	42.4	40.1	53.5	13.6	25.1	7.53	46.1	46.3	59.7
GPTQ + absmax		19.6	11.3	9.23	39.6	49.8	55.1	10.4	7.79	6.49	49.3	59.4	65.6
GPTQ + OptQuant	**I-I**	546	9e3	8e5	33.1	32.0	31.7	255	9e3	4e5	35.1	31.5	31.7
GPTQ + PiSO	**X-X**	12.9	9.19	7.42	47.5	54.6	**63.0**	9.94	7.63	6.32	**52.0**	60.2	**68.3**
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	13.0	9.20	7.51	46.1	53.5	62.4	9.94	7.64	6.32	50.9	59.7	67.5
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	12.8	9.08	7.41	47.0	**55.9**	62.6	9.91	7.65	6.32	51.3	60.0	68.1
Qronos + absmax		15.0	9.83	8.27	40.6	51.4	57.9	10.0	7.64	6.36	50.2	60.1	65.2
Qronos + OptQuant	**I-I**	2e3	9e3	8e5	32.8	31.9	32.3	335	9e3	4e5	34.6	31.8	31.1
Qronos + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	11.6	8.42	7.01	**48.6**	55.5	61.4	9.61	**7.44**	6.19	**51.5**	**61.2**	67.4
Qronos $\odot$ PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	**11.6**	**8.41**	**6.98**	48.0	**55.6**	**62.7**	**9.59**	**7.44**	**6.18**	51.1	60.5	67.0

Table 7: Channel-wise 3-bit and 4-bit quantization results on Qwen-2.5 models reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, Beacon, AWQ). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

		3-bit						**4-bit**					
		WikiText2 (↓)			0-shot (↑)			WikiText2 (↓)			0-shot (↑)		
	Obj.	1.5B	3B	7B	1.5B	3B	7B	1.5B	3B	7B	1.5B	3B	7B
BF16		8.89	7.88	6.92	54.9	56.5	57.7	8.89	7.88	6.92	54.9	56.5	57.7
RTN + absmax		861	5e5	9e7	33.7	32.1	32.7	13.9	1e4	9.13	49.5	31.6	53.5
RTN + OptQuant	**I-I**	142	1e5	535	34.6	32.4	33.5	14.3	13.9	10.9	51.9	50.6	55.0
RTN + PiSO	$\mathbf{X}$-$\mathbf{X}$	16.9	11.9	9.88	46.8	48.1	54.6	10.0	8.60	7.55	54.7	54.6	58.7
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	16.8	12.0	10.0	44.6	46.8	53.2	10.0	8.60	7.56	54.5	55.2	58.0
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.9	11.5	9.50	46.9	50.8	51.8	10.00	8.67	7.51	54.0	55.1	57.6
AWQ		27.0	16.6	20.3	45.3	47.2	49.2	10.7	9.02	7.87	51.7	54.4	56.7
Beacon		380	12.3	12.6	31.6	39.9	40.4	142	10.2	10.8	33.6	44.8	43.2
GPTQ + absmax		14.1	11.5	9.23	41.1	42.0	44.6	9.76	8.54	7.34	53.2	54.7	55.0
GPTQ + OptQuant	**I-I**	29.2	914	8.71	43.7	32.3	52.2	14.2	8.72	9.02	43.2	54.7	**60.1**
GPTQ + PiSO	$\mathbf{X}$-$\mathbf{X}$	11.5	9.43	7.90	48.2	**55.0**	56.5	9.50	8.26	7.17	**55.2**	55.7	59.0
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	11.6	9.45	7.91	47.0	53.3	55.7	9.49	8.25	7.18	53.5	**57.8**	59.5
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	**11.4**	9.41	7.82	**50.6**	52.0	57.0	9.47	8.26	7.15	54.7	56.9	58.5
Qronos + absmax		16.7	10.4	8.86	35.3	41.9	52.2	9.66	8.42	7.30	51.4	54.2	55.6
Qronos + OptQuant	**I-I**	105	168	11.3	33.1	32.0	43.6	16.8	10.1	8.83	41.9	45.0	59.6
Qronos + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	11.8	**9.10**	7.73	43.8	53.5	**58.3**	9.36	**8.19**	7.12	52.0	56.8	58.1
Qronos $\odot$ PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	12.0	**9.10**	**7.73**	39.9	53.8	57.5	**9.34**	8.20	**7.11**	52.5	54.7	58.1

Table 8: Channel-wise quantization results on Llama-3 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). None of these scale selection methods requires calibration data: the grid search of the Brevitas library (Pappalardo et al., 2025) approximates the minimizer of the weight reconstruction objective by evaluating a uniform grid of candidates; while data-free optimizes this objective analytically.

	2-bit				**3-bit**				**4-bit**			
	WikiText2 (↓)		0-shot (↑)		WikiText2 (↓)		0-shot (↑)		WikiText2 (↓)		0-shot (↑)	
	1B	3B	1B	3B	1B	3B	1B	3B	1B	3B	1B	3B
BF16	8.97	7.14	52.6	62.0	8.97	7.14	52.6	62.0	8.97	7.14	52.6	62.0
RTN + grid search	2e6	8e5	**32.9**	32.3	4e3	1e5	32.0	31.7	**108**	**4e3**	35.0	**32.3**
RTN + OptQuant	2e6	8e5	32.8	32.3	9e3	6e4	31.9	**32.4**	136	8e3	33.8	31.8
GPTQ + grid search	**4e3**	**7e3**	32.3	**32.5**	452	**9e3**	32.8	32.4	258	8e3	**35.1**	31.8
GPTQ + OptQuant	5e3	8e3	31.8	32.4	**390**	1e4	**33.3**	32.1	300	9e3	34.6	31.8

Table 9: Channel-wise 4-bit quantization results using the `E2M1` (`FP4`) floating-point format on Llama-3 models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column specifies the scale-optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax).

		WikiText2 (↓)		0-shot (↑)	
	Obj.	1B	3B	1B	3B
BF16		8.97	7.14	52.6	62.0
RTN + absmax		12.6	8.39	49.2	59.5
RTN + PiSO	$\mathbf{X}$-$\mathbf{X}$	10.4	7.75	50.7	60.5
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	10.4	7.75	50.5	61.0
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	10.3	7.69	50.6	59.8
GPTQ + absmax		9.78	7.52	50.8	60.1
GPTQ + PiSO	$\mathbf{X}$-$\mathbf{X}$	9.67	7.52	**51.1**	**62.1**
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	**9.67**	**7.52**	50.5	60.8
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	9.69	7.52	50.5	60.7

Table 10: Comparison against AutoRound (Cheng et al., 2024) for channel-wise weight-only quantization of Llama-3 models, reporting WikiText2 perplexity (↓). All methods share the same protocol: 512 WikiText2 calibration sequences of 2048 tokens, symmetric per-channel weight quantization, and the same zero-shot task suite. For AutoRound, the hyperparameter configuration of Cheng et al. (2024) was adapted. For error correction with PiSO, $+/\odot$ denote decoupled/interleaved (Section 3.3).

	2-bit			**3-bit**			**4-bit**		
	1B	3B	8B	1B	3B	8B	1B	3B	8B
BF16	8.97	7.14	5.89	8.97	7.14	5.89	8.97	7.14	5.89
RTN + absmax	2e5	4e5	4e5	2e3	436	1e3	22.3	9.61	7.83
RTN + PiSO	3e3	3e3	3e3	19.5	10.6	8.99	11.0	7.92	6.64
AutoRound	87.6	37.7	35.7	14.8	10.0	8.72	10.0	7.57	6.48
GPTQ + PiSO	92.5	47.5	32.5	12.9	9.19	7.42	9.94	7.63	6.32
GPTQ $\odot$ PiSO	54.7	29.5	22.3	12.8	9.08	7.41	9.91	7.65	6.32
Qronos + PiSO	216	84.8	76.0	**11.6**	8.42	7.01	9.61	**7.44**	6.19
Qronos $\odot$ PiSO	**27.9**	**18.9**	**16.6**	11.6	**8.41**	**6.98**	**9.59**	7.44	**6.18**

Table 11: Comparison against AutoRound (Cheng et al., 2024) for channel-wise weight-only quantization of Llama-3 models, reporting average zero-shot accuracy (↑). All methods share the same protocol: 512 WikiText2 calibration sequences of 2048 tokens, symmetric per-channel weight quantization, and the same zero-shot task suite. For AutoRound, the hyperparameter configuration of Cheng et al. (2024) was adopted. For error correction with PiSO, $+/\odot$ denote decoupled/interleaved (Section 3.3).

	2-bit			**3-bit**			**4-bit**		
	1B	3B	8B	1B	3B	8B	1B	3B	8B
BF16	52.6	62.0	68.6	52.6	62.0	68.6	52.6	62.0	68.6
RTN + absmax	33.0	32.5	33.4	32.8	35.7	38.0	46.3	55.6	64.7
RTN + PiSO	32.3	31.6	32.0	44.2	52.1	59.9	50.0	60.4	67.3
AutoRound	33.2	34.1	32.5	45.6	53.6	60.0	51.1	59.9	66.4
GPTQ + PiSO	33.1	33.2	34.7	47.5	54.6	**63.0**	**52.0**	60.2	**68.3**
GPTQ $\odot$ PiSO	35.5	36.3	40.2	47.0	**55.9**	62.6	51.3	60.0	68.1
Qronos + PiSO	32.1	33.3	33.0	**48.6**	55.5	61.4	51.5	**61.2**	67.4
Qronos $\odot$ PiSO	**37.5**	**41.4**	**44.0**	48.0	55.6	62.7	51.1	60.5	67.0

Extended Model Coverage. Tables 12 and 13 extend the channel-wise 3-bit evaluation to Qwen2.5-14B-Instruct and to the Gemma-2 Instruct models. The trends match those observed on the smaller models: PiSO recovers near-BF16 quality where the absmax scales collapse, improving RTN from 77.5 / 36.6 to 7.20 / 60.1 and GPTQ from 7.37 / 51.5 to 6.49 / 62.4 on Qwen2.5-14B-Instruct, and RTN from 13.6 / 58.3 to 8.57 / 63.8 and GPTQ from 8.52 / 61.8 to 8.03 / 65.9 on Gemma-2-9B-it.

Table 12: Channel-wise 3-bit quantization results on Qwen2.5-14B-Instruct, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

	Obj.	WikiText2 (↓)	0-shot (↑)
BF16		5.73	61.6
RTN + absmax		77.5	36.6
RTN + OptQuant	**I**-**I**	199	37.5
RTN + PiSO	$\mathbf{X}$-$\mathbf{X}$	7.48	61.7
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	7.49	61.4
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	7.20	60.1
GPTQ + absmax		7.37	51.5
GPTQ + OptQuant	**I**-**I**	7.04	55.7
GPTQ + PiSO	$\mathbf{X}$-$\mathbf{X}$	6.49	62.4
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	6.50	**62.5**
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	**6.47**	61.7

Extended Task Coverage. Table 14 extends the evaluation of channel-wise 3-bit Qwen-2.5 Instruct models to MMLU (Hendrycks et al., 2021), IFEval (Zhou et al., 2023), and LongBench (Bai et al., 2024). The trends are consistent with those observed for WikiText2 perplexity and average zero-shot accuracy: optimizing the scales with PiSO improves over the absmax scales on every task, model, and pipeline. The gap is widest without error correction, where absmax collapses to chance-level MMLU and near-zero LongBench performance (MMLU 25.5 → 55.2 and LongBench 1.2 → 17.0 on the 3B model).

Table 13: Channel-wise 3-bit quantization results on Gemma-2 Instruct models, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

		WikiText2 (↓)		0-shot (↑)	
	Obj.	2B	9B	2B	9B
BF16		10.7	7.59	57.6	65.8
RTN + absmax		41.2	13.6	47.2	58.3
RTN + PiSO	$\mathbf{X}$-$\mathbf{X}$	14.4	8.64	55.2	64.4
RTN + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	14.4	8.68	55.0	64.4
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.3	8.57	54.5	63.8
GPTQ + absmax		13.5	8.52	52.4	61.8
GPTQ + PiSO	$\mathbf{X}$-$\mathbf{X}$	**12.1**	**8.03**	54.9	**65.9**
GPTQ + PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	**12.1**	**8.03**	55.8	65.1
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	12.2	8.06	**56.0**	65.6

Table 14: Channel-wise 3-bit quantization results on Qwen-2.5 Instruct models, reporting WikiText2 perplexity (↓), average zero-shot accuracy (↑), and MMLU, IFEval, and LongBench performance (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax). For PiSO, $+/\odot$ denote decoupled/interleaved optimization (see Section 3.3).

		WikiText2 (↓)		0-shot (↑)		MMLU (↑)		IFEval (↑)		LongBench (↑)	
	Obj.	1.5B	3B	1.5B	3B	1.5B	3B	1.5B	3B	1.5B	3B
BF16		8.89	7.88	54.9	56.5	60.1	65.5	27.0	47.7	28.5	28.4
RTN + absmax		861	5e5	33.7	32.1	24.6	25.5	9.4	11.3	2.6	1.2
RTN + PiSO	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.9	11.5	46.9	50.8	42.6	55.2	18.1	**29.2**	15.0	17.0
GPTQ + absmax		14.1	11.5	41.1	42.0	25.7	29.9	10.9	14.6	13.3	12.8
GPTQ + PiSO	$\mathbf{X}$-$\mathbf{X}$	11.5	9.43	48.2	**55.0**	**45.3**	54.0	**21.3**	25.9	**19.8**	**19.3**
GPTQ $\odot$ PiSO	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	**11.4**	**9.41**	**50.6**	52.0	40.2	**55.6**	16.6	27.7	19.1	18.7

Group-wise Quantization. Tables 15 and 16 report group-wise quantization results, with group sizes 16 and 32, on Llama-3.2-3B and Qwen-2.5-3B, covering all combinations of independent/sequential objectives and decoupled/interleaved integration strategies (Sections 3.2 and 3.3) that are aggregated in the main text (Table 3). Table 17 shows the group-wise quantization results for Llama-3.1-8B and Qwen-2.5-7B for group size 16.

Table 15: **WikiText2 perplexity (↓)** results for group-wise quantization on Llama-3.2-3B and Qwen-2.5-3B with group sizes 16 and 32 (G16/G32). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, AWQ). In the "Var." column, I/S denotes independent/sequential group-wise heuristics (see Section 3.2).

			2-bit				**3-bit**				**4-bit**			
	Var.	Obj.	Llama 3B		Qwen 3B		Llama 3B		Qwen 3B		Llama 3B		Qwen 3B	
			G16	G32	G16	G32	G16	G32	G16	G32	G16	G32	G16	G32
BF16			7.14	7.14	7.88	7.88	7.14	7.14	7.88	7.88	7.14	7.14	7.88	7.88
RTN + absmax			8e5	8e5	3e8	2e8	4e3	11.7	21.0	102	7.69	7.72	8.78	8.91
RTN + OptQuant		**I-I**	8e5	8e5	3e8	3e8	10.1	13.0	38.0	76.1	7.49	7.65	9.30	8.69
RTN + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	6e4	5e4	388	758	8.38	8.95	8.98	9.37	7.35	7.46	8.08	8.18
RTN + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	21.5	32.6	18.0	29.4	8.01	8.44	8.63	9.01	7.28	7.37	8.00	8.05
RTN + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	2e4	2e4	612	774	8.21	8.62	8.85	9.13	7.33	7.42	8.09	8.18
RTN + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	19.4	28.7	17.1	25.9	7.86	8.17	8.58	8.92	7.26	7.33	8.00	8.04
AWQ			8e5	8e5	2e8	2e8	9.61	9.57	11.4	11.4	7.50	7.55	8.37	8.46
GPTQ + absmax			30.3	26.4	34.8	30.4	8.41	8.29	8.86	8.91	7.35	7.38	8.06	8.03
GPTQ + OptQuant		**I-I**	22.1	25.8	17.3	26.4	8.05	8.31	8.54	8.84	7.29	7.33	8.00	**8.01**
GPTQ + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	28.8	32.0	16.6	19.6	8.13	8.53	8.55	8.79	7.30	7.37	8.02	8.03
GPTQ + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	21.5	25.2	15.5	18.1	8.12	8.50	8.56	8.82	7.30	7.37	8.01	8.04
GPTQ ⊙ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	23.9	29.9	16.0	19.2	8.11	8.50	8.52	8.72	7.30	7.37	8.01	8.03
GPTQ ⊙ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	20.7	24.0	15.3	18.1	8.12	8.44	8.53	8.74	7.30	7.36	8.01	8.04
GPTQ ⊡ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	17.4	21.3	14.3	16.9	7.92	8.30	8.48	8.68	7.27	7.34	**7.98**	8.03
GPTQ ⊡ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	17.0	22.3	14.3	16.6	7.92	8.29	8.50	8.70	7.27	7.34	**7.98**	8.02
Qronos + absmax			22.6	24.5	77.3	21.1	8.07	8.08	8.67	8.66	7.32	7.31	8.03	8.03
Qronos + OptQuant		**I-I**	15.1	18.1	17.7	35.9	7.74	**7.91**	8.43	8.64	7.24	**7.27**	7.99	8.02
Qronos + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	23.6	28.3	22.0	23.7	7.75	7.95	8.44	8.62	7.25	7.29	**7.98**	8.02
Qronos + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.6	16.3	13.3	14.7	7.80	7.99	8.49	8.66	7.25	7.30	8.00	8.03
Qronos ⊙ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	15.5	17.9	13.6	15.1	7.76	7.96	8.44	8.61	7.25	7.29	7.99	8.03
Qronos ⊙ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.8	16.3	13.0	14.4	7.82	8.01	8.48	8.66	7.26	7.30	8.00	8.03
Qronos ⊡ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	**12.9**	15.4	**11.9**	**13.2**	**7.72**	7.94	**8.37**	**8.53**	**7.23**	7.28	**7.98**	**8.01**
Qronos ⊡ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	13.7	**15.2**	12.5	13.8	7.73	**7.91**	8.45	8.63	7.24	7.28	7.99	8.03

Table 16: **Average zero-shot accuracy** (↑) results for group-wise quantization on Llama-3.2-3B and Qwen-2.5-3B with group sizes 16 and 32 (G16/G32). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, AWQ). In the "Var." column, I/S denotes independent/sequential group-wise heuristics (see Section 3.2).

			2-bit				**3-bit**				**4-bit**			
	Var.	Obj.	Llama 3B		Qwen 3B		Llama 3B		Qwen 3B		Llama 3B		Qwen 3B	
			G16	G32	G16	G32	G16	G32	G16	G32	G16	G32	G16	G32
BF16			62.0	62.0	56.5	56.5	62.0	62.0	56.5	56.5	62.0	62.0	56.5	56.5
RTN + absmax			32.3	32.2	31.1	31.5	50.7	52.0	46.2	31.6	60.3	59.3	56.2	55.0
RTN + OptQuant		**I-I**	32.2	32.3	31.2	31.3	54.7	51.5	37.7	38.5	61.2	61.0	54.4	54.3
RTN + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	32.5	32.1	32.6	32.7	58.2	57.2	54.3	54.4	61.2	60.8	55.5	55.2
RTN + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	44.3	40.5	38.5	37.1	**59.8**	57.6	54.8	52.8	61.9	61.6	55.6	56.2
RTN + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	33.1	33.4	31.3	32.7	58.9	58.4	54.6	54.8	60.7	60.8	55.8	54.9
RTN + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	42.3	40.0	39.8	38.7	59.1	57.1	56.3	53.3	61.3	61.3	56.5	56.1
AWQ			32.1	32.2	31.6	32.6	55.5	54.9	52.5	51.5	61.0	60.7	54.8	54.7
GPTQ + absmax			39.5	37.9	34.1	33.1	57.8	57.6	53.6	51.7	60.2	60.5	56.2	**57.1**
GPTQ + OptQuant		**I-I**	40.3	37.9	40.9	34.8	59.5	57.5	55.4	54.1	61.8	**62.3**	55.6	54.9
GPTQ + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	37.6	37.6	41.1	**39.2**	57.8	57.7	54.5	54.6	61.7	61.1	56.5	56.6
GPTQ + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	40.3	38.4	39.6	34.5	59.4	57.2	52.9	51.3	61.7	61.2	55.6	55.9
GPTQ ⊙ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	39.7	37.5	38.2	38.3	58.1	57.7	55.3	55.5	61.6	61.1	55.9	55.6
GPTQ ⊙ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	40.9	38.5	40.0	34.9	57.9	55.9	54.1	**56.8**	61.1	61.0	56.1	56.0
GPTQ ⊡ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	45.6	40.4	42.2	37.2	58.8	58.2	54.2	55.9	**62.2**	61.0	56.2	56.9
GPTQ ⊡ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	44.0	41.1	39.0	37.5	59.5	58.8	54.5	55.4	**62.2**	61.2	56.1	56.4
Qronos + absmax			39.3	39.3	32.6	31.9	57.2	**58.9**	52.4	51.0	61.1	61.6	56.6	55.2
Qronos + OptQuant		**I-I**	44.5	41.5	31.9	35.5	58.5	57.0	**56.8**	51.3	61.9	61.5	56.0	55.8
Qronos + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	38.1	36.7	34.2	33.2	59.3	58.0	54.9	54.0	60.2	60.6	56.3	56.2
Qronos + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	43.0	43.0	40.1	33.5	58.5	57.5	53.8	52.9	61.2	60.9	56.4	56.5
Qronos ⊙ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	43.9	42.4	34.6	31.4	59.4	57.8	54.4	55.2	60.7	61.3	**57.4**	56.4
Qronos ⊙ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	43.8	41.7	38.3	31.6	58.5	57.9	54.5	51.9	61.7	60.8	56.8	55.6
Qronos ⊡ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	46.4	**43.7**	**44.3**	33.8	**59.8**	57.2	55.3	50.1	61.9	61.8	55.3	55.0
Qronos ⊡ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	**46.5**	41.9	41.1	34.7	58.1	57.9	52.4	54.1	61.1	61.7	55.9	56.3

Table 17: Group-wise quantization results on Llama 3.1-8B and Qwen-2.5-7B Instruct with group size 16, reporting WikiText2 perplexity (↓) and average zero-shot accuracy (↑). The "Obj." column denotes the scale optimization objective (see Equation 2): $\mathbf{X}$-$\mathbf{X}$ (float activations), $\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$ (self-activation), or $\mathbf{X}$-$\tilde{\mathbf{X}}$ (cross-activation); blank where not applicable (absmax, AWQ). In the "Var." column, I/S denotes independent/sequential group-wise heuristics (see Section 3.2).

			2-bit				**3-bit**				**4-bit**			
			WikiText2 (↓)		0-shot (↑)		WikiText2 (↓)		0-shot (↑)		WikiText2 (↓)		0-shot (↑)	
	Var.	Obj.	8B	7B	8B	7B	8B	7B	8B	7B	8B	7B	8B	7B
BF16			5.89	6.92	68.6	57.7	5.89	6.92	68.6	57.7	5.89	6.92	68.6	57.7
RTN + absmax			2e6	6e6	31.9	32.3	2e6	4e6	33.4	32.9	6.38	7.34	67.0	58.3
RTN + OptQuant		**I-I**	2e6	6e6	32.1	32.3	7.67	8.14	62.7	54.9	6.17	7.13	67.9	57.8
RTN + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	1e6	980	32.0	31.8	6.84	7.44	65.3	56.8	6.05	7.00	68.1	58.0
RTN + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	20.4	12.4	35.1	50.1	6.60	7.26	65.3	56.3	6.02	6.98	68.4	57.3
RTN + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	3e5	308	31.9	32.6	6.74	7.40	65.0	55.7	6.05	7.03	67.9	57.9
RTN + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	14.6	10.8	43.7	50.6	6.49	7.26	65.1	55.8	6.00	6.98	68.2	57.4
AWQ			2e6	6e6	32.0	32.2	2e6	3e6	42.8	32.0	6.29	7.30	66.6	58.4
GPTQ + absmax			22.9	12.7	43.2	36.0	6.79	7.44	65.8	56.5	6.08	7.04	67.8	**58.7**
GPTQ + OptQuant		**I-I**	20.4	10.8	38.0	46.9	6.59	7.27	66.2	56.8	6.02	6.98	67.6	57.1
GPTQ + PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	20.2	11.3	45.1	49.6	6.63	7.25	66.6	56.4	6.02	6.98	67.4	58.2
GPTQ + PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	14.8	10.7	46.2	**53.4**	6.63	7.25	66.5	56.8	6.04	6.98	67.7	57.2
GPTQ ⊙ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	17.5	11.6	46.3	32.4	6.59	7.25	**66.9**	55.4	6.02	6.98	67.7	58.0
GPTQ ⊙ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	15.7	10.6	48.2	52.3	6.64	7.27	65.8	56.2	6.03	6.98	**68.7**	56.3
GPTQ ⊡ PiSO	I	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	15.7	10.1	38.3	52.9	6.51	7.22	66.1	**56.9**	6.01	**6.97**	68.4	57.7
GPTQ ⊡ PiSO	S	$\tilde{\mathbf{X}}$-$\tilde{\mathbf{X}}$	15.6	10.1	44.5	52.3	6.52	7.23	65.8	55.7	6.01	**6.97**	68.4	57.4
Qronos + absmax			29.6	2e4	34.7	33.6	6.68	7.67	63.5	56.1	6.06	7.06	67.4	57.6
Qronos + OptQuant		**I-I**	12.1	3e3	43.7	32.3	6.40	7.26	66.4	55.2	5.99	**6.97**	68.1	57.3
Qronos + PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	21.4	17.6	37.1	37.9	6.42	7.22	66.0	56.8	5.99	**6.97**	67.4	57.7
Qronos + PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	11.4	10.5	48.2	50.7	6.45	7.23	65.0	54.7	6.00	6.98	68.0	57.3
Qronos ⊙ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	13.8	9.92	43.6	32.6	6.41	7.21	66.1	56.2	5.99	**6.97**	67.9	57.8
Qronos ⊙ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	11.9	10.4	47.8	52.4	6.46	7.22	65.2	55.3	6.00	6.98	67.0	57.9
Qronos ⊡ PiSO	I	$\mathbf{X}$-$\tilde{\mathbf{X}}$	**10.1**	**8.91**	36.4	52.2	**6.37**	**7.18**	66.4	**56.9**	**5.98**	**6.97**	67.3	57.2
Qronos ⊡ PiSO	S	$\mathbf{X}$-$\tilde{\mathbf{X}}$	10.9	25.5	**48.5**	39.6	6.40	8.76	65.4	51.0	5.99	7.39	67.8	56.5