SCALE SENSITIVITY IN LOW-BIT POST-TRAINING QUANTIZATION: CURVATURE OF THE QUANTIZATION ERROR LANDSCAPE

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization requires choosing a grid scale, and max-based choices can fail sharply at low bit-widths. For i.i.d. Gaussian weights and activations with sufficiently large effective rank, we prove uniform convergence of the normalized round-to-nearest loss to a Gaussian quantization objective. Its unique optimal scale decreases with the number of levels, while its relative-scale curvature decays approximately exponentially with bit-width. We verify the rank condition for wide random MLPs. Across five LLMs, GPTQ scale sensitivity likewise falls with bit-width: scale choice matters greatly at 2–3 bits but little from 6 bits onward. The Gaussian-optimal scale performs poorly on raw weights; after Hadamard incoherence processing, it matches the best searched rule from 3 bits onward but remains worse at 2 bits.

1 INTRODUCTION

Quantization has emerged as a key approach to reduce the computational cost and memory footprint of modern machine learning models. As these models grow larger, methods that compress them and accelerate inference without retraining have attracted considerable interest in the machine learning community. The success of small-scale language models (LLMs) that can be deployed locally, such as Llama 3.2 (Dubey et al., 2024), DeepSeek (Liu et al., 2024), and BitNet (Ma et al., 2025), has further intensified this interest.

Among various compression strategies, Post-training quantization (PTQ) is especially attractive because it can be applied to a trained model using only a small calibration set, without the need of retraining. It therefore offers a practical route to efficient deployment with modest additional overhead. Despite its widespread adoption, the theoretical understanding of PTQ remains limited. A wide variety of PTQ methods have been proposed, such as GPTQ Frantar et al. (2022) and AWQ Lin et al. (2024), but a mathematical analysis of these methods has only recently begun to emerge Tseng et al. (2024); Chen et al. (2026); Zhang et al. (2026). All of these, however, consider the quantization problem on a fixed grid. A key parameter shared by these algorithms is the quantization scale, which determines the spacing and range of the quantization grid. Although many rules have been proposed for selecting it Banner et al. (2019); Nagel et al. (2021), see also Table 5, it remains unclear which choices best preserve model performance and how sensitive quantization algorithms are to the quantization scale choice. Independently of how the scale is chosen, these methods are surprisingly accurate at 4 or 8 bits. At lower precision, however, performance degradation seems unavoidable Dettmers & Zettlemoyer (2023).

Preprocessing techniques can further reduce the degradation caused by quantization. A prominent example is Hadamard incoherence processing (HIP) (Tseng et al., 2024), which applies a random orthogonal transformation to the weights so as to spread their magnitudes evenly across coordinates. As a consequence, the entries of the transformed weights become approximately Gaussian.

In this work, we take a step towards a mathematical understanding of how scale choice affects these algorithms. We view PTQ as the problem of approximating the linear map of a network layer, in the layerwise reconstruction sense, at a fixed quantization granularity. Our main observation is that, for networks at initialization and under suitable conditions on the calibration data, the reconstruction loss under round-to-nearest (RTN) quantization, viewed as a function of the scale, is well

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.

approximated by a deterministic Gaussian quantization objective. We then analyze the Gaussian quantization problem, characterizing its loss landscape around the optimal scale and its dependence on the bit width. We use this analysis to shed light on the scale sensitivity of existing PTQ methods, especially when combined with HIP preprocessing.

Paper Roadmap. Section 2 reviews related work, and Section 3 introduces the notation and quantization setup. Section 4 establishes uniform convergence to a Gaussian limiting loss, characterizes its landscape, and verifies the convergence assumptions for randomly initialized MLPs with Gaussian data. Section 5 evaluates scale sensitivity in LLM quantization. Proofs and additional experiments appear in the appendix.

Contributions. **1) Uniform Gaussian limit.** Under Gaussian weight and effective-rank assumptions, we prove uniform convergence over scales of the normalized RTN reconstruction loss to a scalar Gaussian objective, with explicit finite-dimensional, high-probability bounds (Theorem 1). We verify the rank condition for wide MLPs at initialization (Theorem 5). **2) Optimal scales and sensitivity.** For an analytic extension of the limiting objective, we prove a unique, nondegenerate minimizer and a strictly decreasing optimal scale as quantizer resolution increases (Proposition 1), extending the discrete monotonicity of Na & Neuhoff (2018). We derive its scale-invariant curvature and numerically observe approximately exponential decay with bit-width. **3) Implications for LLM quantization.** Across several LLMs, GPTQ exhibits greater sensitivity to scale selection at 2–3 bits than at 6–8 bits, with local sensitivity trends consistent with the Gaussian model. HIP substantially improves the performance of the resulting analytic scale rule, which requires no scale search. We also introduce Proxy-static, an efficient GPTQ-aware scale-selection objective that performs strongly in our experiments, particularly without preprocessing.

2 Related Work

Layerwise post-training quantization. For scalability, PTQ often minimizes the layerwise reconstruction error $\|(W - \hat{W})X\|_F^2$ on a fixed quantization grid. GPTQ Frantar et al. (2022), building on Optimal Brain Quantization Frantar & Alistarh (2022) and OBS Hassibi & Stork (1992), quantizes columns sequentially while compensating errors through the remaining floating-point weights. GPTAQ Li et al. (2025), ResComp Li et al. (2026), and QRoNoS Zhang et al. (2026) modify reconstruction and error compensation. Orthogonal preprocessing (Chee et al., 2023; Tseng et al., 2024; Ashkboos et al., 2024; Liu et al., 2025), activation smoothing (Xiao et al., 2023), activation-aware weight scaling Lin et al. (2024), and magnitude reduction (Zhang et al., 2024) make weights and activations easier to quantize. In particular, QuIP Tseng et al. (2024) uses random orthogonal preprocessing to obtain incoherence-based error guarantees for adaptive rounding on a fixed grid. Scale selection remains often heuristic; we study its influence, especially at low precision, without changing the projection procedure. BRECQ Li et al. (2021) combines blockwise reconstruction with learned rounding based on AdaRound Nagel et al. (2020), improving low-bit accuracy at an optimization cost that can be prohibitively expensive for large models.

Scale and clipping selection. OmniQuant Shao et al. (2024) optimizes clipping parameters and equivalent transformations through blockwise reconstruction, avoiding per-weight rounding optimization. PiSO (Amboage et al., 2026) exploits the piecewise-quadratic objective to compute exact channelwise RTN scales, but reports collapse at 3 bits in the data-free setting. ACIQ (Banner et al., 2018) derives Gaussian and Laplacian clipping thresholds, applying them only to activations after finding no benefit for weights. LLAPQ Nahshan et al. (2021) studies the full-model task-loss landscape over quantization parameters. We instead study layerwise reconstruction, establish a scalar asymptotic limit under coordinatewise RTN, and empirically examine scale sensitivity for RTN and GPTQ.

Optimal uniform scalar quantization. Classical scalar quantization theory studies optimal quantizers for specific distributions. Max (Max, 1960) derived optimality conditions and numerical solutions, including for Gaussian sources. For symmetric uniform quantization, Na and Neuhoff (Na & Neuhoff, 2017; 2018) study distortion convexity in the grid scale and prove that the Gaussian-optimal scale decreases when two levels are added. We connect this theory to layerwise reconstruc-

tion through a rigorous asymptotic limit, derive explicit curvature formulas, and compare them with empirical estimates from real quantization landscapes.

3 Notation and Preliminaries

Notation. We use standard notation for sets, Gaussian distributions, and differentiability classes, with $\mathbb{N} = 1, 2, \ldots$ and $[n] = 1, \ldots, n$. Write Id_d for the identity matrix and $\mathbb{S}^{p-1}$ for the Euclidean unit sphere in $\mathbb{R}^p$. Vectors, including matrix rows $M_{i:}$, are viewed as columns. The norms $\|\cdot\|_{op}$, $\|\cdot\|_F$, and $\|\cdot\|_{\psi_2}$ denote the operator, Frobenius, and sub-Gaussian norms, respectively. For a nonzero positive semidefinite matrix M, the effective rank is given by

$$r_{\mathrm{eff}}(M) := \mathrm{Tr}(M)/\|M\|_{op}\,.$$

Symmetric rounding is $\lfloor t \rceil := \mathrm{sign}(t)\lfloor |t| + 1/2 \rfloor$, with $\mathrm{sign}(0) = 0$. We denote the standard Gaussian density and cumulative density function by φ and Φ. Subscripts on $\mathbb{E}, \mathbb{P}$ specify the variable integrated over; primes and subscripts denote ordinary and partial derivatives respectively, e.g., $f_{ssm} = \partial_m \partial_s^2 f$. Constants $\tilde{C}, C, c > 0$ may change from line to line.

Post-training quantization. Post-training quantization (PTQ) maps the weights of a trained neural network onto a low-bit grid without retraining, aiming to preserve its input–output behavior using a small calibration set. Let $\vartheta \in \mathbb{R}^d$ be a vector of full-precision weights, let $B \geq 2$ denote the target bit precision, and set $M_B := 2^{B-1} - 1$. We define the *symmetric quantization cone* $\Gamma_B \subset \mathbb{R}^d$ as the union of all symmetric uniform grids with $2^B - 1$ levels per coordinate:

$$\Gamma_B := \left\{ sv : s > 0,\ v \in \{-M_B, \ldots, M_B\}^d \right\}\,.$$

For a fixed scale $s > 0$, coordinatewise round-to-nearest (RTN) quantization is the Euclidean projection onto the corresponding grid:

$$\Pi_{s,M_B}(\vartheta) := s \left\lfloor \frac{\vartheta}{s} \right\rceil_B\,, \qquad \lfloor t \rceil_B := \begin{cases} -M_B & t < -M_B, \\ \lfloor t \rceil & |t| \leq M_B, \\ M_B & t > M_B. \end{cases}\,. \tag{1}$$

where $\lfloor \cdot \rceil_B$ is applied componentwise. We write $\mathcal{Q}_B := \{\Pi_{s,M_B} : s > 0\}$ for the resulting family of quantization maps.

We study layerwise PTQ through the empirical squared L_2 distance between a full-precision linear map and its quantized approximation, evaluated on the same input activations. Let $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ denote the input space, equipped with a probability measure μ, and let $\mathbb{D}_{\mathrm{cal}} = \{X'_j\}_{j=1}^{N_c}$ consist of N_c i.i.d. samples from μ. For a layer with weights $\vartheta^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, we consider the RTN scale-selection problem of minimizing, over $s > 0$, the reconstruction error

$$\mathcal{E}^{(\ell)}(s, \vartheta^{(\ell)}, X^{\ell-1}) := \left\| \left(\vartheta^{(\ell)} - \Pi_{s,M_B}(\vartheta^{(\ell)})\right) X^{\ell-1} \right\|_F^2\,, \tag{2}$$

where the quantization map acts entrywise. Here, $X^{\ell-1} \in \mathbb{R}^{d_{\ell-1} \times N_c}$ denotes the input activation matrix used to calibrate layer ℓ. Its columns are obtained by propagating the calibration samples through the preceding layers, which may be full-precision or already quantized, depending on the PTQ procedure. We hold $X^{\ell-1}$ fixed when optimizing the scale of the current layer. We write $\mathcal{E}^{(\ell)}(s)$ when the dependence on $\vartheta^{(\ell)}$ and $X^{\ell-1}$ is clear.

This formulation also covers query, key, value, and output projections in attention, using their respective input activations. While (2) uses one scale per layer, other granularities are commonly used. For instance, *per-channel* quantization assigns one scale per channel of $\vartheta^{(\ell)}$, yielding d_ℓ independent row-wise problems. Finer granularities, such as *group quantization*, are also used.

A simple scale-selection rule is the symmetric min-max choice $s^e := \frac{\max_{i,j} |\vartheta_{ij}^{(\ell)}|}{M_B}$, with the maximum taken over the relevant channel or group for finer granularities. This choice uses only the weight range and does not optimize (2). Our experiments show that it can perform well at higher precision, while becoming substantially less reliable at lower bit-widths.

4 THEORETICAL RESULTS

In this section, we study how sensitive the reconstruction error (2) is to the quantization scale s across different bit-widths B. We focus on one output channel and write $\vartheta \in \mathbb{R}^d$ for its full-precision weight vector, where $d = d_{\ell-1}$ is the input dimension. Let $X \in \mathbb{R}^{d\times N_c}$ denote the corresponding calibration activations. For $\mathrm{tr}(XX^\top) > 0$, we consider the normalized reconstruction loss

$$\mathcal{E}(s, \vartheta, X) = \mathcal{E}(s) := \frac{\|\vartheta^\top X - \Pi_{s,M_B}(\vartheta)^\top X\|_2^2}{\mathrm{tr}(XX^\top)} = \frac{\big(\vartheta - \Pi_{s,M_B}(\vartheta)\big)^\top XX^\top \big(\vartheta - \Pi_{s,M_B}(\vartheta)\big)}{\mathrm{tr}(XX^\top)}.$$

The normalization by $\mathrm{tr}(XX^\top)$ removes the overall scale of the calibration activations and, under our assumptions, yields a loss of order one as d grows. For instance, when $XX^\top = \mathrm{Id}_d$, the normalized loss reduces to $\mathcal{E}(s) = \frac{1}{d}\|\vartheta - \Pi_{s,M_B}(\vartheta)\|_2^2$, the mean squared approximation error per coordinate. For fixed $s > 0$ and bit-width B, this admits a nontrivial limit as $d \to \infty$ when $\vartheta \sim \mathcal{N}(0, \mathrm{Id}_d)$. In the rest of this section, we assume that

- (H1) the parameters $\vartheta \overset{iid}{\sim} \mathcal{N}(0, 1)$;
- (H2) the effective rank of $XX^\top$ satisfies $r_{\text{eff}}(XX^\top) \geq C \log^8 d$ for some constant $C > 0$ not dependent on d.

Assumptions (H1) is standard in the literature on wide random neural networks. (H2) asks that the inputs to the group of parameters being quantized are not concentrated on a small number of directions. Section A.4 establishes (H2) with high probability for an MLP at initialization under the stated assumptions on its architecture and isotropic input distribution.

We study the asymptotic limit of $s \mapsto \mathcal{E}(s)$ to gain insight into its loss landscape. In Section 4.1, we establish a high-probability bound yielding uniform convergence in s to the deterministic limiting objective

$$\mathcal{E}^\infty(s) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\big[(Z - \Pi_{s,M_B}(Z))^2\big].$$

We then analyze the landscape of $s \mapsto \mathcal{E}^\infty(s)$ in Section 4.2.

4.1 CONVERGENCE TO THE LIMIT

Since our goal is to study the loss as a function of s, we seek an approximation that holds simultaneously over all $s > 0$. Such a uniform guarantee also applies when s is chosen based on the realization of ϑ and the calibration data X. The following theorem provides this guarantee with an explicit error bound.

Theorem 1. *Let $B \geq 2$, $X \in \mathbb{R}^{d\times N_c}$ be nonzero matrix satisfying (H2), and ϑ satisfying (H1). Then, for all $d \geq 2$, it holds*

$$\mathbb{P}\left(\sup_{s>0} |\mathcal{E}(s) - \mathcal{E}^\infty(s)| > \frac{2 + 12M_B}{\log d}\right) \leq C\, M_B\, d\, e^{-c\log^2 d},$$

where $c, C > 0$ are absolute constants independent on d.

A detailed proof of Theorem 1 appears in Appendix A.2. For fixed B, the theorem bounds the approximation error by a constant multiple of $(\log d)^{-1}$ with probability tending to one as $d \to \infty$. In particular,

$$\sup_{s>0} |\mathcal{E}(s) - \mathcal{E}^\infty(s)| \xrightarrow[d\to\infty]{\mathbb{P}} 0.$$

Let $m := \inf_{s>0} \mathcal{E}(s)$ and $m^\infty := \inf_{s>0} \mathcal{E}^\infty(s)$ denote the optimal empirical and limiting reconstruction losses. We note that both m and m^∞ are finite, since $\mathcal{E}, \mathcal{E}^\infty \geq 0$. Uniform convergence immediately implies $m \xrightarrow{\mathbb{P}} m^\infty$. Convergence of minimizers is established in Corollary 1 in Appendix A.3.

Remark 1. The uniform approximation of Theorem 1 relies on (H2), namely a lower bound on the effective rank of $XX^\top$. In Appendix A.4 we show that this condition holds, with high probability, in the setting of a random multi-layer perceptron with odd activation functions and isotropic inputs.

4.2 LIMITING FUNCTION LANDSCAPE

The uniform convergence established in Theorem 1 motivates the study of the limiting functional

$$\mathcal{E}^{\infty}(M_B, s) := \int_{\mathbb{R}} \big(s\ \lfloor t/s \rceil_B - t\big)^2 \varphi(t)\, dt\,. \tag{3}$$

Since our goal is to understand the loss landscape of $\mathcal{E}^{\infty}$ as a function of the bit precision B, which enters only through M_B, we make this dependence explicit in the notation, that is from now on we write $\mathcal{E}^{\infty}(M_B, s)$ in place of $\mathcal{E}^{\infty}(s)$ used above. We start by noting that $\mathcal{E}^{\infty}(M_B, s)$ splits into the rounding error of an infinite grid, which depends on s only, and a nonnegative clipping term, which is the only part through which the bit precision B enters. More precisely, $\mathcal{E}^{\infty}(M_B, s) = 1 - 4s\sum_{j=0}^{\infty} g\big((j+\frac{1}{2})s\big) + 4s\sum_{j=0}^{\infty} g\big((M_B+j+\frac{1}{2})s\big)$, where $g(x) := \varphi(x) - x\,(1-\Phi(x))$. The derivation of this expression, together with the proof of its well-posedness, is deferred to Corollary 2 in Appendix A.5. For the analysis that follows it is convenient to let the second argument vary continuously. We therefore consider the extension $\bar{\mathcal{E}}^{\infty} : [1,\infty)\times(0,\infty) \to (0,\infty)$ of $\mathcal{E}^{\infty}$, defined by $\bar{\mathcal{E}}^{\infty}(m, s) := 1 - 4s\sum_{j=0}^{\infty} g\big((j+\frac{1}{2})s\big) + 4s\sum_{j=0}^{\infty} g\big((m+j+\frac{1}{2})s\big)$, so that $\bar{\mathcal{E}}^{\infty}(M_B, s) = \mathcal{E}^{\infty}(M_B, s)$ for every $B \geq 2$.

The following proposition shows that $\bar{\mathcal{E}}^{\infty}$ is smooth, that for every m the function $s \mapsto \bar{\mathcal{E}}^{\infty}(m, s)$ has a unique critical point, which is a nondegenerate global minimizer $s^*(m)$, and that the optimal scale $s^*(m)$ is a smooth, strictly decreasing function of m, bounded below by $\sqrt{2}/(m+\frac{1}{2})$.

Proposition 1. *The following properties hold:*

(i) For every $B \geq 2$, the function $\bar{\mathcal{E}}^{\infty}(m, s)$ is smooth, e.g. $\bar{\mathcal{E}}^{\infty}(m, s) \in \mathcal{C}^{\infty}([1,\infty)\times(0,\infty))$.

(ii) For every real $m \geq 1$, the function $s \mapsto \bar{\mathcal{E}}^{\infty}(m, s)$ has a unique critical point on $(0,\infty)$. This point is its unique global minimizer $s^*(m)$ and is nondegenerate:

$$\partial_s^2 \bar{\mathcal{E}}^{\infty}\big(m, s^*(m)\big) > 0.$$

(iii) The map $m \mapsto s^*(m)$ is continuously differentiable and strictly decreasing on $[1,\infty)$. Moreover, for every $m \geq 1$,

$$\frac{ds^*}{dm}(m) = -\frac{\bar{\mathcal{E}}^{\infty}_{ms}\big(m, s^*(m)\big)}{\bar{\mathcal{E}}^{\infty}_{ss}\big(m, s^*(m)\big)} < 0, \qquad s^*(m) > \frac{\sqrt{2}}{m+\frac{1}{2}}.$$

Proposition 1 describes the landscape of the limiting energy $\bar{\mathcal{E}}^{\infty}(m, \cdot)$ as a function of the scale s for a fixed level parameter m. Item (ii) shows that it has a unique critical point, which is a nondegenerate global minimum. Item (iii) shows that increasing the number of levels strictly decreases the optimal scale $s^*(m)$. However, at fixed m, decreasing the scale trades finer resolution for a narrower representable range and hence greater clipping error. The optimal scale balances these effects and satisfies $s^*(m) > \sqrt{2}/(m+\frac{1}{2})$.

Beyond existence and uniqueness of the optimal scale $s^*(m)$, we are interested in how sensitive the energy is to scale choice. Since the optimal scale changes with the number of levels, we consider relative rather than absolute scale perturbations. The ordinary second derivative measures sensitivity to additive changes in s; multiplying it by s^2 instead measures sensitivity to proportional changes.

Definition 1. *Let $m \in [1,\infty)$ be such that $\bar{\mathcal{E}}^{\infty}(m, \cdot)$ is twice differentiable at $s^*(m)$. We define its* scale-invariant curvature *by $\kappa(m, s^*(m)) := s^2 \bar{\mathcal{E}}^{\infty}_{ss}(m, s^*(m))$.*

At the minimizer $s^* = s^*(m)$, a relative perturbation $s = s^*(1+\varepsilon)$ gives $\bar{\mathcal{E}}^{\infty}(m, s^*(1+\varepsilon)) - \bar{\mathcal{E}}^{\infty}(m, s^*) = \frac{1}{2}\kappa(m, s^*)\varepsilon^2 + o(\varepsilon^2)$. We invite the reader to Remark A.5.2 for a more detailed discussion on the role of $s^(*)(m)$ in the definition of κ. Thus, $\kappa(m, s^*)$ quantifies the local increase in reconstruction loss caused by a relative error in the scale. An explicit computation of $\kappa(m, s^*(m))$, for an integer $m > 0$, leads to

$$\kappa(m, s) = 4s\sum_{j=0}^{m-1} \Big[2\big(j+\tfrac{1}{2}\big)s\Big(1-\Phi\big((j+\tfrac{1}{2})s\big)\Big) - \big(j+\tfrac{1}{2}\big)^2 s^2\, \varphi\big((j+\tfrac{1}{2})s\big)\Big]\,.$$

The derivation of this formula can be found in Appendix A.5.1. A numerical evaluation for $B = 2, \ldots, 8$ shows that $\kappa(M_B, s^*(M_B))$ is exponentially decreasing with B as shown in Figure 4. We refer to Algorithm 1 in Appendix A.5.1 for the algorithm we have used for the numerical evaluation.

This shows that the loss landscape is sharper near its minimum at low bit-widths and flatter at high bit-widths, when measured with respect to relative changes in the quantization scale.

5 Experiments

The previous section showed that, under suitable assumptions, the layerwise PTQ reconstruction loss converges to the simpler asymptotic loss (3). We now test whether predictions from this limiting landscape carry over to trained models. In particular, we want to answer the following questions: **1)** Does the greater local scale sensitivity of the limiting objective at lower precision translate into greater sensitivity to scale choice under GPTQ? **2)** Does the analytic scale $s = s^*(M_B)\hat{\sigma}$ remain near-optimal for the reconstruction loss of trained, finite-width layers? Here, $\hat{\sigma}$ is the RMS of the weights in the channel being quantized, and $s^*(M_B)$ is computed by solving $\mathcal{E}_s^\infty(M_B, s^*(M_B)) = 0$, using the explicit expression in (28).

5.1 Scale sensitivity of LLM post-training quantization

We quantize several LLMs using GPTQ, with a calibration set of 128 randomly sampled windows of 2,048 tokens from the C4 training split. We use per-channel quantization and vary only the scale-selection rule. Our theory concerns coordinatewise RTN, whereas GPTQ updates the remaining weights to compensate for quantization errors. The two procedures coincide on the same fixed grid when $XX^\top$ is diagonal, but GPTQ is not generally covered by our theory. These experiments therefore test whether the predicted scale sensitivity extends to GPTQ. We consider five different scale-selection rules, ranging from no search (Min-max) through search without activation data (WMSE, Shrink-2.4) to Hessian-weighted search (Proxy-static, RTNH), as specified in Table 5.

Remark 2. Existing scale rules target the RTN loss, which need not align with GPTQ's compensated loss without preprocessing. Our *Proxy-static* rule, see Appendix B.1.3, reuses GPTQ's inverse-Hessian factorization to account for the compensation: it is the best rule without HIP (Table 1) but gives no clear gain with HIP (Table 2).

For each model and precision, Figure 1 reports the performance range: the difference between the maximum and minimum NLL across the five scale-selection rules. This measures how much the quantized model performance depends on the choice of scale-selection rule. We observe, that from

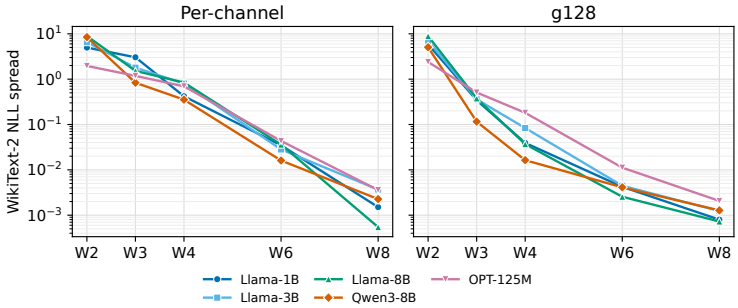


Figure 1: Log-scale GPTQ sensitivity to the scale rule across bit per channel (left) and group size 128 (right) for different models. NLL spread is the max minus min WikiText-2 NLL across Min-max, Shrink-2.4, WMSE, Proxy-static, and RTNH.

W6 onward, all scale rules give essentially the same NLL. The scale matters most at W2 and W3, consistent with κ decreasing exponentially in the number of bits, i.e., the theoretical loss landscape being more sensitive to relative scale perturbations at low precision. Amboage et al. (2026) reports a related trend, i.e. data-aware scale optimization pays off increasingly as the bit-width decreases. We note that, for group size 128, the performance differences between scale-selection rules diminish

even more rapidly with bit-width, consistent with the greater flexibility of fitting scales to smaller groups of weights.

Remark 3. The NLL spread across scale-selection rules reflects both the model's sensitivity to scale changes and the differences between the scales selected by the rules. It shows the practical importance of scale selection in low-precision settings, but does not establish higher curvature of the local GPTQ reconstruction loss. Appendix B.2.4 complements this comparison by measuring sensitivity under fixed relative scale perturbations.

Table 1: Llama-3.1-8B, W3 per-channel GPTQ without HIP. Lower perplexity and higher accuracy indicate better performance. Proxy-static, the GPTQ-aware scale-selection rule introduced in Appendix B.1.3, consistently outperforms the other tested rules in this setting.

Rule	WikiText-2 PPL	C4 PPL	PIQA	ARC-E	ARC-C	HellaSwag	WinoGrande	BoolQ	Mean
Shrink-2.4	14.050	17.177	0.705	0.553	0.369	0.708	0.666	0.795	0.632
Proxy-static	**11.756**	**16.328**	**0.729**	**0.635**	**0.412**	**0.721**	**0.685**	**0.798**	**0.663**
Min-max	41.473	45.506	0.624	0.408	0.282	0.472	0.515	0.524	0.471
Analytic	930,372.777	643,344.609	0.514	0.254	0.268	0.267	0.507	0.622	0.405

Table 1 contrasts the best and worst of the five heuristic scale-selection rules for Llama-3.1-8B at W3 without HIP, together with zero-shot accuracy on six benchmarks to assess whether perplexity differences carry over to downstream tasks. To address the second question, we also include the analytic scale $s^*(M_B)\hat{\sigma}$.

5.2 Preprocessing and scale sensitivity

Scale selection strongly affects performance, and the asymptotically optimal scale fails on the unprocessed model. Since trained weights need not be iid Gaussian, we ask whether preprocessing can preserve the model's function while bringing its weights closer to the assumptions of Theorem 1 and making $s^*(M_B)\hat{\sigma}$ near-optimal for the finite-width reconstruction objective.

Figure 2 shows that, after HIP (see Appendix B.4 for a short description), the analytic rule $s^*(M_B)\hat{\sigma}$ consistently matches the best tested heuristic in NLL across the tested models at W3 and above, without activation-based scoring or scale search.

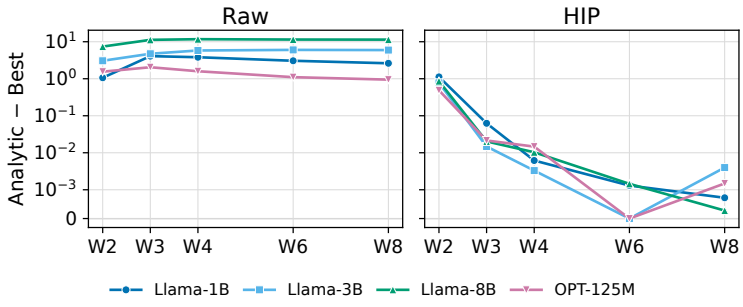


Figure 2: HIP substantially reduces the NLL gap between the analytic rule $s^*(M_B)\hat{\sigma}$ and the best of five heuristics under per-channel GPTQ across tested models and precisions. We report the positive part of this gap, setting it to zero whenever the analytic rule matches or outperforms all five heuristics.

Table 2: Llama-3.1-8B, W3 per-channel GPTQ with HIP. Lower is better for perplexity; higher is better for accuracy. Best results are shown in bold.

Rule	WikiText-2 PPL	C4 PPL	PIQA	ARC-E	ARC-C	HellaSwag	WinoGrande	BoolQ	Mean
Shrink-2.4	**9.095**	**14.452**	0.785	0.744	**0.508**	**0.740**	0.717	0.817	0.718
Proxy-static	9.229	14.618	0.781	0.757	0.488	0.739	**0.733**	0.817	0.719
Min-max	15.396	22.775	0.705	0.544	0.332	0.620	0.595	0.621	0.569
Analytic	9.214	14.602	**0.786**	**0.761**	0.506	0.737	0.711	**0.824**	**0.721**

QuIP# Tseng et al. (2024) likewise selects the scales of its vector codebooks from the Gaussian quantization error. Our analysis makes this connection rigorous for uniform scalar quantization of the layerwise reconstruction objective, and our experiments show that the resulting analytic scale transfers, after HIP, to GPTQ (and to QRoNoS and ResComp, see the appendix) with competitive performance and no scale search.

5.3 Local reconstruction sensitivity

We now examine scale sensitivity directly at the level of the reconstruction objective. For each weight matrix, we evaluate 32 predetermined rows using the calibration activations from a sequential GPTQ run. We repeat the experiment across five models, five precisions, and three calibration draws.

For each row, let $\hat{s}$ denote the best searched GPTQ scale and $L(s)$ its RMS-normalized, Hessian-weighted reconstruction loss. We measure sensitivity to relative scale perturbations through

$$C_{\mathrm{sym}}(\delta) := \frac{L(\hat{s}(1+\delta)) + L(\hat{s}(1-\delta)) - 2L(\hat{s})}{\delta^2}. \tag{4}$$

For a twice-differentiable loss, this converges to $\hat{s}^2 L''(\hat{s})$ as $\delta \to 0$. For the empirical GPTQ landscapes, we interpret it as a finite-window sensitivity measure.

Figure 3 shows that the median empirical GPTQ sensitivity[1] decreases approximately exponentially with bit-width over the tested range, consistently across all three perturbation windows. The theoretical Gaussian curvature exhibits the same trend, and at $\delta = 0.05$, the empirical and theoretical quantities also agree closely in magnitude across models. This behavior mirrors the global sensitivity trend in Figure 1. Together, these observations suggest that the Gaussian landscape provides a useful model for the bit-width dependence of GPTQ sensitivity even without HIP preprocessing, despite substantial differences between the empirically optimal scales and their Gaussian predictions.

Figure 15 illustrates the reconstruction landscape for one weight row across precisions, with and without HIP preprocessing. Figure 13 further shows that median empirical GPTQ sensitivity after HIP follows a similar bit-width dependence to that observed without preprocessing. This suggests that the reduced performance spread across scale-selection rules after HIP could reflect better agreement among the scales selected by these rules, and not a flatter local landscape.

Note that, HIP's rotation alone preserves the effective rank of the input Gram matrix $H = XX^\top$. Motivated by the tighter convergence bound at higher effective rank (Theorem 1), we test whether transformations that change the spectrum improve analytic scale prediction.

We study W2 per-channel GPTQ on Llama-3.1-8B-Instruct, using a single unrotated sequential quantization stream. We partition the model's 32 blocks into ten fixed depth bins. Within each bin, we select the matrix with the largest normalized reconstruction error after per-row scale search, then its highest-error row. We compare four settings: no preprocessing, HIP, diagonal scaling followed by HIP, and whitening followed by HIP (B.4). All settings use the same pre-target weights and incoming activations; preprocessing is applied only to the target.

Relative to HIP alone, diagonal scaling and whitening reduce the median scale gap $\left|\log\left(\frac{s_{\mathrm{search}}}{s_{\mathrm{an}}}\right)\right|$, where $s_{\mathrm{an}} = s^*(M_2)\hat{\sigma}$, from 0.187 to 0.0298 and 0.00406, improving agreement in $8/10$ and

[1]For each model and bit-width, we pool observations from 32 predetermined rows per matrix across all studied matrices and three calibration seeds.

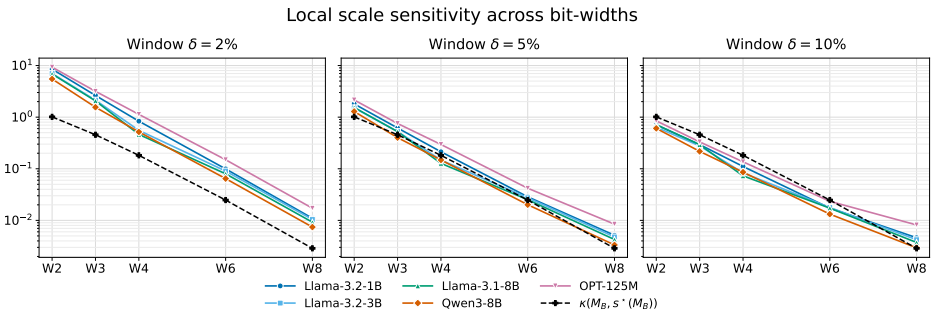


Figure 3: Colored curves: median finite-window GPTQ sensitivity $C_{\mathrm{sym}}(\delta)$ in logscale at rowwise scales optimized by grid search for different δ. Dashed black: Gaussian curvature $\kappa(M_B, s^\star(M_B))$ computed using Algorithm 1. Same exponential decay with number of bits as the empirical GPTQ sensitivity. The dependence of $C_{\mathrm{sym}}(\delta)$ on the perturbation window indicates that these finite-window measurements cannot be interpreted as a single local quadratic curvature.

$10/10$ selected rows, respectively (Figure 10). Here, s_{search} is the best searched GPTQ scale, not a certified global optimum. However, using a common normalization across settings, the best searched reconstruction error increases in every row, by median factors of 3.63 and 20.8 relative to HIP alone (Table 9).

These results reveal a trade-off in the selected cases: the transformations bring the best searched scale closer to the analytic prediction, but worsen the reconstruction error attained by GPTQ even after scale optimization.

6 Limitation and Future Work

We establish the first rigorous connection between Gaussian scalar quantization and layerwise post-training quantization at random initialization. Under Gaussian weights and an effective-rank condition on calibration activations, the normalized round-to-nearest reconstruction loss converges uniformly in scale to the Gaussian objective (Theorem 1); this condition holds with high probability for wide MLPs (Theorem 5). The objective has a unique, nondegenerate minimizer whose scale-invariant curvature decreases with bit-width (Proposition 1; Section 4.2), explaining strong scale sensitivity at 2–3 bits and weaker sensitivity from 6 bits onward, consistent with LLM experiments. The analysis also shed light on why HIP improves quantization: most scale-selection rules target near-optimal scales for Gaussian weights, and HIP brings weights closer to this regime. We stress the fact that convergence is not guarantee but Gaussianity of the weights alone for iterative PTQ. Finally, our experiments reveal that naively increasing the effective rank improves scale predictability but removes low-eigenvalue directions that GPTQ exploits for error compensation.

Limitations & Future works. Our theory concerns randomly initialized networks and does not model training or its effects on weight distributions and activation geometry. We therefore do not establish whether the Gaussian and effective-rank assumptions hold after training; the experiments on pretrained LLMs provide empirical evidence beyond our theoretical guarantees. Moreover, the analysis is specific to round-to-nearest (RTN) quantization and does not cover other strategies, including error-compensation methods such as GPTQ. Extending the theory to trained networks and more general quantization methods remains an open direction. In our experiments HIP makes weight distributions more Gaussian, improving alignment with (H1) and GPTQ performance, whereas the transformations we tested to increase the effective rank, motivated by (H2), degrade performance when combined with HIP. This calls for preprocessing that improves Gaussianity and effective rank jointly without sacrificing GPTQ's error compensation. On the theoretical side, extending our analysis to trained weights and to compensated methods such as GPTQ would clarify how scale selection, spectral structure and error compensation interact. Finally, the approximately exponential decay of relative-scale sensitivity in the Gaussian model provides a quantitative reference for future PTQ methods: a natural goal is lower sensitivity at low precision, under matched bit-widths and loss normalization, without loss of reconstruction quality or end-to-end performance.

AI USE STATEMENT

In this work, we used generative AI tools for polishing code, improving sentence clarity, and refining grammar. Moreover, discussion with AI has been used to provide insight on how to prove positiveness and decreasing behaviour of (ii) and (iii) in Proposition 1. The conversation with AI tools has not produced the proof but just the idea, that has been carefully written by the authors. Therefore, we have not used generative AI tools for creating results and proof in a copy and paste manner. We have carefully reviewed all AI-assisted work: in particular, the LLM-generated part of the code used in the experiments was verified and tested for correctness. We take responsibility for the final content of this work, including text and claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on the theoretical analysis of scale sensitivity and quantization objective in machine learning model quantization and does not involve experiments on human subjects, sensitive personal data, or applications with direct societal risks. The datasets referenced are publicly available, and no private or restricted data was used. Potential ethical concerns related to misuse are minimal, as the contributions are mainly theoretical.

REPRODUCIBILITY STATEMENT

We have taken multiple steps to ensure reproducibility of our results. All theoretical claims are accompanied by rigorous proofs, presented in detail in the appendix. Assumptions underlying the theorems are explicitly stated, and definitions are given in full to allow independent verification. The code used for the experiment is provided as anonymous supplementary material at `https://anonymous.4open.science/r/SOQ-345F`.

REFERENCES

Juan Amboage, Pablo Monteagudo-Lago, Ian Colbert, Giuseppe Franco, and Nicholas Fraser. Optimal post-training quantization scales and where to find them. *arXiv preprint arXiv:2606.10890*, 2026.

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

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

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32, 2019.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36:4396–4429, 2023.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of llm quantization: Gptq as babai's nearest plane algorithm. In *International Conference on Learning Representations*, volume 2026, pp. 122653–122695, 2026.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pp. 7750–7774. PMLR, 2023.

Abhimanyu Dubey, Abhinav Jauhri, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. URL `https://doi.org/10.48550/arXiv.2407.21783`.

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. *arXiv preprint arXiv:2210.17323*, 2022.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

Shuaiting Li, Juncan Deng, Kedong Xu, Rongtao Deng, Hong Gu, Minghan Jiang, Haibin Shen, and Kejie Huang. Rethinking residual errors in compensation-based llm quantization. *arXiv preprint arXiv:2604.07955*, 2026.

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

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. Gptaq: Efficient finetuning-free quantization for asymmetric calibration. *arXiv preprint arXiv:2504.02692*, 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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 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 *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025.

Shuming Ma, Hongyu Wang, Shaohan Huang, Xingxing Zhang, Ying Hu, Ting Song, Yan Xia, and Furu Wei. Bitnet b1. 58 2b4t technical report. *arXiv preprint arXiv:2504.12285*, 2025.

Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960.

Sangsin Na and David L Neuhoff. On the convexity of the mse distortion of symmetric uniform scalar quantization. *IEEE Transactions on Information Theory*, 64(4):2626–2638, 2017.

Sangsin Na and David L Neuhoff. Monotonicity of step sizes of mse-optimal symmetric uniform scalar quantizers. *IEEE Transactions on Information Theory*, 65(3):1782–1792, 2018.

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.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

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

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

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Proceedings of machine learning research*, 235:48630, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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

Aozhong Zhang, Naigang Wang, Yanxia Deng, Xin Li, Zi Yang, and Penghang Yin. Magr: Weight magnitude reduction for enhancing post-training quantization. *Advances in neural information processing systems*, 37:85109–85130, 2024.

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In *International Conference on Learning Representations*, volume 2026, pp. 87775–87798, 2026.

APPENDIX

A Theory **13**

A.1 Known results . . . 13
A.2 Proofs of Theorem 1 . . . 16
A.3 Proof of Corollary 1 . . . 21
A.4 Verifying (H2) for Multi-Layer Perceptrons at initialization . . . 22
A.5 Proof of Proposition 1 . . . 28
A.5.1 Computation of $\kappa(m)$ for $m > 0$ integer . . . 37
A.5.2 Why relative curvature κ? . . . 38
A.6 Technical Lemmas . . . 39

B Experiments **40**

B.1 Experimental Details . . . 41
B.1.1 Models . . . 41
B.1.2 Quantization setup . . . 41
B.1.3 Scale-selection rules . . . 42
B.1.4 Compute resources and software . . . 43
B.2 Global scale-sensitivity . . . 43
B.2.1 Sensitivity to scale selection with HIP preprocessing . . . 44
B.2.2 Sensitivity to scale selection with GPTQ extensions . . . 45
B.2.3 Scale sensitivity without min-max . . . 46
B.2.4 End-to-end sensitivity to controlled scale perturbations . . . 46
B.2.5 Calibration-Set Size . . . 47
B.3 Zero-shot evaluation . . . 47
B.4 Preprocessing and effective rank . . . 48
B.5 Quantization error landscapes . . . 52
B.5.1 Gaussian landscapes and convergence . . . 52
B.6 Local Reconstruction Sensitivity with HIP . . . 52

A THEORY

A.1 KNOWN RESULTS

In this section we collect the probabilistic tools used throughout the paper. Throughout the paper the sub-gaussian and sub-exponential norms are understood in the following Orlicz sense, which is the one used in Vershynin (2018). Most of the results appearing in this section are taken from Vershynin (2018).

Definition 2. *For a real random variable Y we set*

$$\|Y\|_{\psi_2} := \inf\left\{u > 0 : \ \mathbb{E}\left[e^{Y^2/u^2}\right] \leq 2\right\},$$

with the convention $\inf \emptyset = +\infty$*, and we call Y sub-gaussian if* $\|Y\|_{\psi_2} < \infty$*.*

Definition 3. *For a real random variable Y we set*

$$\|Y\|_{\psi_1} := \inf\left\{u > 0 : \ \mathbb{E}\left[e^{|Y|/u}\right] \le 2\right\},$$

with the convention $\inf \emptyset = +\infty$, and we call Y sub-exponential if $\|Y\|_{\psi_1} < \infty$.

Gaussian random variables are sub-gaussian (Vershynin, 2018, Ex. 2.5.8); since the value of their ψ_2-norm enters the constants of our bounds, we compute it exactly.

Lemma 1. *If $Y \sim \mathcal{N}(0, \tau^2)$ with $\tau > 0$, then $\|Y\|_{\psi_2} = \sqrt{8/3}\,\tau$. Moreover, $\|aY\|_{\psi_2} = |a|\,\|Y\|_{\psi_2}$ for every $a \in \mathbb{R}$.*

Proof. For $u > \sqrt{2}\,\tau$ one has $\mathbb{E}[e^{Y^2/u^2}] = (1 - 2\tau^2/u^2)^{-1/2}$, and $\mathbb{E}[e^{Y^2/u^2}] = +\infty$ otherwise. Hence $\mathbb{E}[e^{Y^2/u^2}] \le 2$ if and only if $1 - 2\tau^2/u^2 \ge 1/4$, i.e. $u^2 \ge \frac{8}{3}\tau^2$. The homogeneity is immediate from the definition. □

The square of a sub-gaussian random variable is sub-exponential, with an exact relation between the two norms.

Lemma 2. *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

The following lemma shows that centering does not increase the sub-exponential norm by more than a constant factor, which we make explicit.

Lemma 3. *Let X be a sub-exponential random variable. Then $X - \mathbb{E}X$ is sub-exponential as well, and*

$$\|X - \mathbb{E}X\|_{\psi_1} \ \le \ \left(1 + \frac{1}{\log 2}\right)\|X\|_{\psi_1} \ \le \ 3\,\|X\|_{\psi_1}.$$

Proof. Since $\|\cdot\|_{\psi_1}$ is a norm, the triangle inequality gives $\|X - \mathbb{E}X\|_{\psi_1} \le \|X\|_{\psi_1} + \|\mathbb{E}X\|_{\psi_1}$. For a deterministic $a \in \mathbb{R}$ we have $e^{|a|/t} \le 2$ if and only if $t \ge |a|/\log 2$, hence $\|a\|_{\psi_1} = |a|/\log 2$. It remains to bound $|\mathbb{E}X|$ by $\|X\|_{\psi_1}$. Let $t > \|X\|_{\psi_1}$, so that $\mathbb{E}e^{|X|/t} \le 2$. Using $e^x \ge 1 + x$ we get $2 \ge \mathbb{E}e^{|X|/t} \ge 1 + \mathbb{E}|X|/t$, i.e. $\mathbb{E}|X| \le t$. Letting $t \downarrow \|X\|_{\psi_1}$ yields $|\mathbb{E}X| \le \mathbb{E}|X| \le \|X\|_{\psi_1}$. Combining the three estimates,

$$\|X - \mathbb{E}X\|_{\psi_1} \ \le \ \|X\|_{\psi_1} + \frac{|\mathbb{E}X|}{\log 2} \ \le \ \Big(1 + \frac{1}{\log 2}\Big)\|X\|_{\psi_1},$$

and $1 + 1/\log 2$. □

We shall use the following concentration inequalities for sums and quadratic forms of independent sub-gaussian or sub-exponential random variables.

Theorem 2 (Bernstein's inequality)**.** *Let $X_1, \dots, X_N$ be independent, mean-zero, sub-exponential random variables, and let $K := \max_i \|X_i\|_{\psi_1}$. Then, for every $t \ge 0$,*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{N} X_i\Big| \ge t\right) \ \le \ 2\exp\left[-c\min\left(\frac{t^2}{K^2 N},\ \frac{t}{K}\right)\right].$$

Theorem 3 (Hanson–Wright inequality)**.** *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent components satisfying $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \le K$ for every $i \in [n]$, and let $A \in \mathbb{R}^{n \times n}$. Then, for every $t \ge 0$,*

$$\mathbb{P}\left(\left|X^\top A X - \mathbb{E}\left[X^\top A X\right]\right| > t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|A\|_F^2},\ \frac{t}{K^2\|A\|_{op}}\right\}\right),$$

where $\|A\|_F$ and $\|A\|_{op}$ denote the Frobenius and the operator norm of A.

We recall some tail bounds for the sup-norm and the Euclidean norm of a standard Gaussian vector. The first is a union bound over the standard Gaussian tail $\mathbb{P}(|g| > t) \le 2e^{-t^2/2}$, $t \ge 0$ (Vershynin, 2018, Prop. 2.1.2); the second is the sharp χ^2 tail bound of Laurent and Massart (Laurent & Massart, 2000, Lemma 1).

Lemma 4. *Let $X \sim \mathcal{N}(0, \mathrm{Id}_d)$. Then, for every $d \geq 2$,*

$$\mathbb{P}\left(\|X\|_\infty > \log d\right) \;\leq\; 2d\, e^{-\log^2 d/2},$$

and, for every $x > 0$,

$$\mathbb{P}\left(\|X\|_2^2 \geq d + 2\sqrt{dx} + 2x\right) \leq e^{-x}, \qquad \mathbb{P}\left(\|X\|_2^2 \leq d - 2\sqrt{dx}\right) \leq e^{-x}.$$

In particular, choosing $x = d/16$, each of the two inequalities $\frac{1}{2}d \leq \|X\|_2^2 \leq 2d$ fails with probability at most $e^{-d/16}$.

Proof. For the first claim, by a union bound and the Gaussian tail bound,

$$\mathbb{P}\left(\|X\|_\infty > \log d\right) \leq \sum_{i=1}^{d} \mathbb{P}\left(|X_i| > \log d\right) \leq 2d\, e^{-\log^2 d/2}.$$

The two-sided bound on $\|X\|_2^2 \sim \chi^2_d$ is in (Laurent & Massart, 2000, Lemma 1). For the last claim, with $x = d/16$ one has $d - 2\sqrt{dx} = d/2$ and $d + 2\sqrt{dx} + 2x = \frac{13}{8}d \leq 2d$. □

Next, we recall that the operator norm of an $m \times n$ random matrix with independent sub-gaussian entries is of order $\sqrt{m} + \sqrt{n}$ with high probability.

Theorem 4. *Let A be an $m \times n$ random matrix with independent, mean-zero, sub-gaussian entries A_{ij}, and let $K := \max_{i,j} \|A_{ij}\|_{\psi_2}$. Then, for any $t > 0$,*

$$\|A\|_{op} \leq CK\left(\sqrt{m} + \sqrt{n} + t\right)$$

with probability at least $1 - 2\exp(-t^2)$.

Finally, we recall the Gaussian concentration inequality for Lipschitz functions. These references state the inequality for Euclidean-Lipschitz functions of a standard Gaussian vector, but, since the Frobenius norm is the Euclidean norm on $\mathbb{R}^{m\times n} \cong \mathbb{R}^{mn}$ and the entries of G are i.i.d. $\mathcal{N}(0,1)$, the matrix version below is the same statement with the same constants.

Lemma 5 (Theorem 2.26 in Wainwright (2019))**.** *Let $F\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ be K-Lipschitz with respect to the Frobenius norm, i.e. $|F(A) - F(B)| \leq K\|A - B\|_F$ for all $A, B \in \mathbb{R}^{m\times n}$, and let $G \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, for every $t \geq 0$,*

$$\mathbb{P}\left\{|F(G) - \mathbb{E}F(G)| > t\right\} \leq 2\exp\left(-\frac{t^2}{2K^2}\right).$$

Finally, we recall a known result on randomly initialized MLPs, which says that, with high probability over the weights, the Euclidean norm of each hidden layer grows at most like the square root of the width.

Lemma 6. *Under (H1), (H3) and (H4), for all $x \in \mathcal{X}$ and $\ell \in [L-1]$, there exists a constant $C > 0$ not dependent on d such that*

$$\mathbb{P}_\vartheta\left(\|\alpha^{(\ell)}(x, \vartheta^{(\ell)})\|_2 \geq C\sqrt{d'}\right) \leq 2\ell e^{-d'}\,.$$

Proof. We proceed by induction. We start by the base case $\ell = 1$. We notice that, using Lemma 4 and Theorem 4, for $x \in \mathcal{X}$, it holds

$$\begin{aligned}
\left\|\alpha^{(1)}\left(x, \vartheta^{(1)}\right)\right\|_2^2 &= \sum_{i=1}^{d_1} \sigma\left(\frac{1}{\sqrt{d_0}}\vartheta^{(1)}_{i:}x\right)^2 \\
&\leq L_\sigma^2 \sum_{i=1}^{d_1}\left(\frac{1}{\sqrt{d_0}}\vartheta^{(1)}_{i:}x\right)^2 \\
&\leq L_\sigma^2 \left\|\frac{1}{\sqrt{d_0}}\vartheta^{(1)}\right\|_{op}^2 \|x\|_2^2 \leq L_\sigma^2 \left\|\vartheta^{(1)}\right\|_{op}^2 \|x\|_2^2\,.
\end{aligned}$$

Since the entries of $\vartheta^{(1)}$ are independent, centered and, by Lemma 1 applied with $\tau^2 = 1$, $\|\vartheta_{ij}^{(1)}\|_{\psi_2} = \sqrt{8/3}$ for all $i \in [d_1]$, $j \in [d_0]$. Combining Theorem 4 with $K = \sqrt{8/3}$ and $t = \sqrt{d_1}$ and Lemma 4, we obtain that there exists a constant $C > 0$ not dependent on d such that

$$\mathbb{P}_\vartheta \left(\left\| \alpha^{(1)} \left(x, \vartheta^{(1)} \right) \right\|_2 \geq C\sqrt{d} \right) \leq 2e^{-d} .$$

Inductively, assume that, for all $x \in \mathcal{X}$, there exists constants $C, c > 0$ not dependent on d such that $\left\| \alpha^{(\ell-1)} \left(x, \vartheta^{(\ell-1)} \right) \right\|_2 \leq C\sqrt{d}$ with probability at least $1 - 2(L-1)e^{-cd}$. Then, using Theorem 4 with $K = \sqrt{8/3}$ and $t = \sqrt{d_\ell}$ as in the base case, together with the independence of $\vartheta^{(\ell)}$ from $\alpha^{(\ell-1)}$, we conclude that

$$\begin{aligned}
&\mathbb{P}_\vartheta \left(\left\| \alpha^{(\ell)} \left(x, \vartheta^{(\ell)} \right) \right\|_2 \geq C\sqrt{d} \right) \\
&\quad \leq \mathbb{P}_\vartheta \left(L_\sigma \left\| \frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \right\|_{op} \|\alpha^{(\ell-1)}\|_2 \geq C\sqrt{d} \right) \\
&\quad \leq \mathbb{P}_\vartheta \left(\left\| \frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \right\|_{op} \geq C \right) + 2(L-1)e^{-d'} \\
&\quad \leq 2Le^{-d} ,
\end{aligned}$$

for some constant $C > 0$ not dependent on d. □

A.2 Proofs of Theorem 1

To improve readability, we sometimes omit the constants which do not depend on d and N_c form the formulas, since they don't affect the asymptotic behavior. We define, for each $\ell \in [L]$, the quantization residual $r^{(\ell)} : (0, \infty) \times \mathbb{R} \to \mathbb{R}$ as

$$r(s, t) := t - s \lfloor t/s \rceil_B . \tag{5}$$

Whenever r is applied to a multidimensional object, it acts componentwise. In the next lemma, we show some basic properties of the quantized residual evaluated at initialization.

Lemma 7. *Let $\vartheta \in \mathbb{R}^d$ with $\vartheta_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Then, for all $s > 0$, the residual satisfies the following properties*

1. *$r(s, \vartheta_i)$ is odd, i.e. $r(s, \vartheta_i) = -r(s, -\vartheta_i)$, for all $i \in [d]$.*
2. *$|r(s, \vartheta_i)| \leq |\vartheta_i|$ for all $i \in [d]$.*
3. *For any fixed $x \in \mathbb{R}$, the function $s \mapsto r^2(s, x)$ is $2M_B|x|$-Lipschitz continuous.*
4. *$r(s, \vartheta_i)$, $i \in [d]$, are independent, centered and sub-Gaussian; more precisely, for all $i \in [d]$, it holds*

$$\mathbb{E}_\vartheta [r(s, \vartheta_i)] = 0 , \qquad \|r(s, \vartheta_i)\|_{\psi_2} \leq \sqrt{\frac{8}{3}} .$$

Proof. We prove each property separately.

1. Property 1. follows immediately by oddness and symmetry of $\lfloor \cdot \rceil_B$. Indeed, for all $i \in [d]$, we have
$$r(s, -\vartheta_i) = -\vartheta_i - \Pi_{s, M_B}(-\vartheta_i) = -\vartheta_i + \Pi_{s, M_B}(\vartheta_i) = -r(s, \vartheta_i) .$$
2. For all $i \in [d]$, it holds
$$|r(s, \vartheta_i)| = |\vartheta_i - \Pi_{s, M_B}(\vartheta_i)| \overset{(*)}{=} \operatorname{dist}(\vartheta_i, s\{-M_B, \dots, M_B\}) \leq |\vartheta_i| , \tag{6}$$
where $(*)$ follows from the definition (1) of $\lfloor \cdot \rceil_B$, while the last inequality holds because 0 is a quantization level, i.e. $0 \in s\{-M_B, \dots, M_B\}$.

3. Let us assume without loss of generality that $x > 0$. The case with $x < 0$ follows by oddness of Π_{s,M_B}. Let us fix $n \in [M_B]$, and denote with $\sigma_n = \frac{x}{n-\frac{1}{2}}$. On $(\sigma_{n+1}, \sigma_n]$ (with $\sigma_{M_B+1} := 0$) we have $\lfloor x/s \rceil_B = n$, hence $r^2(s,x) = (x - ns)^2$ there; on $(\sigma_n, \sigma_{n-1}]$ (with $\sigma_0 := \infty$) we have $\lfloor x/s \rceil_B = n-1$, hence $r^2(s,x) = (x-(n-1)s)^2$ there. Both expressions are polynomials in s, so $r^2(\cdot, x)$ is continuous on each of the two half-open intervals, and in particular left-continuous at σ_n. For right-continuity at σ_n we use $x = (n - \frac{1}{2})\sigma_n$ to compute

$$\lim_{s \to \sigma_n^+} r^2(s,x) = \big(x - (n-1)\sigma_n\big)^2 = \Big(\frac{\sigma_n}{2}\Big)^2 = \big(x - n\sigma_n\big)^2 = r^2(\sigma_n, x).$$

Hence $r^2(\cdot, x)$ is continuous at every σ_n, and therefore on all of $(0, \infty)$. On each interval (σ_{n+1}, σ_n), $r^2(s,x)$ is differentiable, with derivative

$$\left| \frac{d}{ds} r^2(s,x) \right| = 2n\big|x - sn\big| \le 2M_B|x|\,,$$

where we used $n \le M_B$ and $|x - sn| = |r(s,x)| \le |x|$ by Property 2. A continuous function on an interval that is $\mathcal{C}^1$ off finitely many points with derivative bounded by L is L-Lipschitz (apply the mean value theorem between consecutive breakpoints and sum), which gives the claim.

4. Independence follows from the independence of the components of ϑ and from the point-wise nature of $r(s,\vartheta)$. Since the Gaussian distribution is symmetric and Π_{s,M_B} is odd, we have that

$$\mathbb{E}_\vartheta\left[r(s,\vartheta_i)\right] = \mathbb{E}_\vartheta\left[\vartheta_i\right] - \mathbb{E}_\vartheta\left[\Pi_{s,M_B}(\vartheta_i)\right] = 0\,.$$

From the point-wise bound in Property 2, we get $\mathbb{E}_\vartheta\big[e^{r^2(s,\vartheta_i)/u^2}\big] \le \mathbb{E}_\vartheta\big[e^{\vartheta_i^2/u^2}\big]$ for every $u > 0$, hence

$$\begin{aligned}\|r(s,\vartheta_i)\|_{\psi_2} &= \inf\Big\{u > 0 :\ \mathbb{E}_{\vartheta_i}\Big[e^{r^2(s,\vartheta_i)/u^2}\Big] \le 2\Big\} \\ &\le \inf\Big\{u > 0 :\ \mathbb{E}_{\vartheta_i}\Big[e^{\vartheta_i^2/u^2}\Big] \le 2\Big\} = \|\vartheta_i\|_{\psi_2} = \sqrt{\frac{8}{3}}\,,\end{aligned}$$

where the last equality is Lemma 1 applied with $\tau^2 = 1$. As i was arbitrary, this concludes the proof.

□

For clarity, we recall here Theorem 1.

Theorem 1. *Let $B \ge 2$, $X \in \mathbb{R}^{d \times N_c}$ be nonzero matrix satisfying (H2), and ϑ satisfying (H1). Then, for all $d \ge 2$, it holds*

$$\mathbb{P}\left(\sup_{s>0} |\mathcal{E}(s) - \mathcal{E}^\infty(s)| > \frac{2 + 12M_B}{\log d}\right) \le C\, M_B\, d\, e^{-c\log^2 d}\,,$$

where $c, C > 0$ are absolute constants independent on d.

Proof. Let us set $H := XX^\top \in \mathbb{R}^{d\times d}$ and abbreviate $r_{\text{eff}} := r_{\text{eff}}(H)$. We note that

$$\begin{aligned}\mathcal{E}(s) &= \frac{1}{\mathrm{Tr}(H)} \sum_{i=1}^d H_{ii}\big(\vartheta_i - \Pi_{s,M_B}(\vartheta_i)\big)^2 \\ &\quad + \frac{1}{\mathrm{Tr}(H)} \sum_{i=1}^d \sum_{j\ne i}^d H_{ij}\big(\vartheta_i - \Pi_{s,M_B}(\vartheta_i)\big)\big(\vartheta_j - \Pi_{s,M_B}(\vartheta_j)\big) \\ &= D_B(s,\vartheta,X) + O_B(s,\vartheta,X)\,,\end{aligned} \tag{7}$$

where

$$D_B(s,\vartheta,X) := \frac{1}{\mathrm{Tr}(H)}\sum_{i=1}^{d} H_{ii}\big(\vartheta_i - \Pi_{s,M_B}(\vartheta_i)\big)^2$$

$$O_B(s,\vartheta,X) := \frac{1}{\mathrm{Tr}(H)}\sum_{i=1}^{d}\sum_{j\neq i}^{d} H_{ij}\big(\vartheta_i - \Pi_{s,M_B}(\vartheta_i)\big)\big(\vartheta_j - \Pi_{s,M_B}(\vartheta_j)\big)\,.$$

Since the ϑ_i are independent and, by Lemma 7, the residuals $r(s,\vartheta_i)$ are centered, we have $\mathbb{E}_\vartheta[O_B(s,\vartheta,X)] = 0$. Hence, it holds

$$\mathcal{E}^\infty(s) = \mathbb{E}_\vartheta\big[D_B(s,\vartheta,X)\big]$$
$$\mathcal{E}(s) - \mathcal{E}^\infty(s) = D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)] + O_B(s,\vartheta,X).$$

Let us start by controlling $D_B(s,\vartheta,X)$. Let us fix $s>0$. We note that

$$\mathbb{E}_\vartheta[D_B(s,\vartheta,X)] = \frac{1}{\mathrm{Tr}(H)}\sum_{i=1}^{d} H_{ii}\,\mathbb{E}_\vartheta\big[r(s,\vartheta_i)^2\big] \;=\; \mathbb{E}_{\vartheta\sim\mathcal{N}(0,1)}\big[r(s,\vartheta)^2\big]\,,$$

since the ϑ_i are identically distributed. We note that H is positive semidefinite, so $0\le H_{ii}\le \|H\|_{op}$ for every $i\in[d]$, and therefore

$$\|\,\mathrm{diag}(H)\|_F^2 = \sum_{i=1}^{d} H_{ii}^2 \le \|H\|_{op}\,\mathrm{Tr}(H), \qquad \|\,\mathrm{diag}(H)\|_{op} \le \|H\|_{op}. \tag{8}$$

By Lemma 7 the coordinates of $r(s,\vartheta)$ are independent, centered and sub-Gaussian with $\|r(s,\vartheta_i)\|_{\psi_2} \le K := \sqrt{\frac{8}{3}}$. Set $\varepsilon_d := \sqrt{\log^6 d / r_{\text{eff}}}$. Therefore, by Theorem 3 applied to the matrix $\mathrm{diag}(H)$ with threshold $\mathrm{Tr}(H)\varepsilon_d$, and (8), for all $s>0$, we have that

$$\begin{aligned}
&\mathbb{P}_\vartheta\left(\big|D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]\big| \ge \varepsilon_d\right)\\
&\quad= \mathbb{P}_\vartheta\left(\Big|r(s,\vartheta)^\top\,\mathrm{diag}(H)\,r(s,\vartheta) - \mathbb{E}_\vartheta\left[r(s,\vartheta)^\top\,\mathrm{diag}(H)\,r(s,\vartheta)\right]\Big| \ge \mathrm{Tr}(H)\varepsilon_d\right)\\
&\quad\le 2\exp\left(-c\min\left\{\frac{\mathrm{Tr}(H)^2\varepsilon_d^2}{K^4\,\|H\|_{op}\,\mathrm{Tr}(H)},\ \frac{\mathrm{Tr}(H)\,\varepsilon_d}{K^2\,\|H\|_{op}}\right\}\right)\\
&\quad= 2\exp\left(-c\min\left\{\tfrac{9}{64}\,r_{\text{eff}}\,\varepsilon_d^2,\ \tfrac{3}{8}\,r_{\text{eff}}\,\varepsilon_d\right\}\right)\\
&\quad\le 2\exp\left(-c'\min\left\{\log^6 d,\ \sqrt{r_{\text{eff}}}\,\log^3 d\right\}\right)\,,
\end{aligned} \tag{9}$$

where $c' = 3c/64$. Moreover, by Lemma 7, the function $s\mapsto r(s,\vartheta)^2$ is Lipschitz continuous, that is

$$|r(s,\vartheta)^2 - r(t,\vartheta)^2| \le 2M_B|\vartheta||t-s|\,.$$

Therefore, using $H_{ii}\ge 0$, $\sum_i H_{ii} = \mathrm{Tr}(H)$ and $\mathbb{E}_{\vartheta\sim\mathcal{N}(0,1)}[|\vartheta|] = \sqrt{2/\pi}$, we obtain that

$$\begin{aligned}
&|D_B(s,\vartheta,X) - D_B(t,\vartheta,X)| \le |s-t|\,\frac{2M_B}{\mathrm{Tr}(H)}\sum_{i=1}^{d} H_{ii}|\vartheta_i| \le 2M_B\,\|\vartheta\|_\infty\,|s-t|\,,\\
&|\mathbb{E}_\vartheta[D_B(s,\vartheta,X)] - \mathbb{E}_\vartheta[D_B(t,\vartheta,X)]| \le |s-t|\,2M_B\,\mathbb{E}|\vartheta_1| \le 2M_B\,|s-t|\,.
\end{aligned} \tag{10}$$

Let us denote, for simplicity, $\beta := \log d$. Let us consider the partition of $(0,2\beta]$ constructed via the grid $s_k := 2k\frac{\beta}{d}$ for $k=0,\dots,d$, whose mesh is $\eta := s_{k+1}-s_k = \frac{2\beta}{d}$. Let us now fix $k\in\{0,\dots,d-1\}$ and $s\in[s_k,s_{k+1}]$. Then, using (10), it holds

$$\begin{aligned}
&|D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]|\\
&\quad\le |D_B(s,\vartheta,X) - D_B(s_{k+1},\vartheta,X)| + |D_B(s_{k+1},\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_{k+1},\vartheta,X)]|\\
&\qquad + |\mathbb{E}_\vartheta[D_B(s_{k+1},\vartheta,X)] - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]|\\
&\quad\le 2\eta\,M_B\|\vartheta\|_\infty + 2\eta\,M_B + |D_B(s_{k+1},\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_{k+1},\vartheta,X)]|\\
&\quad= 4M_B\,\frac{\beta}{d}\,\|\vartheta\|_\infty + 4M_B\,\frac{\beta}{d} + |D_B(s_{k+1},\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_{k+1},\vartheta,X)]|\,.
\end{aligned}$$

Since, by Lemma 4, $\|\vartheta\|_\infty \le \beta$ with probability at least $1 - 2d\, e^{-\log^2 d/2}$, and $\beta/d = \log d/d \le \beta^2$ for $d \ge 3$, on the event $\{\|\vartheta\|_\infty \le \beta\}$ it holds

$$\begin{aligned}&\sup_{s\in(0,2\beta]} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]| \\ &\quad\le 8M_B\,\frac{\log^2 d}{d} + \max_{k\in[d]} |D_B(s_k,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_k,\vartheta,X)]|\,.\end{aligned} \tag{11}$$

It remains to cover $s > 2\beta$. Since $\Pi_{s,M_B}(t) = 0$ whenever $s > 2|t|$, on $\{\|\vartheta\|_\infty < \beta\}$ we have $r(s,\vartheta_k) = \vartheta_k = r(2\beta,\vartheta_k)$ for every $k \in [d]$ and every $s > 2\beta$, hence $D_B(s,\vartheta,X) = D_B(s_d,\vartheta,X)$ for all $s > 2\beta$, where $s_d = 2\beta$. Moreover, by Lemma 7, $0 \le r(s,\vartheta_k)^2 \le \vartheta_k^2$ for all $s > 0$, and $r(s,\vartheta_k) = \vartheta_k = r(2\beta,\vartheta_k)$ when $|\vartheta_k| < \beta$ and $s > 2\beta$, so that $|r(s,\vartheta_k)^2 - r(s_d,\vartheta_k)^2| \le \vartheta_k^2 \mathbf{1}_{\{|\vartheta_k|\ge\beta\}}$. Therefore

$$\begin{aligned}\sup_{s>2\beta} \big|\mathbb{E}_\vartheta[D_B(s,\vartheta,X)] - \mathbb{E}_\vartheta[D_B(s_d,\vartheta,X)]\big| &\le \sum_{k=1}^d \mathbb{E}\big[\vartheta_k^2 \mathbf{1}_{\{|\vartheta_k|\ge\beta\}}\big] = d\,\mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[G^2\mathbf{1}_{\{|G|\ge\log d\}}\big] \\ &\le 4d\log d\, e^{-\log^2 d/2} \le 4\,\frac{\log^2 d}{d}\end{aligned} \tag{12}$$

for $d \ge 27$, where we used $\mathbb{E}[G^2\mathbf{1}_{\{|G|\ge a\}}] \le 4a\, e^{-a^2/2}$ for $a \ge 1$ with $a = \log d$, and in the last step that $e^{2u-u^2/2} \le u$ for $u = \log d \ge 4$. Hence, on $\{\|\vartheta\|_\infty < \beta\}$,

$$\begin{aligned}&\sup_{s>0} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]| \\ &\quad= \max\Big\{ \sup_{s\in(0,2\beta]} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]|\,,\ \sup_{s>2\beta} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]| \Big\} \\ &\quad\le \max\Big\{ \sup_{s\in(0,2\beta]} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]|\,,\ |D_B(s_d,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_d,\vartheta,X)]| \\ &\qquad + \sup_{s>2\beta} |\mathbb{E}_\vartheta[D_B(s_d,\vartheta,X)] - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]| \Big\} \\ &\quad\overset{(*)}{\le} 12M_B\,\frac{\log^2 d}{d} + \max_{k\in[d]} |D_B(s_k,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_k,\vartheta,X)]|\,,\end{aligned}$$

where in the second step we used that $D_B(s,\vartheta,X) = D_B(s_d,\vartheta,X)$ for $s > 2\beta$ on $\{\|\vartheta\|_\infty < \beta\}$, and in $(*)$ we used (11) for the first term, (12) for the last term, $M_B \ge 2$, and the fact that s_d is one of the grid points, so that $|D_B(s_d,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_d,\vartheta,X)]| \le \max_{k\in[d]} |D_B(s_k,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_k,\vartheta,X)]|$. Combining (9) and (11), we can conclude that

$$\begin{aligned}&\mathbb{P}_\vartheta\left(\sup_{s>0} |D_B(s,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s,\vartheta,X)]| > \varepsilon_d + 12M_B\,\frac{\log^2 d}{d}\right) \\ &\quad\le \mathbb{P}_\vartheta\big(\|\vartheta\|_\infty > \beta\big) + \sum_{k=1}^d \mathbb{P}_\vartheta\left(|D_B(s_k,\vartheta,X) - \mathbb{E}_\vartheta[D_B(s_k,\vartheta,X)]| > \varepsilon_d\right) \\ &\quad\le 2d\, e^{-\log^2 d/2} + 2d\exp\left(-c'\min\left\{\log^6 d,\ \sqrt{r_{\text{eff}}}\,\log^3 d\right\}\right) \\ &\quad\le 4d\exp\left(-c''\min\left\{\log^2 d,\ \sqrt{r_{\text{eff}}}\,\log^3 d\right\}\right)\,,\end{aligned} \tag{13}$$

which concludes the bound on $D_B(s,\vartheta,X)$.

It remains now to bound $O_B(s,\vartheta,X)$. Let $\xi \in \{-1,1\}^d$ have i.i.d. $\mathrm{Rad}(1/2)$ coordinates, independent of ϑ. Since the coordinates of ϑ are independent and symmetric, $\xi \odot \vartheta \overset{d}{=} \vartheta$ as random vectors; moreover $r(s,\xi_i\vartheta_i) = \xi_i\, r(s,\vartheta_i)$ by item 1 of Lemma 7. Writing $\widetilde{H} := H - \mathrm{diag}(H)$, it holds, for every $s > 0$,

$$\sum_{i=1}^{d}\sum_{j\neq i}^{d} H_{ij}\, r(s,\xi_i\vartheta_i)\, r(s,\xi_j\vartheta_j) = \sum_{i,j=1}^{d} \xi_i\, r(s,\vartheta_i)\, \widetilde{H}_{ij}\, r(s,\vartheta_j)\, \xi_j$$
$$= \xi^\top \operatorname{diag}(r(s,\vartheta))\, \widetilde{H}\, \operatorname{diag}(r(s,\vartheta))\, \xi\,,$$

and consequently, as processes indexed by $s>0$,

$$\big(O_B(s,\vartheta,X)\big)_{s>0} \overset{d}{=} \Big(\frac{d}{\operatorname{Tr}(H)}\, \xi^\top \operatorname{diag}(r(s,\vartheta))\, \widetilde{H}\, \operatorname{diag}(r(s,\vartheta))\, \xi\Big)_{s>0},$$

so that it suffices to bound the supremum of the right-hand side. Since $\widetilde{H}$ is equal to H on the off-diagonal terms and equal to zero on the diagonal terms, and H is positive semidefinite with diagonal entries in $[0, \|H\|_{op}]$, it follows that $\|\widetilde{H}\|_F \le \|H\|_F$ and, by Weyl's inequality, $\|\widetilde{H}\|_{op} \le \|H\|_{op}$. For fixed $\vartheta \in \mathbb{R}^d$, there are at most $M_B d$ positive breakpoints,

$$\sigma_{i,k} = \frac{|\vartheta_i|}{k+1/2}, \qquad 1\le i\le d, \quad 0\le k\le M_B-1\,,$$

which we sort increasingly as $0<\tau_1<\cdots<\tau_K$, $K\le M_B d$, and we set $\tau_0:=0$. Let us define

$$p(s) := \sum_{i,j=1}^{d} \xi_i\, r(s,\vartheta_i)\, \widetilde{H}_{ij}\, r(s,\vartheta_j)\, \xi_j \qquad s>0.$$

For each $s\in(\tau_k,\tau_{k+1})$ the residual $r(s,\vartheta_i)$ is affine, and therefore

$$p(s) = p_k(s) := \sum_{i,j=1}^{d} \widetilde{H}_{ij}\, \xi_i\xi_j\, (\vartheta_i - s\, n_i)(\vartheta_j - s\, n_j)\,,$$

where $n_i := \lfloor \vartheta_i/s \rceil_B$, which is constant for $s\in(\tau_k,\tau_{k+1})$. Thus p_k is a polynomial of degree at most 2, defined on all of $\mathbb{R}$, and it coincides with p on the open interval (τ_k,τ_{k+1}). Therefore, it holds

$$\sup_{s>0}|p(s)| = \max\Big\{\max_{k=0,\dots,K-1}\ \sup_{s\in(\tau_k,\tau_{k+1})}|p_k(s)|,\ \max_{k=1,\dots,K}|p(\tau_k)|,\ \sup_{s>\tau_K}|p(s)|\Big\}.$$

For $s>\tau_K = 2\|\vartheta\|_\infty$ every rounding vanishes, so $r(s,\vartheta)=\vartheta$ and $p(s)=\sum_{i,j}\widetilde{H}_{ij}\xi_i\xi_j\vartheta_i\vartheta_j$ is constant. By Lemma 10, applied to p_k on the closed interval $[\tau_k,\tau_{k+1}]$, we have

$$\sup_{s\in(\tau_k,\tau_{k+1})}|p_k(s)| \le 3\max\big\{|p_k(\tau_k)|,\ |p_k(\tau_{k+1})|,\ |p_k((\tau_k+\tau_{k+1})/2)|\big\}.$$

Therefore, we have that

$$\sup_{s>0}|p(s)|$$
$$\le 3\max\Big\{\max_{k<K}\max\big\{|p_k(\tau_k)|, |p_k(\tau_{k+1})|, |p_k(\tfrac{\tau_k+\tau_{k+1}}{2})|\big\},\ \max_{k\le K}|p(\tau_k)|,\ \Big|\sum_{i,j=1}^{d}\widetilde{H}_{ij}\xi_i\xi_j\vartheta_i\vartheta_j\Big|\Big\}.$$

Each quantity inside the maximum is of the form $\xi^\top C\xi$ with $C=\operatorname{diag}(b)\,\widetilde{H}\,\operatorname{diag}(b)$, where $b\in\mathbb{R}^d$ collects either the residuals $r(s_0,\vartheta_i)$ at a node s_0, or their one-sided limits at a breakpoint, or ϑ_i itself. Since 0 is a quantization level, $|r(s,\vartheta_i)|\le|\vartheta_i|$, and at a breakpoint the one-sided limits equal $\pm\sigma_{i,k}/2$ with $\sigma_{i,k}\le 2|\vartheta_i|$; hence $|b_i|\le|\vartheta_i|$ in all cases. Thus there are at most $4M_Bd+1$ matrices $C_j\in\mathbb{R}^{d\times d}$ of the form $C_j=\operatorname{diag}(b^{(j)})\,\widetilde{H}\,\operatorname{diag}(b^{(j)})$ with $|b_i^{(j)}|\le|\vartheta_i|$, such that

$$\sup_{s>0}\Big|\xi^\top\ \operatorname{diag}(r(s,\vartheta))\,\widetilde{H}\,\operatorname{diag}(r(s,\vartheta))\,\xi\Big| \le 3\max_{j\in[4M_Bd+1]}|\xi^\top C_j\xi|\,. \tag{14}$$

This includes rounding ties. Since $\widetilde{H}$ has zero diagonal, so does every C_j, and hence

$$\mathbb{E}_\xi[\xi^\top C_j\xi] = \operatorname{Tr} C_j = 0.$$

On the event $\{\|\vartheta\|_\infty \le \beta\}$, which has probability at least $1 - 2d\, e^{-\log^2 d/2}$, using $|b_i^{(j)}| \le \|\vartheta\|_\infty \le \beta$, the submultiplicativity of the operator norm and $\|H\|_F^2 \le \|H\|_{op} \operatorname{Tr}(H)$, it holds that

$$\|C_j\|_{op} \le \|\vartheta\|_\infty^2 \, \|\widetilde{H}\|_{op} \le \beta^2 \, \|H\|_{op} = \log^2 d \|H\|_{op} \,,$$
$$\|C_j\|_F^2 \le \|\vartheta\|_\infty^4 \|\widetilde{H}\|_F^2 \le \beta^4 \|H\|_{op} \operatorname{Tr}(H) = \log^4 d \|H\|_{op} \operatorname{Tr}(H) \,.$$

We note that the ξ_i are independent, centered, with $\|\xi_i\|_{\psi_2}$ an absolute constant. Hence, on $\{\|\vartheta\|_\infty \le \beta\}$, by Theorem 3 applied conditionally on ϑ, (14) and a union bound, it holds

$$\begin{aligned}
&\mathbb{P}_\xi\Big(\sup_{s>0} \frac{1}{\operatorname{Tr}(H)} \big|\xi^\top \operatorname{diag}(r(s,\vartheta))\, \widetilde{H}\, \operatorname{diag}(r(s,\vartheta))\, \xi\big| > \varepsilon_d \Big) \\
&\quad \le \sum_{j=1}^{4M_B d+1} \mathbb{P}_\xi\Big(|\xi^\top C_j \xi| > \frac{\operatorname{Tr}(H)}{3} \varepsilon_d \Big) \\
&\quad \le 2(4M_B d+1) \exp\Big(- c \min\Big\{ \frac{\operatorname{Tr}(H)^2 \varepsilon_d^2/9}{\log^4 d \, \|H\|_{op} \operatorname{Tr}(H)}, \frac{\operatorname{Tr}(H)\varepsilon_d/3}{\log^2 d \, \|H\|_{op}} \Big\} \Big) \qquad (15) \\
&\quad = 2(4M_B d+1) \exp\Big(- c \min\Big\{ \frac{r_{\text{eff}}\, \varepsilon_d^2}{9 \log^4 d}, \frac{r_{\text{eff}}\, \varepsilon_d}{3 \log^2 d} \Big\} \Big) \\
&\quad = 2(4M_B d+1) \exp\Big(- c \min\Big\{ \frac{\log^2 d}{9}, \frac{\log d \sqrt{r_{\text{eff}}}}{3} \Big\} \Big) \,.
\end{aligned}$$

Integrating over ϑ and using the equality in law established above,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{s>0} |O_B(s,\vartheta,X)| > \varepsilon_d \Big) \\
&\quad \le \mathbb{P}_\vartheta\big(\|\vartheta\|_\infty > \beta \big) + 2(4M_B d+1) \exp\Big(- \frac{c}{9} \min\big\{ \log^2 d,\ \log d \sqrt{r_{\text{eff}}} \big\} \Big).
\end{aligned}$$

Combining (7), (13) and (15), we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s>0} \, |\mathcal{E}(s) - \mathcal{E}^\infty(s)| > 2\sqrt{\frac{\log^6 d}{r_{\text{eff}}(H)}} + 12 M_B \, \frac{\log^2 d}{d} \right) \\
&\qquad \le C\, M_B\, d\, \exp\left\{ -c \min\left\{ \log^2 d,\ \log d \sqrt{r_{\text{eff}}(H)} \right\} \right\}
\end{aligned}$$

for absolute constants $c, C > 0$. In particular, if $r_{\text{eff}}(H) \ge \log^8 d$, then $\min\{\log^2 d, \log d\sqrt{r_{\text{eff}}}\} = \log^2 d$ and $\sqrt{\log^6 d / r_{\text{eff}}} \le 1/\log d$, so that

$$\mathbb{P}\left(\sup_{s>0} \, |\mathcal{E}(s) - \mathcal{E}^\infty(s)| > \frac{2 + 12 M_B}{\log d} \right) \le C\, M_B\, d\, e^{-c \log^2 d} \,,$$

where $c, C > 0$ are universal constants independent of d. □

A.3 Proof of Corollary 1

Corollary 1 follows directly from the uniform approximation in Theorem 1, which controls the difference between the infima of $\mathcal{E}$ and $\mathcal{E}^\infty$ and ensures that any approximate minimizer of $\mathcal{E}$ also approximately minimizes $\mathcal{E}^\infty$.

Corollary 1. *Under the assumptions of Theorem 1, set $\varepsilon_d := \frac{2+12M_B}{\log d}$. Then, with probability at least $1 - C\, M_B\, d\, e^{-c \log^2 d}$, the following statements hold simultaneously:*

1. *The optimal values satisfy*
$$\big| m - m^\infty \big| \ \le \ \varepsilon_d \,.$$
2. *For every $\eta \ge 0$ and every $\widehat{s} > 0$ satisfying $\mathcal{E}(\widehat{s}) \le m + \eta$ one has*
$$0 \ \le \ \big(\mathcal{E}^\infty(\widehat{s}) - m^\infty \big) \ \le \ 2\varepsilon_d + \eta.$$
In particular, any exact minimizer $\widehat{s}$ of $\mathcal{E}$ satisfies this bound with $\eta = 0$.

Proof. By Theorem 1, with probability at least $1 - CM_B d\, e^{-c\log^2 d}$ it holds

$$\sup_{s>0} \left|\mathcal{E}(s) - \mathcal{E}^\infty(s)\right| \;\le\; \varepsilon_d\,.$$

We work on this event. For every $s > 0$, we have

$$\mathcal{E}^\infty(s) - \varepsilon_d \;\le\; \mathcal{E}(s) \;\le\; \mathcal{E}^\infty(s) + \varepsilon_d\,.$$

Taking infima over $s > 0$ yields

$$m^\infty - \varepsilon_d \;\le\; m \;\le\; m^\infty + \varepsilon_d,$$

which proves the first point of the corollary.

Now let $\eta \ge 0$ and let $\widehat{s} > 0$ satisfy $\mathcal{E}(\widehat{s}) \le m + \eta$. The uniform error bound and the first assertion give

$$\begin{aligned}\mathcal{E}^\infty(\widehat{s}) &\le \mathcal{E}(\widehat{s}) + \varepsilon_d \\ &\le m + \eta + \varepsilon_d \\ &\le m^\infty + \eta + 2\varepsilon_d.\end{aligned}$$

The observation that $\mathcal{E}^\infty(\widehat{s}) \ge m^\infty$ concludes the proof. □

A.4 Verifying (H2) for Multi-Layer Perceptrons at initialization

The uniform approximation of Theorem 1 relies on (H2), namely a lower bound on the effective rank of $XX^\top$. In this section we show that this condition holds, with high probability, in the setting of a random multi-layer perceptron with isotropic data for channelwise quantization, i.e., for layer ℓ, we quantize independently the rows $\vartheta^{(\ell)}_{i:}$ and X is given by the output of the previous layer (not quantized). We begin by defining the considered model.

Definition 4. *Let $L > 1$ be fixed. A* multi-layer perceptron *(MLP) of depth L is defined as the parametric function $f_\vartheta : \mathbb{R}^{d_0} \to \mathbb{R}$ obtained via the recursion*

$$\begin{aligned}&\alpha^{(0)}(x) := x, \quad \alpha^{(\ell)}(x) := \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}}\vartheta^{(\ell)}\alpha^{(\ell-1)}(x)\right) \quad \text{for } \ell = 1,\dots,L-1, \\ &f_\vartheta(x) := \frac{1}{\sqrt{d_{L-1}}}\vartheta^{(L)\top}\alpha^{(L-1)}(x),\end{aligned} \tag{16}$$

Here, $\vartheta^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ for $\ell = 1,\dots,L-1$, and $\vartheta^{(L)} \in \mathbb{R}^{d_{L-1}}$. The nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ acts componentwise, is L_σ-Lipschitz, and satisfies $\sigma(0) = 0$.

Rather than assuming (H2) directly, we give explicit conditions on the network architecture and calibration distribution under which it holds with high probability.

Assumptions We assume that:

- (H3) $d_k = \alpha_k d$ for some constant $\alpha_\ell > 0$ and $\ell \in [L]$, and we denote with $d := \min_{k=1,\dots,L-1} d_k$.
- (H4) the calibration dataset $\mathbb{D}_{cal} = \{X'_i\}_{i=1}^{N_c}$ is the realization of N_c iid copies of $X' \sim \mathcal{N}(0, \mathrm{Id}_{d_0})$. We denote with with $X' = [X'_1 \dots, X'_{N_c}] \in \mathbb{R}^{d_0 \times N_c}$. We assume that $N_c \ge C\log^8 d$ for some constant $C > 0$ not dependent on d.

Assumption (H3) places the hidden-layer widths in a proportional regime. Assumption (H4) models the calibration inputs as isotropic Gaussian and requires the sample size to grow at least polylogarithmically with the minimum hidden width. This permits calibration sets that are asymptotically smaller than the layer widths. We additionally assume that the activation is odd and Lipschitz, covering choices such as $\tanh$ but excluding ReLU.

The results of Theorem 1 relies on the feature Gram matrix $\alpha^{(\ell)}(X')\,\alpha^{(\ell)}(X')^\top$ having large effective rank. The following theorem shows that this condition is fulfilled, with high probability, for the

(iterative) PTQ problem of a MLP with Gaussian weights and Gaussian calibration inputs, as soon as the width is large enough.

In the rest of this section, to improve readability, we sometimes omit the constants which do not depend on d and N_c form the formulas, since they don't affect the asymptotic behavior. We start by controlling, with high probability, the operator norm $\left\|\alpha^{(\ell)}(X')\,\alpha^{(\ell)}(X')^\top\right\|_{\mathrm{op}}$ from above and the trace $\mathrm{tr}\left(\alpha^{(\ell)}(X')\,\alpha^{(\ell)}(X')^\top\right)$ from below, which are the two quantities entering the effective rank.

Lemma 8. *Let us assume (H1), (H3) and (H4) and let $\ell \in [L-1]$. Then, there exists universal constants $C, c > 0$ such that*

$$\mathbb{P}_{\vartheta,X'}\left(\left\|\alpha^{(\ell)}(X')\,\alpha^{(\ell)}(X')^\top\right\|_{\mathrm{op}} > C(d+N_c)\right) \;\le\; 2\cdot 9^{-c(d+N_c)} + 2\ell e^{-d}\,.$$

Proof. Let us fix $\ell \in [d]$, $u \in \mathbb{S}^{d_\ell - 1}$ and $v \in \mathbb{S}^{N_c-1}$. Let us fix $\vartheta^{(k)} \in \mathbb{R}^{d_k \times d_{k-1}}$ for $k \in [\ell]$ and define $F_{u,v} : \mathbb{R}^{d_0 \times N_c} \to \mathbb{R}$ as follows

$$F_{u,v}(X) := u^\top \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}}\vartheta^{(\ell)}\alpha^{(\ell-1)}(X)\right) v\,,$$

where $\alpha^{(\ell-1)}(X) \in \mathbb{R}^{d_{\ell-1}\times N_c}$ is the output of the $(\ell-1)$-th layer evaluated on the calibration dataset (with the convention $\alpha^{(0)}(X) = X$). We note that $F_{u,v}$ is Lipschitz. Indeed, since $\|u\|_2 = \|v\|_2 = 1$ and σ is L_σ-Lipschitz and applied entrywise, for all $X, Y \in \mathbb{R}^{d_0\times N_c}$ it holds

$$\begin{aligned}
|F_{u,v}(X) - F_{u,v}(Y)| &\le \left\|\sigma\left(\frac{1}{\sqrt{d_{\ell-1}}}\vartheta^{(\ell)}\alpha^{(\ell-1)}(X)\right) - \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}}\vartheta^{(\ell)}\alpha^{(\ell-1)}(Y)\right)\right\|_F \\
&\le L_\sigma \left\|\frac{1}{\sqrt{d_{\ell-1}}}\vartheta^{(\ell)}\right\|_{op} \|\alpha^{(\ell-1)}(X) - \alpha^{(\ell-1)}(Y)\|_F \\
&\le L_\sigma^\ell \prod_{k=1}^{\ell} \left\|\frac{1}{\sqrt{d_{k-1}}}\;\vartheta^{(k)}\right\|_{op} \|X - Y\|_F =: \Lambda(\vartheta)\,\|X-Y\|_F\,,
\end{aligned}$$

where we used $\|MN\|_F \le \|M\|_{op}\|N\|_F$ and, in the last step, we iterated the previous inequality over the layers $\ell-1, \dots, 1$. Moreover, since the activation σ is odd, it follows that $F_{u,v}(-X) = -F_{u,v}(X)$. Now, since by (H4), $X'_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ which is symmetric, it holds that

$$\mathbb{E}_{X'}[F_{u,v}(X')] = 0\,.$$

Therefore, using Lemma 5, for all $t > 0$, it holds

$$\mathbb{P}_{X'}\big(|F_{u,v}(X')| > t\big) \;\le\; 2\exp\left\{-\frac{t^2}{2\Lambda(\vartheta)^2}\right\}. \tag{17}$$

We now control $\Lambda(\vartheta)$. Since the entries of $\vartheta^{(k)}$ are i.i.d. $\mathcal{N}(0,1)$, we have $K = \max_{i,j}\|\vartheta^{(k)}_{ij}\|_{\psi_2} = \sqrt{8/3}$, and Theorem 4 with $t = \sqrt{d}$ gives, for every $k \in [\ell]$ (recall $d_k = \alpha d$),

$$\mathbb{P}_\vartheta\left(\|\vartheta^{(k)}\|_{op} > 3CK\sqrt{d}\right) \;\le\; \mathbb{P}_\vartheta\left(\|\vartheta^{(k)}\|_{op} > CK\big(\sqrt{d_k} + \sqrt{d_{k-1}} + \sqrt{d}\big)\right) \;\le\; 2e^{-d}\,.$$

Hence, by a union bound, the event

$$\mathcal{E} := \left\{\max_{k\in[\ell]} \|\vartheta^{(k)}\|_{op} \le 3CK\sqrt{d}\right\}$$

satisfies $\mathbb{P}_\vartheta(\mathcal{E}^c) \le 2\ell e^{-d}$, and on $\mathcal{E}$, under hypothesis (H3), it holds $\Lambda(\vartheta) \le \Lambda := (3CKL_\sigma)^\ell \prod_{k=1}^{\ell}\sqrt{\frac{1}{\alpha_{k-1}}}$. In particular, by (17), for every $\vartheta \in \mathcal{E}$ and every $t > 0$,

$$\mathbb{P}_{X'}\big(|F_{u,v}(X')| > t\big) \;\le\; 2\exp\left\{-\frac{t^2}{2\Lambda^2}\right\}. \tag{18}$$

Let $\mathcal{U} \subset \mathbb{S}^{d_\ell - 1}$, $\mathcal{V} \subset \mathbb{S}^{N_c - 1}$ be the $\frac{1}{4}$-nets of Lemma 11, so that $|\mathcal{U}| \le 9^{d_\ell}$ and $|\mathcal{V}| \le 9^{N_c}$. By Lemma 11,

$$\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} \le 2 \max_{(u,v) \in \mathcal{U} \times \mathcal{V}} |F_{u,v}(X')| \,,$$

so, for every $t > 0$, we have

$$\left\{ \left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right\} \subseteq \bigcup_{(u,v) \in \mathcal{U} \times \mathcal{V}} \{ |F_{u,v}(X')| > t \},$$

and by (18) and a union bound, for every $\vartheta \in \mathcal{E}$,

$$\begin{aligned} \mathbb{P}_X \left(\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right) &\le |\mathcal{U}|\,|\mathcal{V}| \cdot 2 \exp\left\{ -\frac{t^2}{2\Lambda^2} \right\} \\ &\le 2 \cdot 9^{\,d_\ell + N_c} \exp\left\{ -\frac{t^2}{2\Lambda^2} \right\}. \end{aligned}$$

Let us choose $t^2 := 4 \log 9 \; (d_\ell + N_c)\, \Lambda^2$. Then $\dfrac{t^2}{2\Lambda^2} = 2(d_\ell + N_c) \log 9$, so the right-hand side equals $2 \cdot 9^{d_\ell + N_c} \cdot 9^{-2(d_\ell + N_c)} = 2 \cdot 9^{-(d_\ell + N_c)}$. Finally, since X' and ϑ are independent,

$$\begin{aligned} &\mathbb{P}_{\vartheta, X'} \left(\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right) \\ &\quad \le \mathbb{E}_\vartheta \left[\mathbf{1}_{\mathcal{E}} \, \mathbb{P}_{X'} \left(\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right) \right] + \mathbb{P}_\vartheta(\mathcal{E}^c) \\ &\quad \le 2 \cdot 9^{-(d_\ell + N_c)} + 2\ell e^{-d} \,. \end{aligned}$$

Finally, since X and ϑ are independent,

$$\begin{aligned} &\mathbb{P}_{\vartheta, X'} \left(\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right) \\ &\quad \le \mathbb{E}_\vartheta \left[\mathbf{1}_{\mathcal{E}} \, \mathbb{P}_{X'} \left(\left\| \sigma\left(\frac{1}{\sqrt{d_{\ell-1}}} \vartheta^{(\ell)} \alpha^{(\ell-1)}(X') \right) \right\|_{\mathrm{op}} > 2t \right) \right] + \mathbb{P}_\vartheta(\mathcal{E}^c) \\ &\quad \le 2 \cdot 9^{-(d_\ell + N_c)} + 2\ell e^{-d} \,. \end{aligned}$$

We note that $\|\alpha^{(\ell)}(X')\, \alpha^{(\ell)}(X')^\top\|_{op} = \|\alpha^{(\ell)}(X)\|_{op}^2$. Therefore, under (H3), there exists universal constants $C, c > 0$ such that

$$\begin{aligned} \mathbb{P}_{\vartheta, X'} \left(\left\| \alpha^{(\ell)}(X')\, \alpha^{(\ell)}(X')^\top \right\|_{\mathrm{op}} > C\,(d + N_c) \right) &\le \mathbb{P}_{\vartheta, X'} \left(\left\| \alpha^{(\ell)}(X') \right\|_{\mathrm{op}} > 2t \right) \\ &\le 2 \cdot 9^{-(d_\ell + N_c)} + 2\ell e^{-d} \\ &\le 2 \cdot 9^{-c(d + N_c)} + 2\ell e^{-d} \,, \end{aligned}$$

which concludes the proof. □

The next lemma shows that, with high probability, a lower bound on the trace $\mathrm{tr}\left(\alpha^{(\ell)}(X')\, \alpha^{(\ell)}(X')^\top\right)$ for all $\ell \in [L-1]$.

Lemma 9. *Let us assume (H1), (H3) and (H4). Then there exist universal constants $a, C, c > 0$ not depending on d, such that for every $\ell \in \{0, \dots, L-1\}$, it holds*

$$\mathbb{P}\left(\mathrm{tr}\left(\alpha^{(\ell)}(X')\, \alpha^{(\ell)}(X')^\top \right) < a\, d N_c \right) \le C N_c\, e^{-cd} \,.$$

Proof. Let $\ell \in \{0, \dots, L-1\}$. We note that

$$\operatorname{tr}\left(\alpha^{(\ell)}(X')\alpha^{(\ell)}(X')^\top\right) = \|\alpha^{(\ell)}(X')\|_F^2 = \sum_{j=1}^{N_c} \|\alpha_{:j}^{(\ell)}(X')\|_2^2 = \sum_{j=1}^{N_c} \|\alpha^{(\ell)}(X'_{:j})\|_2^2 \,,$$

so that a lower bound on the trace follows from a lower bound on $\|\alpha^{(\ell)}(X'_{:j})\|_2^2$ holding simultaneously for all $j \in [N_c]$. For $\ell \geq 1$, $i \in [d_\ell]$ and $j \in [N_c]$, let us define the random variables

$$Z_i^j := \frac{1}{\sqrt{d_{\ell-1}}}\, \vartheta_{i:}^{(\ell)} \alpha^{(\ell-1)}(X'_{:j}) \,,$$
$$Y_i^j := \sigma\big(Z_i^j\big)^2 \,.$$

We note that $\|\alpha^{(\ell)}(X'_{:j})\|_2^2 = \sum_{i=1}^{d_\ell} Y_i^j$. Let $\mathcal{F}_{\ell-1}$ denote the σ-algebra generated by $X', \vartheta^{(1)}, \dots, \vartheta^{(\ell-1)}$. Since $\vartheta^{(\ell)}$ is independent of $\mathcal{F}_{\ell-1}$, the random variables $Y_1^j, \dots, Y_{d_\ell}^j$ are i.i.d., and, using (H3), we have

$$0 \leq Y_i^j \leq L_\sigma^2 \big(Z_i^j\big)^2 \,.$$

Moreover, conditioned on $\mathcal{F}_{\ell-1}$, we have that $Z_i^j \sim \mathcal{N}\left(0, \frac{\|\alpha^{(\ell-1)}(X'_{:j})\|_2^2}{d_{\ell-1}}\right)$. Let us define following iterative constants

$$a_0 := \frac{1}{2}, \quad b_0 := 2$$
$$m_\ell = \inf_{h \in [a_{\ell-1}, b_{\ell-1}]} \mathbb{E}_{Z\sim\mathcal{N}(0,h)}\big[\sigma^2(Z)\big] \,,$$
$$b_\ell := 2L_\sigma^2\, b_{\ell-1} \,,$$
$$a_\ell := \tfrac{1}{2}\, m_\ell \,.$$

and the event

$$E_\ell := \bigcap_{j=1}^{N_c} \Big\{ a_\ell d_\ell \leq \|\alpha^{(\ell)}(X'_{:j})\|_2^2 \leq b_\ell d_\ell \Big\} \,.$$

We estimate now the probability of E_ℓ. We proceed by induction over ℓ. The base case $\ell = 0$ follows from Lemma 4. Indeed, since $\alpha^{(0)}(X'_{:j}) = X'_{:j} \sim \mathcal{N}(0, \mathrm{Id}_{d_0})$ by (H4), for each $j \in [N_c]$ each of the two inequalities

$$\frac{1}{2} d_0 \leq \|\alpha^{(0)}(X'_{:j})\|_2^2 = \|X'_{:j}\|_2^2 \leq 2d_0$$

fails with probability at most $e^{-d_0/16}$. Therefore, by a union bound,

$$\mathbb{P}\left(E_0^c\right) \leq \sum_{j=1}^{N_c} \left[\mathbb{P}\left(\|X'_{:j}\|_2^2 < \frac{1}{2} d_0 \right) + \mathbb{P}\left(\|X'_{:j}\|_2^2 > 2d_0 \right) \right] \leq 2N_c e^{-d_0/16} \,.$$

Let now $\ell \geq 1$. As inductive hypothesis, we assume that there exists $c_1, \dots, c_{\ell-1} > 0$, depending only on the activation σ and on the layer index, such that the event

$$E_{\ell-1} = \bigcap_{j=1}^{N_c} \Big\{ a_{\ell-1} d_{\ell-1} \leq \|\alpha^{(\ell-1)}(X'_{:j})\|_2^2 \leq b_{\ell-1} d_{\ell-1} \Big\}$$

satisfies

$$\mathbb{P}\left(E_{\ell-1}^C\right) \;\leq\; 2N_c\, e^{-d_0/16} + N_c \sum_{k=1}^{\ell-1} \left(2e^{-c_k d_k} + e^{-d_k/16} \right) \,.$$

Then, we have that

$$\begin{aligned}\mathbb{P}\left(E_\ell^C\right) &\le \mathbb{P}\left(E_\ell^C \;\middle|\; E_{\ell-1}\right) + \mathbb{P}\left(E_{\ell-1}^C\right) \\ &\le \sum_{j=1}^{N_c}\left[\mathbb{P}\left(\|\alpha^{(\ell)}(X'_{:j})\|_2^2 < a_\ell d_\ell \;\middle|\; E_{\ell-1}\right) + \mathbb{P}\left(\|\alpha^{(\ell)}(X'_{:j})\|_2^2 > b_\ell d_\ell \;\middle|\; E_{\ell-1}\right)\right] \\ &\quad + \mathbb{P}\left(E_{\ell-1}^C\right).\end{aligned} \tag{19}$$

Since $E_{\ell-1} \in \mathcal{F}_{\ell-1}$, for every event A we have $\mathbb{P}(A \,|\, E_{\ell-1}) = \mathbb{E}\big[\mathbf{1}_{E_{\ell-1}}\mathbb{P}(A \,|\, \mathcal{F}_{\ell-1})\big]/\mathbb{P}(E_{\ell-1})$, so it suffices to bound $\mathbb{P}(A \,|\, \mathcal{F}_{\ell-1})$ uniformly on the event $E_{\ell-1}$.

By Lemma 4, we notice that on $E_{\ell-1}$, it holds

$$\begin{aligned}\mathbb{P}\left(\|\alpha^{(\ell)}(X'_{:j})\|_2^2 > b_\ell d_\ell \;\middle|\; \mathcal{F}_{\ell-1}\right) &\le \mathbb{P}\left(L_\sigma^2 \sum_{i=1}^{d_\ell}\left(Z_i^j\right)^2 > 2L_\sigma^2 b_{\ell-1} d_\ell \;\middle|\; \mathcal{F}_{\ell-1}\right) \\ &\le \mathbb{P}\left(\sum_{i=1}^{d_\ell}\sqrt{d_{\ell-1}}\Big(\frac{Z_i^j}{\|\alpha^{(\ell-1)}(X'_{:j})\|_2}\Big)^2 > 2d_\ell \;\middle|\; \mathcal{F}_{\ell-1}\right) \\ &\le e^{-d_\ell/16},\end{aligned} \tag{20}$$

where we used $Y_i^j \le L_\sigma^2 (Z_i^j)^2$ and $b_\ell = 2L_\sigma^2 b_{\ell-1}$ in the first step, $s_j^2 \le b_{\ell-1}$ in the second step, and the fact that, conditioned on $\mathcal{F}_{\ell-1}$, $d_{\ell-1}\frac{Z_1^j}{\|\alpha^{(\ell-1)}(X'_{:j})\|_2}, \dots, d_{\ell-1}\frac{Z_{d_\ell}^j}{\|\alpha^{(\ell-1)}(X'_{:j})\|_2}$ are iid $\mathcal{N}(0,1)$ in the last step.

Let us define

$$m_\ell := \inf_{h\in[a_{\ell-1},b_{\ell-1}]} \mathbb{E}_{Z\sim\mathcal{N}(0,h)}\left[\sigma^2(Z)\right]. \tag{21}$$

We note that $m_\ell > 0$. Indeed, the map $h \mapsto \mathbb{E}_{Z\sim\mathcal{N}(0,h)}[\sigma^2(Z)] = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\sigma^2(\sqrt{h}\,g)]$ is continuous on $(0,\infty)$, and it is strictly positive for every $h > 0$: otherwise $\sigma(\sqrt{h}\,g) = 0$ almost surely, i.e. $\sigma = 0$ Lebesgue-almost everywhere, which is excluded by assumption. Hence the infimum over the compact interval $[a_{\ell-1}, b_{\ell-1}] \subset (0,\infty)$ is attained and positive.

Note also that $m_\ell \le L_\sigma^2 b_{\ell-1}$, since $\mathbb{E}_{Z\sim\mathcal{N}(0,h)}[\sigma^2(Z)] \le L_\sigma^2 h$. We notice that on $E_{\ell-1}$,

$$\mathbb{E}\big[Y_i^j \;\big|\; \mathcal{F}_{\ell-1}\big] = \mathbb{E}_{Z\sim\mathcal{N}(0,s_j^2)}\left[\sigma^2(Z)\right] \ge m_\ell,$$

by definition of m_l. Conditioned on $\mathcal{F}_{\ell-1}$, we note that $Y_1^j, \dots, Y_{d_\ell}^j$ are sub-exponential i.i.d. random variables with

$$\begin{aligned}\left\|Y_i^j - \mathbb{E}[Y_i^j \mid \mathcal{F}_{\ell-1}]\right\|_{\psi_1} &\le C\,\|Y_i^j\|_{\psi_1} \\ &\le CL_\sigma^2\left\|(Z_i^j)^2\right\|_{\psi_1} \\ &= CL_\sigma^2\|Z_i^j\|_{\psi_2}^2 \\ &= \tfrac{8}{3}\,CL_\sigma^2\frac{\|\alpha^{(\ell-1)}(X'_{:j})\|_2^2}{d_{\ell-1}} \\ &\le \tfrac{8}{3}\,CL_\sigma^2 b_{\ell-1},\end{aligned}$$

where $C \ge 1$ is an absolute constant, the ψ_1- and ψ_2-norms are taken with respect to the conditional distribution given $\mathcal{F}_{\ell-1}$, and we used Lemma 2, Lemma 3, and Lemma 1 with $\tau = \frac{\|\alpha^{(\ell-1)}(X'_{:j})\|_2^2}{d_{\ell-1}}$. By applying Theorem 2 conditioned on $\mathcal{F}_{\ell-1}$ to the centered variables $Y_i^j - \mathbb{E}[Y_i^j \,|\, \mathcal{F}_{\ell-1}]$, with $t := \frac{m_\ell}{2}d_\ell$, it follows that on $E_{\ell-1}$,

$$\begin{aligned}\mathbb{P}\left(\|\alpha^{(\ell)}(X'_{:j})\|_2^2 < a_\ell d_\ell \;\middle|\; \mathcal{F}_{\ell-1}\right) &\le \mathbb{P}\left(\left|\sum_{i=1}^{d_\ell} Y_i^j - d_\ell\,\mathbb{E}[Y_1^j \mid \mathcal{F}_{\ell-1}]\right| \ge \frac{m_\ell}{2}d_\ell \;\middle|\; \mathcal{F}_{\ell-1}\right) \\ &\le 2\exp\left\{-c\min\left(\frac{t^2}{d_\ell K_\ell^2}, \frac{t}{K_\ell}\right)\right\},\end{aligned}$$

where $c > 0$ is an absolute constant, and in the first step we used that, on $E_{\ell-1}$, $a_\ell d_\ell = \frac{m_\ell}{2} d_\ell \leq d_\ell \, \mathbb{E}[Y_1^j \mid \mathcal{F}_{\ell-1}] - \frac{m_\ell}{2} d_\ell$, since $\mathbb{E}[Y_1^j \mid \mathcal{F}_{\ell-1}] \geq m_\ell$. We now compute the minimum in the exponent. Plugging in $t = \frac{m_\ell}{2} d_\ell$, we have

$$\min\left(\frac{t^2}{d_\ell K_\ell^2}, \frac{t}{K_\ell}\right) = d_\ell \min\left\{\left(\frac{m_\ell}{2K_\ell}\right)^2, \frac{m_\ell}{2K_\ell}\right\} = d_\ell \left(\frac{m_\ell}{2K_\ell}\right)^2,$$

where the last equality holds because $m_\ell \leq L_\sigma^2 b_{\ell-1} \leq K_\ell$ gives $\frac{m_\ell}{2K_\ell} \leq \frac{1}{2} < 1$, so that the square is the smaller of the two terms. Therefore, on $E_{\ell-1}$, we have

$$\mathbb{P}\left(\|\alpha^{(\ell)}(X'_{:j})\|_2^2 < a_\ell d_\ell \,\middle|\, \mathcal{F}_{\ell-1}\right) \leq 2\exp\{-c_\ell \, d_\ell\}\,, \tag{22}$$

where $c_\ell = c\left(\frac{m_\ell}{2K_\ell}\right)^2 > 0$ depends only on σ and ℓ, and not on d_ℓ or N_c.

Using (19), (20), (22) and taking the union over $j \in [N_c]$, we obtain

$$\mathbb{P}\left(E_\ell^C\right) \;\leq\; 2N_c \, e^{-d_0/16} + N_c \sum_{k=1}^{\ell} \left(2e^{-c_k d_k} + e^{-d_k/16}\right).$$

Under (H3), there exist universal constants $C, c > 0$ which depends only on σ and κ such that

$$\mathbb{P}\left(E_\ell^c\right) \;\leq\; C N_c \, e^{-cd}\,.$$

Finally, let $a := \alpha_\ell \min_{0\leq \ell \leq L-1} a_\ell > 0$. Since on the event E_ℓ we have $\|\alpha^{(\ell)}(X'_{:j})\|_2^2 \geq a_\ell d_\ell \geq a \, d_\ell$ for every $j \in [N_c]$, the first display of the proof yields the inclusion

$$E_\ell \;\subseteq\; \left\{\mathrm{tr}\left(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right) \;\geq\; a \, d N_c\right\},$$

and therefore

$$\mathbb{P}\left(\mathrm{tr}\left(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right) \;<\; a \, d_\ell N_c\right) \;\leq\; \mathbb{P}\left(E_\ell^c\right) \;\leq\; C N_c \, e^{-cd}\,.$$

This concludes the proof. □

Theorem 5. *Let us assume (H1), (H3) and (H4), and let $\ell \in [L-1]$. Then there exist constants $C, c, \bar{C} > 0$ not dependent on d, and an absolute constant $D > 0$, such that for all $d > D$, with probability at least $1 - \bar{C} N_c \, e^{-c\,d}$ over ϑ and X,*

$$r_{\mathit{eff}}\big(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\big) \;\geq\; C \log^8 d\,.$$

Proof. By definition of the effective rank,

$$r_{\mathrm{eff}}\big(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\big) = \frac{\mathrm{tr}\left(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right)}{\left\|\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right\|_{\mathrm{op}}}\,,$$

By Lemma 8, there exist constants $C, c > 0$ such that

$$\mathbb{P}_{\vartheta, X'}\left(\left\|\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right\|_{\mathrm{op}} > C(d + N_c)\right) \;\leq\; 2 \cdot 9^{-c(d+N_c)} + 2\ell e^{-d}\,.$$

On the other hand, by Lemma 9, there exist constants $a, c', C' > 0$ such that

$$\mathbb{P}_{\vartheta, X'}\left(\mathrm{tr}\left(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right) \;<\; a \, d N_c\right) \;\leq\; C' N_c \, e^{-c' d}\,.$$

Therefore, by a union bound, with probability at least

$$1 - 2 \cdot 9^{-c(d+N_c)} - 2\ell e^{-d} - C' N_c e^{-c' d} \;\geq\; 1 - \bar{C} N_c \, e^{-\bar{c}\, d}$$

with $\bar{c} := \min\{c \log 9,\, 1,\, c'\}, \bar{C} := 2 + 2L + C'$, it holds that

$$r_{\mathrm{eff}}\big(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\big) = \frac{\mathrm{tr}\left(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right)}{\left\|\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\right\|_{\mathrm{op}}} \;\geq\; \tilde{C} \, \frac{d N_c}{d + N_c} \;\geq\; \frac{\tilde{C}}{2} \, \min\{d, N_c\}\,,$$

with $\tilde{C} := a\kappa / C$, where in the last step we have used the fact that $\frac{dN_c}{d+N_c} \geq \frac{1}{2}\min\{d, N_c\}$. Therefore, since by (H4) $N_c \geq \log^8 d$, and $\log^8 d \leq d$ for d large enough, we conclude that, with probability at least $1 - \bar{C} N_c e^{-\bar{c} d}$,

$$r_{\mathrm{eff}}\big(\alpha^{(\ell)}(X') \, \alpha^{(\ell)}(X')^\top\big) \;\geq\; \frac{\tilde{C}}{2} \, \log^8 d\,.$$

□

A.5 Proof of Proposition 1

We start by proving that $\mathcal{E}^\infty$ separates the two sources of quantization error. The first is the rounding error, which does not depend on the bit precision and would be present even with an infinite grid; the second is the clipping error, due to inputs falling outside the representable range $[-M_B s, M_B s]$, and it is the only term through which M_B enters. We record this decomposition in the following corollary, which will be the starting point of our analysis of the loss landscape.

Corollary 2. *For every $B \geq 2$ and $s > 0$,*

$$\mathcal{E}^\infty(M_B, s) = \mathcal{J}_\infty(s) + 4s \sum_{j=0}^{\infty} g\big(\big(M_B + j + \tfrac{1}{2}\big)\, s\big)\,,$$

with

$$\mathcal{J}_\infty(s) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[(Z - s\lfloor Z/s \rceil)^2] = 1 - \sum_{j=0}^{\infty} 4s\, g\big((j+\tfrac{1}{2})s\big)\,,$$

where $g(x) := \varphi(x) - x(1-\Phi(x))$, and both series converge absolutely.

Proof. For a natural number $n \geq 0$, let $\Pi_{s,n}(z) := s\max\{-n, \min\{n, \lfloor z/s \rceil\}\}$ denote the quantizer with $2n+1$ levels, so that $\mathcal{E}^\infty(n,s) = \mathbb{E}[(Z - \Pi_{s,n}(Z))^2]$. Let us fix $n \geq 0$, and notice that the quantizers $\Pi_{s,n}$ and $\Pi_{s,n+1}$ agree on $\{|z| \leq (n+\frac{1}{2})s\}$. On $\{z > (n+\frac{1}{2})s\}$ one has $\Pi_{s,n}(z) = ns$, while $\lfloor z/s \rceil \geq n+1$ there, so $\Pi_{s,n+1}(z) = (n+1)s$; the case $z < -(n+\frac{1}{2})s$ is symmetric. It holds

$$\begin{aligned}
\mathcal{E}^\infty(n,s) - \mathcal{E}^\infty(n+1,s) &= 2\int_{(n+1/2)s}^{\infty} \Big[(z-ns)^2 - \big(z-(n+1)s\big)^2\Big]\varphi(z)\,dz \\
&\overset{(*)}{=} 4s\int_{(n+\frac{1}{2})s}^{\infty} (z - (n+\tfrac{1}{2})s)\varphi(z)\,dz \\
&= 4s\big[\varphi((n+\tfrac{1}{2})s) - (n+\tfrac{1}{2})s\ \big(1 - \Phi\big((n+\tfrac{1}{2})s\big)\big)\big] \\
&= 4s\, g\big((n+\tfrac{1}{2})s\big)\,,
\end{aligned}$$

where in (*) we have used the fact that

$$(z-ns)^2 - \big(z-(n+1)s\big)^2 = 2zs - 2ns^2 - s^2 = 2s\Big(z - \big(n+\tfrac{1}{2}\big)\, s\Big),$$

and $\int_x^\infty z\varphi(z)dz = \varphi(x)$ and $\int_x^\infty \varphi(z)\ dz = 1 - \Phi(x)$. Let us fix $N > M_B$. Summing on $n = M_B, \ldots, N-1$ gives

$$\mathcal{E}^\infty(M_B, s) - \mathcal{E}^\infty(N, s) = \sum_{n=M_B}^{N-1} 4s\, g\big((n+\tfrac{1}{2})s\big)\,.$$

Since $0 \leq g(x) \leq \varphi(x)$ for $x \geq 0$, the series converges absolutely. Moreover $\big(Z - \Pi_{s,N}(Z)\big)^2 \leq Z^2 \in L^1$ and $\Pi_{s,N}(Z) \to s\lfloor Z/s \rceil$ point-wise as $N \uparrow \infty$, so dominated convergence gives $\mathcal{E}^\infty(N,s) \to \mathcal{J}_\infty(s)$. Letting $N \uparrow \infty$ and substituting $j = n - M_B$ concludes the proof.
Finally, since $\Pi_{s,0} \equiv 0$ and hence $\mathcal{E}^\infty(0,s) = \mathbb{E}[Z^2] = 1$, we can use the above formula for $\mathcal{E}^\infty(M_B,s) - \mathcal{E}^\infty(N,s)$ with M_B replaced by 0 and $N \to \infty$ to prove that

$$\mathcal{J}_\infty(s) = 1 - \sum_{n=0}^{\infty} 4s\, g\big((n+\tfrac{1}{2})s\big)\,.$$

□

For clarity, we restate Proposition 1 below.

Proposition 1. *The following properties hold:*

(i) For every $B \geq 2$, the function $\bar{\mathcal{E}}^\infty(m,s)$ is smooth, e.g. $\bar{\mathcal{E}}^\infty(m,s) \in \mathcal{C}^\infty([1,\infty)\times(0,\infty))$.

(ii) For every real $m \geq 1$, the function $s \mapsto \bar{\mathcal{E}}^{\infty}(m,s)$ has a unique critical point on $(0,\infty)$. This point is its unique global minimizer $s^*(m)$ and is nondegenerate:

$$\partial_s^2 \bar{\mathcal{E}}^{\infty}\big(m, s^*(m)\big) > 0.$$

(iii) The map $m \mapsto s^*(m)$ is continuously differentiable and strictly decreasing on $[1,\infty)$. Moreover, for every $m \geq 1$,

$$\frac{ds^*}{dm}(m) = -\frac{\bar{\mathcal{E}}^{\infty}_{ms}\big(m, s^*(m)\big)}{\bar{\mathcal{E}}^{\infty}_{ss}\big(m, s^*(m)\big)} < 0, \qquad s^*(m) > \frac{\sqrt{2}}{m+\frac{1}{2}}.$$

Proof. We prove each of the four points separately.

(i) The function g is $\mathcal{C}^\infty$, with $g' = -(1-\Phi)$, $g'' = \varphi$ and $g^{(k)} = p_k\varphi$ for a polynomial p_k when $k \geq 2$; in particular $|g^{(k)}(x)| \leq C_k(1+x)^k\varphi(x)$ for $x \geq 1$ and all $k \geq 0$, using $g(x) \leq \varphi(x)/x^2$ and $1-\Phi(x) \leq \varphi(x)/x$ by Lemma 12. Hence every partial derivative in (m,s) of the summand $s\,g\big((m+j+\frac{1}{2})s\big)$ is bounded by a summable function, on $\{s \geq s_0,\, m \geq 1\}$. All series of partial derivatives thus converge uniformly on compact subsets of $[1,\infty)\times(0,\infty)$, and term-wise differentiation gives $\bar{\mathcal{E}}^{\infty} \in \mathcal{C}^\infty\big([1,\infty)\times(0,\infty)\big)$.

(ii) We start by noticing that

$$\begin{aligned} g(as) = \varphi(as) - as\big(1-\Phi(as)\big) &= \int_{as}^{\infty} t\varphi(t)dt - as\int_{as}^{\infty}\varphi(t)dt \\ &= \int_{as}^{\infty}(t-as)\varphi(t)dt \\ &= s^2\int_a^{\infty}(u-a)\varphi(su)du \\ &= s^2\int_0^{\infty}(u-a)_+\varphi(su)du\,. \end{aligned}$$

Therefore, by definition of $\bar{\mathcal{E}}^{\infty}$ (cf. Corollary 2), we can write

$$\begin{aligned} \bar{\mathcal{E}}^{\infty}(m,s) &= 1 - 4s\sum_{j=0}^{\infty}\big[g\big((j+\tfrac{1}{2})s\big) - g\big((j+m+\tfrac{1}{2})s\big)\big] \\ &= 1 - 4s\sum_{j=0}^{\infty} s^2\int_0^{\infty}\Big[\big(u-(j+\tfrac{1}{2})\big)_+ - \big(u-(j+m+\tfrac{1}{2})\big)_+\Big]\varphi(su)du \\ &= 1 - 4s^3\int_0^{\infty}\sum_{j=0}^{\infty}\Big[\big(u-(j+\tfrac{1}{2})\big)_+ - \big(u-m-(j+\tfrac{1}{2})\big)_+\Big]\varphi(su)du \\ &= 1 - 4s^3\int_0^{\infty} D_m(u)\varphi(su)du\,, \end{aligned}$$

where $D_m(u) := \sum_{j=0}^{\infty} \left[\left(u - (j+\frac{1}{2})\right)_+ - \left(u - m - (j+\frac{1}{2})\right)_+ \right]$. Differentiating in s, it holds

$$
\begin{aligned}
\partial_s \bar{\mathcal{E}}^{\infty}(m,s) &= -12s^2 \int_0^{\infty} D_m(u)\varphi(su)du + 4s^4 \int_0^{\infty} u^2 D_m(u)\varphi(su)du \\
&= 4s^2 \int_0^{\infty} (s^2u^2 - 3) D_m(u)\varphi(su)du \\
&= -4s^2 \int_0^{\infty} (1 - s^2u^2) D_m(u)\varphi(su)du - 8s^2 \int_0^{\infty} D_m(u)\varphi(su)du \\
&\overset{(*)}{=} -4s^2 \int_0^{\infty} D_m(u) \frac{d}{du}\left(u\varphi(su)\right) du - 8s^2 \int_0^{\infty} D_m(u)\varphi(su)du \\
&= -4s^2 \left[D_m(u)u\varphi(su)\Big|_0^{\infty} - \int_0^{\infty} uD'_m(u)\varphi(su)\, du \right] - 8s^2 \int_0^{\infty} D_m(u)\varphi(su)du \\
&= 4s^2 \int_0^{\infty} [uD'_m(u) - 2D_m(u)]\varphi(su)du
\end{aligned}
$$

where in (*) we used the fact that

$$
\frac{d}{du}\left(u\varphi(su)\right) = (1 - s^2u^2)\varphi(us)\,.
$$

Integrating by parts we have

$$
\begin{aligned}
&\int [uD'_m(u) - 2D_m(u)]du \\
&\quad = uD_m(u) - 3\int D_m(u)du \\
&\quad = uD_m(u) - 3\sum_{j=0}^{\infty} \int \left[\left(u - (j+\tfrac{1}{2})\right)_+ - \left(u - m - (j+\tfrac{1}{2})\right)_+ \right] du \\
&\quad = uD_m(u) - \frac{3}{2}\sum_{j=0}^{\infty} \left[\left(u - (j+\tfrac{1}{2})\right)_+^2 - \left(u - m - (j+\tfrac{1}{2})\right)_+^2 \right].
\end{aligned}
\tag{23}
$$

Using again integration by parts, we get that

$$
\begin{aligned}
&\partial_s \bar{\mathcal{E}}^{\infty}(m,s) \\
&\quad = -4s^2 \int_0^{\infty} \left(uD_m(u) - \frac{3}{2}\sum_{j=0}^{\infty} \left[\left(u - (j+\tfrac{1}{2})\right)_+^2 - \left(u - m - (j+\tfrac{1}{2})\right)_+^2 \right] \right) \frac{d}{du}\varphi(su)\, du \\
&\quad = 4s^4 \int_0^{\infty} uP_m(u)\varphi(su)\, du
\end{aligned}
$$

where $P_m(u) := uD_m(u) - \frac{3}{2}\sum_{j=0}^{\infty}\left[\left(u-(j+\frac{1}{2})\right)_+^2 - \left(u-m-(j+\frac{1}{2})\right)_+^2\right]$. Now we notice that, for $0 < u < m + \frac{1}{2}$, it holds

$$\begin{aligned}
P_m(u) &= \sum_{j=0}^{\infty} u\left(u-(j+\tfrac{1}{2})\right)_+ - u\left(u-m-(j+\tfrac{1}{2})\right)_+ - \frac{3}{2}\left(u-(j+\tfrac{1}{2})\right)_+^2 \\
&\quad + \frac{3}{2}\left(u-m-(j+\tfrac{1}{2})\right)_+^2 \\
&= \sum_{j=0}^{\infty}\left(u - \frac{3}{2}\left(u-(j+\tfrac{1}{2})\right)_+\right)\left(u-j-\tfrac{1}{2}\right)_+ \\
&= \sum_{j=0}^{\lfloor u-1/2\rfloor_+}\left(u - \frac{3}{2}\left(u-j-\tfrac{1}{2}\right)\right)\left(u-j-\tfrac{1}{2}\right) \\
&= \sum_{j=0}^{\lfloor u-1/2\rfloor_+}\left(-\frac{u}{2} + \frac{3}{2}j + \frac{3}{4}\right)\left(u-j-\tfrac{1}{2}\right),
\end{aligned}$$

where $\sum_{j=0}^{\lfloor u-1/2\rfloor_+}$ denotes the sum over the indices $j \geq 0$ with $u - j - \frac{1}{2} > 0$ (empty if $u < \frac{1}{2}$). Writing $n := \lfloor u - 1/2\rfloor_+ + 1$ for the number of such indices, the last sum equals $\frac{n}{2}\left(\frac{1}{4} - (u-n)^2\right)$, which is nonnegative since $|u-n| \leq \frac{1}{2}$; hence $P_m \geq 0$ on $(0, m+\frac{1}{2})$, and it vanishes exactly at the half-integers.

On the other hand, for all $m > \frac{1}{2}$, $u > m + 1$, calling $n = \lfloor u - 1/2\rfloor_+ + 1$ and $r = \lfloor u - m - 1/2\rfloor_+ + 1$ we have that

$$\begin{aligned}
P_m(u) &= \sum_{j=0}^{\infty} u\left(u-(j+\frac{1}{2})\right)_+ - u\left(u-m-(j+\frac{1}{2})\right)_+ - \frac{3}{2}\left(u-(j+\frac{1}{2})\right)_+^2 \\
&\quad + \frac{3}{2}\left(u-m-(j+\frac{1}{2})\right)_+^2 \\
&= \sum_{j=0}^{\lfloor u-1/2\rfloor_+}\left(-\frac{u}{2} + \frac{3}{2}j + \frac{3}{4}\right)\left(u-j-\frac{1}{2}\right) \\
&\quad + \sum_{j=0}^{\lfloor u-m-1/2\rfloor_+}\left(\frac{u}{2} - \frac{3}{2}\left(j+\frac{1}{2}\right) - \frac{3}{2}m\right)\left(u-m-j-\frac{1}{2}\right)_+ \\
&= \sum_{j=0}^{n-1}\left(-\frac{u^2}{2} + 2u\left(j+\frac{1}{2}\right) - \frac{3}{2}\left(j+\frac{1}{2}\right)^2\right) \\
&\qquad + \sum_{j=0}^{r-1}\left(\frac{(u-m)^2}{2} - 2(u-m)\left(j+\frac{1}{2}\right) + \frac{3}{2}\left(j+\frac{1}{2}\right)^2 - m\left(u-m-\left(j+\frac{1}{2}\right)\right)\right) \\
&= \frac{n}{2}\left(\frac{1}{4} - (u-n)^2\right) + \frac{r(u-m)^2}{2} - r^2(u-m) + \frac{r(4r^2-1)}{8} - m\left(r(u-m) - \frac{r^2}{2}\right) \\
&= \frac{n}{2}\left(\frac{1}{4} - (u-n)^2\right) - \frac{r}{2}\left(\frac{1}{4} - (u-m-r)^2\right) - m\left(r(u-m) - \frac{r^2}{2}\right) \\
&\overset{(*)}{\leq} \frac{n}{8} - m\left(u-m-\frac{1}{2}\right) \\
&\leq \left(\frac{1}{8} - m\right)(u-m-1) + \frac{3}{16}(1-2m) \\
&< 0\,,
\end{aligned}$$

where in (*) we have noticed that

$$-\frac{1}{2} \le u-m-r = u-m-(\lfloor u-m-1/2\rfloor_+ +1) \le \frac{1}{2}$$
$$\implies \frac{r}{2}\left(\frac{1}{4}-(u-m-r)^2\right) \ge 0$$
$$\left(r(u-m)-\frac{r^2}{2}\right) = \sum_{j=0}^{r-1}\left(u-m-j-\frac{1}{2}\right) \ge u-m-\frac{1}{2}\,,$$

and in the following step we used $n \le u+\frac{1}{2}$. We consider now what happens when $m+\frac{1}{2} < u < \min\{m+1, \lfloor m\rfloor + \frac{3}{2}\}$. On this interval, it holds

$$\begin{aligned} D_m(u) &= \sum_{j=0}^{\infty}\left[\left(u-(j+\tfrac{1}{2})\right)_+ - \left(u-m-(j+\tfrac{1}{2})\right)_+\right] \\ &= \sum_{j=0}^{\lfloor m\rfloor}(u-j-\frac{1}{2}) - (u-m-\frac{1}{2}) \\ &= (\lfloor m\rfloor+1)u - \frac{(\lfloor m\rfloor+1)^2}{2} - u + m + \frac{1}{2} \\ &= \lfloor m\rfloor u - \frac{\lfloor m\rfloor(\lfloor m\rfloor+2)}{2} + m \end{aligned}$$

and, in particular $D'_m(u) = \lfloor m\rfloor$. Note also that

$$\frac{d}{du}\sum_{j=0}^{\infty}\left[\left(u-(j+\tfrac{1}{2})\right)_+^2 - \left(u-m-(j+\tfrac{1}{2})\right)_+^2\right] = 2D_m(u)$$

Let us compute

$$\begin{aligned} \frac{d}{du}P_m(u) &= D_m(u) + uD'_m(u) - 3D_m(u) \\ &= uD'_m(u) - 2D_m(u) \\ &= u\lfloor m\rfloor - 2\left(\lfloor m\rfloor u - \frac{\lfloor m\rfloor(\lfloor m\rfloor+2)}{2} + m\right) \\ &= \lfloor m\rfloor(\lfloor m\rfloor+2-u) - 2m \\ &\le \lfloor m\rfloor(\lfloor m\rfloor+\frac{3}{2}-m) - 2m \\ &= \lfloor m\rfloor\left(\frac{3}{2}-(m-\lfloor m\rfloor)\right) - 2\lfloor m\rfloor - 2(m-\lfloor m\rfloor) \\ &= -\frac{\lfloor m\rfloor}{2} - (\lfloor m\rfloor+2)(m-\lfloor m\rfloor) < 0 \end{aligned}$$

for all $m > 0$. We notice that $P_m(m+\frac{1}{2}) = \frac{\lfloor m\rfloor+1}{2}(m-\lfloor m\rfloor)(\lfloor m\rfloor+1-m)$ is positive for non-integer m and 0 for integer m.

Now, it follows that

– If $\min\{m+1, \lfloor m\rfloor+\frac{3}{2}\} = m+1$, then

$$P_m(m+1) \le \frac{3}{16}(1-2m) < 0\,.$$

– If $\min\{m+1, \lfloor m \rfloor + \frac{3}{2}\} = \lfloor m \rfloor + \frac{3}{2}$, then

$$
\begin{aligned}
P_m\big(\lfloor m \rfloor + \tfrac{3}{2}\big) &= \big(\lfloor m \rfloor + \tfrac{3}{2}\big) \sum_{j=0}^{\infty} \Big[\big(\lfloor m \rfloor + 1 - j\big)_+ - \big(\lfloor m \rfloor + 1 - m - j\big)_+ \Big] \\
&\quad - \frac{3}{2} \sum_{j=0}^{\infty} \Big[\big(\lfloor m \rfloor + 1 - j\big)_+^2 - \big(\lfloor m \rfloor + 1 - m - j\big)_+^2 \Big] \\
&= \big(\lfloor m \rfloor + \tfrac{3}{2}\big) \Big[\sum_{j=0}^{\lfloor m \rfloor} \big(\lfloor m \rfloor + 1 - j\big) - \big(\lfloor m \rfloor + 1 - m\big) \Big] \\
&\quad - \frac{3}{2} \Big[\sum_{j=0}^{\lfloor m \rfloor} \big(\lfloor m \rfloor + 1 - j\big)^2 - \big(\lfloor m \rfloor + 1 - m\big)^2 \Big] \\
&= \big(\lfloor m \rfloor + \tfrac{3}{2}\big) \Big[\frac{(\lfloor m \rfloor + 1)(\lfloor m \rfloor + 2)}{2} - \big(\lfloor m \rfloor + 1 - m\big) \Big] \\
&\quad - \frac{3}{2} \Big[\frac{(\lfloor m \rfloor + 1)(\lfloor m \rfloor + 2)(2\lfloor m \rfloor + 3)}{6} - \big(\lfloor m \rfloor + 1 - m\big)^2 \Big] \\
&= -\big(\lfloor m \rfloor + \tfrac{3}{2}\big)\big(\lfloor m \rfloor + 1 - m\big) + \frac{3}{2}\big(\lfloor m \rfloor + 1 - m\big)^2 \\
&= \big(\lfloor m \rfloor + 1 - m\big) \Big[-\lfloor m \rfloor - \tfrac{3}{2} + \frac{3}{2}\big(\lfloor m \rfloor + 1 - m\big) \Big] \\
&= -\frac{1}{2}\big(\lfloor m \rfloor + 1 - m\big)\big(3m - \lfloor m \rfloor\big) \;<\; 0\,,
\end{aligned}
$$

where in the second equality we used that $0 < \lfloor m \rfloor + 1 - m \leq 1$, so that $(\lfloor m \rfloor + 1 - m - j)_+$ vanishes for $j \geq 1$, and the last inequality holds since $\lfloor m \rfloor + 1 - m > 0$ and $3m - \lfloor m \rfloor \geq 2m > 0$.

It remains to check what happens for $\lfloor m \rfloor + \frac{3}{2} \le u \le m+1$. On the interval, it holds

$$
\begin{aligned}
P_m(u) &= \sum_{j=0}^{\infty} \Big[u\big(u-j-\tfrac{1}{2}\big)_+ - \frac{3}{2}\big(u-j-\tfrac{1}{2}\big)_+^2 \Big] \\
&\quad - \sum_{j=0}^{\infty} \Big[u\big(u-m-j-\tfrac{1}{2}\big)_+ - \frac{3}{2}\big(u-m-j-\tfrac{1}{2}\big)_+^2 \Big] \\
&= \sum_{j=0}^{\lfloor m \rfloor+1} \Big[u\big(u-j-\tfrac{1}{2}\big) - \frac{3}{2}\big(u-j-\tfrac{1}{2}\big)^2 \Big] - u\big(u-m-\tfrac{1}{2}\big) + \frac{3}{2}\big(u-m-\tfrac{1}{2}\big)^2 \\
&= \sum_{j=0}^{\lfloor m \rfloor+1} \Big[-\frac{u^2}{2} + 2u\big(j+\tfrac{1}{2}\big) - \frac{3}{2}\big(j+\tfrac{1}{2}\big)^2 \Big] \\
&\quad + \big(u-m-\tfrac{1}{2}\big)\Big[\frac{1}{2}\big(u-m-\tfrac{1}{2}\big) - \frac{1}{2} - m\Big] \\
&= -\frac{\lfloor m \rfloor + 2}{2}\, u^2 + (\lfloor m \rfloor + 2)^2\, u - \frac{(\lfloor m \rfloor + 2)\big[4(\lfloor m \rfloor + 2)^2 - 1\big]}{8} \\
&\quad - \frac{1}{2}\big(u-m-\tfrac{1}{2}\big)\big(m+\tfrac{3}{2}-u\big) - m\big(u-m-\tfrac{1}{2}\big) \\
&= \frac{\lfloor m \rfloor + 2}{2}\big(u-\lfloor m \rfloor - \tfrac{3}{2}\big)\big(\lfloor m \rfloor + \tfrac{5}{2} - u\big) \\
&\quad - \frac{1}{2}\big(u-m-\tfrac{1}{2}\big)\big(m+\tfrac{3}{2}-u\big) - m\big(u-m-\tfrac{1}{2}\big) \\
&\overset{(*)}{\le} \frac{u-m-\frac{1}{2}}{2}\Big[(\lfloor m \rfloor + 2)\big(m-\lfloor m \rfloor\big) - \tfrac{1}{2} - 2m\Big] \\
&= -\frac{u-m-\frac{1}{2}}{2}\Big[\lfloor m \rfloor\big(2-m+\lfloor m \rfloor\big) + \tfrac{1}{2}\Big] \;<\; 0\,,
\end{aligned}
$$

where in (*) we used

$$
\begin{aligned}
0 \le u - \lfloor m \rfloor - \tfrac{3}{2} &= (u-m-\frac{1}{2}) - (\lfloor m \rfloor + 1 - m) \\
&\le (u-m-\frac{1}{2}) - (\lfloor m \rfloor + 1 - m)(u-m-\frac{1}{2}) \\
&= \left[1 - (\lfloor m \rfloor + 1 - m)\right](u-m-\frac{1}{2}) \\
&= (m - \lfloor m \rfloor)(u-m-\frac{1}{2})
\end{aligned}
$$

together with

$$
\begin{aligned}
&0 \le \lfloor m \rfloor + \tfrac{5}{2} - u \le 1 \\
&m + \tfrac{3}{2} - u \ge m + \frac{3}{2} - (m+1) = \frac{1}{2}.
\end{aligned}
$$

Therefore, for all $m > 1/2$, P_m has a unique zero u_m^* in $[m+\frac{1}{2}, \min\{m+1, \lfloor m \rfloor + \frac{3}{2}\}]$, and moreover $P_m \ge 0$ on $(0, u_m^*)$ and $P_m < 0$ on (u_m^*, ∞). Then

$$
\begin{aligned}
\partial_s\, \bar{\mathcal{E}}^{\infty}(m,s) &= 4s^4 \int_0^{\infty} u P_m(u) \varphi(su)\, du \\
&= \frac{4s^4}{\sqrt{2\pi}} e^{-s^2 u_m^{*2}/2} \int_0^{\infty} u P_m(u) e^{-s^2(u^2 - u_m^{*2})/2} du
\end{aligned}
$$

Notice now that $\frac{4s^4}{\sqrt{2\pi}} e^{-s^2 u_m^{*2}/2} > 0$ for all m, s, and $\int_0^{\infty} u P_m(u) e^{-s^2(u^2-u_m^{*2})/2} du$ is strictly increasing in s, since its derivative in s is $\int_0^{\infty} u\, s\, (u_m^{*2} - u^2) P_m(u) e^{-s^2(u^2-u_m^{*2})/2} du$ and the integrand is nonnegative and not identically zero, as $(u_m^{*2} - u^2)$ and $P_m(u)$ have

the same sign. Therefore, for all $m > 1/2$, $s \mapsto \bar{\mathcal{E}}_s^{\infty}(m,s)$ has at most one zero. For all $m > 1/2$, a minimum exists because

$$\begin{aligned}\bar{\mathcal{E}}^{\infty}(m,s) &= 1 - 4s^3 \int_0^{\infty} D_m(u)\varphi(su)du \\ &\leq 1 - 4s^3 \int_{1/2}^{\infty} D_m(u)\varphi(su)du \\ &\leq 1 - 4s^3 \int_{m+\frac{1}{2}}^{\infty} m\varphi(su)du \\ &< 1\end{aligned}$$

while $\lim_{s\to 0} \bar{\mathcal{E}}^{\infty}(m,s) = \lim_{s\to\infty} \bar{\mathcal{E}}^{\infty}(m,s) = 1$; indeed, $D_m(u) = 0$ for $u < \frac{1}{2}$ and $D_m(u) \leq m(u+\frac{1}{2})$, so that $4s^3\int_0^{\infty} D_m(u)\varphi(su)\,du \leq 4ms^3\int_{1/2}^{\infty}(u+\frac{1}{2})\varphi(su)\,du$, which tends to 0 both as $s \to 0$ and as $s \to \infty$.

We note now that at the critical point $s^*(m)$ we have

$$\begin{aligned}0 &= \bar{\mathcal{E}}_s^{\infty}(m, s^*(m)) \\ &= \frac{4s^*(m)^4}{\sqrt{2\pi}} e^{-s^{*2}(m)u_m^{*2}/2} \int_0^{\infty} uP_m(u)e^{-s^{*2}(m)(u^2-u_m^{*2})/2}du \\ &\implies \int_0^{\infty} uP_m(u)e^{-s^*(m)^2(u^2-u_m^{*2})/2}\,du = 0\,.\end{aligned}$$

Finally, at $s^*(m)$, it holds

$$\begin{aligned}&\bar{\mathcal{E}}_{ss}^{\infty}(m, s^*(m)) \\ &\quad= \frac{4s^*(m)^4}{\sqrt{2\pi}} e^{-s^{*2}(m)u_m^{*2}/2} \frac{d}{ds}\left(\int_0^{\infty} uP_m(u)e^{-s^2(u^2-u_m^{*2})/2}du\right)\Big|_{s=s^*(m)} \\ &\quad= \frac{4s^*(m)^4}{\sqrt{2\pi}} e^{-s^{*2}(m)u_m^{*2}/2} \int_0^{\infty} u\, s^*(m)(u_m^{*2}-u^2)P_m(u)e^{-s^{*2}(m)(u^2-u_m^{*2})/2}du \\ &\quad> 0\end{aligned}$$

since $(u_m^{*2} - u^2)$ and $P_m(u)$ have the same sign for all $u \in (0,\infty)$, and their product is not identically zero.

(iii) The function $z \mapsto \big(z - s\lfloor z/s \rceil\big)^2$ is even and s-periodic, and equals z^2 on $(-\frac{s}{2}, \frac{s}{2})$; its Fourier expansion is therefore the rescaled expansion of x^2 on $(-\pi,\pi)$, namely

$$\big(z - s\lfloor z/s \rceil\big)^2 = \frac{s^2}{12} + \frac{s^2}{\pi^2}\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}\cos\Big(\frac{2\pi kz}{s}\Big),$$

with absolute and uniform convergence. Taking expectations with respect to $Z \sim \mathcal{N}(0,1)$ and using $\mathbb{E}[\cos(tZ)] = e^{-t^2/2}$ with $t = 2\pi k/s$ yields

$$\begin{aligned}\mathcal{J}_{\infty}(s) = \mathbb{E}\Big[\big(Z - s\lfloor Z/s \rceil\big)^2\Big] &= \frac{s^2}{12} + \frac{s^2}{\pi^2}\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}\, e^{-2\pi^2k^2/s^2} \\ &= \frac{s^2}{12} + \frac{s^2}{\pi^2}\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2} e^{-k^2\tau(s)}\,,\end{aligned}$$

where $\tau(s) := \frac{2\pi^2}{s^2}$. Therefore, for all $s > 0$, we have that

$$\begin{aligned}
\mathcal{J}'_\infty(s) &= \frac{s}{6} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\frac{d}{ds}\left(s^2 e^{-k^2\tau(s)}\right) \\
&= \frac{s}{6} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left[2se^{-k^2\tau(s)} - s^2 e^{-k^2\tau(s)}k^2\tau'(s)\right] \\
&= \frac{s}{6} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left[2se^{-k^2\tau(s)} + s^2 e^{-k^2\tau(s)}\frac{4\pi^2k^2}{s^3}\right] \\
&= \frac{s}{6} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}2s\left[1 + \frac{2\pi^2k^2}{s^2}\right]e^{-k^2\tau(s)} \\
&= \frac{s}{6} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}(-1)^k 2s\left[\frac{1}{k^2} + \tau(s)\right]e^{-k^2\tau(s)} \\
&< \frac{s}{6}\,,
\end{aligned}$$

since $\left[\frac{1}{k^2} + \tau(s)\right]e^{-k^2\tau(s)}$ is strictly decreasing to 0 in k and $-(1+\tau)e^{-\tau} < 0$. Recall now that $g(x) = \varphi(x) - x(1-\Phi(x))$ and $g'(x) = -x\varphi(x) - (1-\Phi(x)) + x\varphi(x) = -(1-\Phi(x))$, hence

$$\begin{aligned}
\bar{\mathcal{E}}^\infty_s(m,s) = \mathcal{J}'_\infty(s) + 4\sum_{j=0}^{\infty}\Bigg[&\varphi\left(\left(m+\frac{1}{2}+j\right)s\right) - \\
&2s\left(m+\frac{1}{2}+j\right)\left(1-\Phi\left(\left(m+\frac{1}{2}+j\right)s\right)\right)\Bigg]\,.
\end{aligned} \tag{24}$$

We note that, whenever $y \geq \sqrt{2}$, we have

$$\varphi(y) - 2y(1-\Phi(y)) < \frac{1-y^2}{1+y^2}\varphi(y) \leq -\frac{1}{3}\varphi(y) \tag{25}$$

where we have used Lemma 12. At $s = \frac{\sqrt{2}}{m+\frac{1}{2}}$, it holds $\left(m+\frac{1}{2}+j\right)\frac{\sqrt{2}}{m+\frac{1}{2}} \geq \sqrt{2}$ for all $j \geq 0$. Therefore, combining the fact that all the terms in the sum appearing in (24) are negative and (25), and keeping only the term $j = 0$, we obtain

$$\bar{\mathcal{E}}^\infty_s\left(m, \frac{\sqrt{2}}{m+\frac{1}{2}}\right) < \frac{\sqrt{2}}{6m+3} - \frac{4}{3}\varphi(\sqrt{2}) \leq \frac{\sqrt{2}}{9} - \frac{4}{3}\varphi(\sqrt{2}) < 0$$

for all $m \geq 1$, since $\frac{\sqrt{2}}{9}$ and $\frac{4}{3}\varphi(\sqrt{2}) = \frac{4}{3}\frac{e^{-1}}{\sqrt{2\pi}}$. By (ii) of Proposition 1, the function $\bar{\mathcal{E}}^\infty_s(m,\cdot)$ has a unique zero in $s^*(m)$ with $\bar{\mathcal{E}}^\infty_{ss}(m, s^*(m)) > 0$. By continuity it holds that

$$\begin{aligned}
\bar{\mathcal{E}}^\infty_s(m,s) &< 0 \qquad s < s^*(m) \\
\bar{\mathcal{E}}^\infty_s(m,s) &> 0 \qquad s > s^*(m)\,,
\end{aligned}$$

and, since $\bar{\mathcal{E}}^\infty_s\left(m, \frac{\sqrt{2}}{m+\frac{1}{2}}\right) < 0$, we know that $\frac{\sqrt{2}}{m+\frac{1}{2}} < s^*(m)$. Let us call now $y_j = (m+j+\frac{1}{2})s^*(m)$. Since $y_j = (m+j+\frac{1}{2})s^*(m) > \frac{\sqrt{2}}{m+\frac{1}{2}}(m+j+\frac{1}{2}) \geq \sqrt{2}$, for all $j \geq 0$, we have that

$$y_j\varphi(y_j) - 2(1-\Phi(y_j)) \overset{(*)}{>} \left(y_j - \frac{2}{y_j}\right)\varphi(y_j) > 0\,,$$

where in (*) we have used again Lemma 12. Since $g' = -(1-\Phi)$, we have $\bar{\mathcal{E}}^\infty_m(m,s) = -4s^2\sum_{j=0}^{\infty}\left(1 - \Phi\big((m+j+\frac{1}{2})s\big)\right)$, and differentiating in s gives $\bar{\mathcal{E}}^\infty_{ms}(m,s) =$

$4s\sum_{j=0}^{\infty}\big[x_j\varphi(x_j)-2(1-\Phi(x_j))\big]$ with $x_j=(m+j+\frac{1}{2})s$. Therefore,

$$\bar{\mathcal{E}}^{\infty}_{ms}(m,s^*(m))=4s^*(m)\sum_{j=0}^{\infty}\Big[y_j\varphi(y_j)-2(1-\Phi(y_j))\Big]>0\,. \tag{26}$$

By the implicit function theorem, since $\bar{\mathcal{E}}^{\infty}_{ss}\big(m,s^*(m)\big)>0$ for every $m\geq 1$, we have that $s^*(\cdot)\in\mathcal{C}^{\infty}([1,\infty))$, and

$$\frac{ds^*}{dm}(m)=-\frac{\bar{\mathcal{E}}^{\infty}_{ms}\big(m,s^*(m)\big)}{\bar{\mathcal{E}}^{\infty}_{ss}\big(m,s^*(m)\big)}. \tag{27}$$

Combining (26), (27) and $\bar{\mathcal{E}}^{\infty}_{ss}\big(m,s^*(m)\big)>0$, we conclude that

$$\frac{ds^*}{dm}(m)=-\frac{\bar{\mathcal{E}}^{\infty}_{ms}\big(m,s^*(m)\big)}{\bar{\mathcal{E}}^{\infty}_{ss}\big(m,s^*(m)\big)}<0\,.$$

□

A.5.1 Computation of $\kappa(m)$ for $m>0$ integer

Let $y_j:=j+\frac{1}{2}$. We recall that

$$\bar{\mathcal{E}}^{\infty}(m,s)=1-4\sum_{j=0}^{\infty}s\,g(y_js)+4\sum_{j=0}^{\infty}s\,g\big((m+j+\tfrac{1}{2})s\big)\,,$$

where $g(x)=\varphi(x)-x\big(1-\Phi(x)\big)$. Since we are interested in integer m, we note that the two series telescope, because $m+j+\frac{1}{2}=y_{j+m}$, and the expression reduces to the finite sum

$$\mathcal{E}^{\infty}(m,s)=1-4s\sum_{j=0}^{m-1}g(y_js)\,.$$

Moreover, it holds $g'(x)=-(1-\Phi(x))$ and $g''(x)=\varphi(x)$. Then, for all $c>0$, it holds that

$$\begin{aligned}\big[s\,g(cs)\big]'&=g(cs)+cs\,g'(cs)\\&=\varphi(cs)-cs\big(1-\Phi(cs)\big)-cs\big(1-\Phi(cs)\big)\\&=\varphi(cs)-2cs\big(1-\Phi(cs)\big)\,,\end{aligned}$$

and

$$\begin{aligned}\big[s\,g(cs)\big]''&=c\,g'(cs)+c\,g'(cs)+c^2s\,g''(cs)\\&=-2c\big(1-\Phi(cs)\big)+c^2s\,\varphi(cs)\\&=c\Big[cs\,\varphi(cs)-2\big(1-\Phi(cs)\big)\Big].\end{aligned}$$

In particular, a direct computation leads to

$$\begin{aligned}\partial_s\,\mathcal{E}^{\infty}(m,s)&=-4\sum_{j=0}^{m-1}\big[s\,g(y_js)\big]'\\&=4\sum_{j=0}^{m-1}\Big[2y_js\big(1-\Phi(y_js)\big)-\varphi(y_js)\Big]\,.\end{aligned} \tag{28}$$

and

$$\begin{aligned}\mathcal{E}^{\infty}_{ss}(m,s)&=-4\sum_{j=0}^{m-1}\big[s\,g(y_js)\big]''\\&=-4\sum_{j=0}^{m-1}y_j\Big[y_js\,\varphi(y_js)-2\big(1-\Phi(y_js)\big)\Big]\,,\end{aligned}$$

and, consequently,

$$\kappa(m,s) = s^2\,\mathcal{E}^{\infty}_{ss}(m,s) = 4s\sum_{j=0}^{m-1}\Big[2y_j s\big(1-\Phi(y_j s)\big) - (y_j s)^2\,\varphi(y_j s)\Big].$$

The following algorithm is used for the numerical evaluation of the scale-invariant curvature $\kappa(M_B, s^*(M_B))$ reported in Figure 4.

Algorithm 1 Computation of $\kappa(M_B, s^*(M_B))$ for $B \in \{2,3,4,5,8,9\}$. The functions `E_s` and `kappa` implement the finite sums of Section A.5.1; `s_star` locates the unique zero of $\mathcal{E}^{\infty}_{s}(m,\cdot)$ by Brent's method on the bracket $[\sqrt{2}/(m+\frac{1}{2}), 20]$.

```
import numpy as np
from scipy.stats import norm
from scipy.optimize import brentq

phi, Q = norm.pdf, norm.sf   # Q = 1 - Phi

def E_s(m, s):
    t = (np.arange(m) + 0.5) * s
    return -4 * np.sum(phi(t) - 2 * t * Q(t))

def kappa(m, s):
    t = (np.arange(m) + 0.5) * s
    return 4 * s * np.sum(2 * t * Q(t) - t**2 * phi(t))

def s_star(m):
    lo, hi = np.sqrt(2) / (m + 0.5), 20.0
    return brentq(lambda s: E_s(m, s), lo, hi, xtol=1e-12)

bits = [2, 3, 4, 5, 8, 9]
for B in bits:
    m = 2 ** (B - 1) - 1
    s = s_star(m)
    print(B, m, s, kappa(m, s))
```

The following figure shows the scale-invariant curvature κ of the limiting objective $s \mapsto \mathcal{E}^{\infty}(M_B, s)$, evaluated at its minimizer $s^*(M_B)$, as a function of the bit-width B. In particular, we note that κ decreases monotonically with B, so the objective is sharply peaked around the optimal scale at low precision and increasingly flat as B grows, which is why the choice of scale matters most at W2 and W3.

A.5.2 Why relative curvature κ?

The scale-invariant curvature κ is a natural quantity for studying sensitivity to relative scale errors. To illustrate this, fix the bit-width and let $\vartheta \sim \mathcal{N}(0, I_d)$ and $\vartheta_\sigma = \sigma\vartheta$, with $\sigma > 0$. Define the variance-normalized loss

$$\mathcal{E}_\sigma(s) := \frac{1}{d\sigma^2}\mathbb{E}\Big[\big\|\vartheta_\sigma - s\,\lfloor \vartheta_\sigma/s \rceil_B\big\|_2^2\Big].$$

Since $\vartheta_\sigma/s = \vartheta/(s/\sigma)$, we obtain

$$\mathcal{E}_\sigma(s) = \mathcal{E}_1(s/\sigma), \qquad s^*_\sigma = \sigma s^*,$$

where s^* minimizes $\mathcal{E}_1$. Consequently, the same relative scale error ε produces the same variance-normalized loss:

$$\mathcal{E}_\sigma(s^*_\sigma(1+\varepsilon)) = \mathcal{E}_1(s^*(1+\varepsilon)).$$

Although the ordinary second derivative depends on σ,

$$\mathcal{E}''_\sigma(s^*_\sigma) = \sigma^{-2}\mathcal{E}''_1(s^*),$$

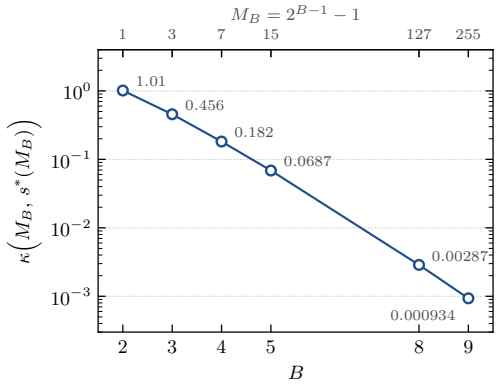


Figure 4: Scale-invariant curvature $\kappa(M_B, s^*(M_B))$ of $\mathcal{E}^\infty(M_B, s^*(M_B))$. κ is decreasing as B increases.

the scale-invariant curvature does not:

$$\kappa = (s_\sigma^*)^2 \mathcal{E}_\sigma''(s_\sigma^*) = (s^*)^2 \mathcal{E}_1''(s^*).$$

It therefore governs the local loss increase under relative scale perturbations:

$$\mathcal{E}_\sigma(s_\sigma^*(1+\varepsilon)) - \mathcal{E}_\sigma(s_\sigma^*) = \frac{\kappa}{2}\varepsilon^2 + o(\varepsilon^2).$$

This invariance motivates using relative perturbations to assess sensitivity to scale choices, including those produced by the rules in Table 5.

In practice, the scale-selection rules in Table 5 return different positive scales for the same quantization setting. Their discrepancies can be expressed as relative deviations from a common reference scale:

$$s_{\text{rule}} = s_{\text{ref}}(1 + \varepsilon_{\text{rule}}).$$

Relative perturbations therefore provide a common way to assess the effect of these differences across weight magnitudes. When the reference scale minimizes the smooth normalized loss, κ governs the resulting loss increase up to second order.

A.6 Technical Lemmas

The following elementary fact allows us to control the supremum of a quadratic polynomial over an interval by its values at three points.

Lemma 10. *Let $a < b$, $m := \frac{a+b}{2}$, and let p be a real polynomial of degree at most 2. Then*

$$\sup_{s\in[a,b]} |p(s)| \;\le\; 3\max\left\{|p(a)|,\, |p(m)|,\, |p(b)|\right\}.$$

Proof. Define the Lagrange basis polynomials associated with the nodes a, m, b,

$$L_a(s) := \frac{(s-m)(s-b)}{(a-m)(a-b)}, \qquad L_m(s) := \frac{(s-a)(s-b)}{(m-a)(m-b)}, \qquad L_b(s) := \frac{(s-a)(s-m)}{(b-a)(b-m)}.$$

Each has degree 2, equals 1 at its own node and 0 at the other two. Hence $q := p(a)L_a + p(m)L_m + p(b)L_b$ is a polynomial of degree at most 2 that agrees with p at a, m and b; the difference $p - q$ has degree at most 2 and three distinct roots, so $p = q$ identically, i.e.

$$p(s) = p(a)\, L_a(s) + p(m)\, L_m(s) + p(b)\, L_b(s) \qquad \text{for all } s \in \mathbb{R}. \tag{29}$$

We now bound the basis polynomials on $[a, b]$. The affine change of variable $s = m + \frac{b-a}{2}\, x$ maps $[a, b]$ onto $x \in [-1, 1]$ and the nodes a, m, b to $-1, 0, 1$, and transforms the basis polynomials into

$$L_a = \frac{x(x-1)}{2}, \qquad L_m = 1 - x^2, \qquad L_b = \frac{x(x+1)}{2}.$$

On $[-1,1]$ the function $x \mapsto x(x-1)$ attains its minimum $-\frac{1}{4}$ at $x = \frac{1}{2}$ and its maximum 2 at $x = -1$, so $|L_a| \le 1$; clearly $L_m \in [0,1]$; and $|L_b| \le 1$ by the symmetry $x \mapsto -x$. Therefore, by (29) and the triangle inequality, for every $s \in [a,b]$,

$$\begin{aligned} |p(s)| &\le |p(a)|\,|L_a(s)| + |p(m)|\,|L_m(s)| + |p(b)|\,|L_b(s)| \\ &\le |p(a)| + |p(m)| + |p(b)| \\ &\le 3\max\big\{|p(a)|, |p(m)|, |p(b)|\big\}, \end{aligned}$$

and the claim follows by taking the supremum over $s \in [a,b]$. □

We will control operator norms through the following standard ε-net argument, which reduces the supremum over the unit spheres to a maximum over a finite set of cardinality exponential in the dimension (see (**?**, Sec. 4.4)).

Lemma 11. *Let $p, q \ge 1$. There exist $\frac{1}{4}$-nets $\mathcal{U} \subset \mathbb{S}^{p-1}$ and $\mathcal{V} \subset \mathbb{S}^{q-1}$ (with respect to the Euclidean norm) with $|\mathcal{U}| \le 9^p$ and $|\mathcal{V}| \le 9^q$, and for every $M \in \mathbb{R}^{p\times q}$*

$$\|M\|_{\mathrm{op}} \;\le\; 2 \max_{(u,v)\in\mathcal{U}\times\mathcal{V}} |u^\top M v|.$$

Proof. The cardinality bound is the standard volumetric estimate $\mathcal{N}(\mathbb{S}^{n-1}, \varepsilon) \le (1 + 2/\varepsilon)^n$ with $\varepsilon = \frac{1}{4}$. Let $u \in \mathbb{S}^{p-1}$, $v \in \mathbb{S}^{q-1}$ and pick $u_0 \in \mathcal{U}$, $v_0 \in \mathcal{V}$ with $\|u - u_0\|_2 \le \frac{1}{4}$, $\|v - v_0\|_2 \le \frac{1}{4}$. Since $u^\top M v - u_0^\top M v_0 = (u - u_0)^\top M v + u_0^\top M (v - v_0)$,

$$|u^\top M v| \;\le\; |u_0^\top M v_0| + \big(\|u - u_0\|_2 + \|v - v_0\|_2\big)\,\|M\|_{\mathrm{op}} \;\le\; \max_{(a,b)\in\mathcal{U}\times\mathcal{V}} |a^\top M b| + \tfrac{1}{2}\,\|M\|_{\mathrm{op}}.$$

Taking the supremum over u, v and rearranging (all quantities are finite) gives the claim. □

Our analysis of the loss landscape of $\mathcal{E}^\infty$ repeatedly requires sharp two-sided bounds on the Gaussian tail $1 - \Phi$, which are most conveniently expressed through the inverse Mills ratio $\lambda := \varphi/(1-\Phi)$. The lemma below collects the elementary properties of λ that we shall need in the following lemma.

Lemma 12. *Let $\lambda(t) := \varphi(t)/\big(1 - \Phi(t)\big)$ for $t \ge 0$. Then*

(i) $\lambda \in \mathcal{C}^\infty([0,\infty))$ and $\lambda(0) = 2\varphi(0) = \sqrt{2/\pi}$;

(ii) $\lambda'(t) = \lambda(t)\big(\lambda(t) - t\big)$ for all $t \ge 0$;

(iii) $t < \lambda(t) < t + \dfrac{1}{t}$ for all $t > 0$; equivalently, $\dfrac{t}{1+t^2}\,\varphi(t) < 1 - \Phi(t) < \dfrac{\varphi(t)}{t}$.

Proof. (i) is clear since $1 - \Phi > 0$ is smooth. For (ii), using $\varphi'(t) = -t\varphi(t)$ and $(1-\Phi)' = -\varphi$,

$$\lambda'(t) = \frac{-t\varphi(t)\big(1-\Phi(t)\big) + \varphi(t)^2}{\big(1-\Phi(t)\big)^2} = -t\,\lambda(t) + \lambda(t)^2\,.$$

For (iii), the upper bound on the tail follows from $1 - \Phi(t) = \int_t^\infty \varphi(u)\,du < \int_t^\infty \frac{u}{t}\varphi(u)\,du = \varphi(t)/t$. For the lower bound let $h(t) := 1 - \Phi(t) - \frac{t}{1+t^2}\varphi(t)$; then $h(t) \to 0$ as $t \to \infty$ and a direct computation gives $h'(t) = -\frac{2\varphi(t)}{(1+t^2)^2} < 0$, so $h > 0$ on $(0,\infty)$. Rearranging the two tail bounds gives $t < \lambda(t) < t + 1/t$. □

B Experiments

In this section, we recall some quantization algorithm and preprocessing that are used for the experiments. Moreover, we report additional experiments that assess what our theory captures.

B.1 Experimental Details

This appendix specifies the setup of all experiments in Section 5 and Appendix B. The released code (see the Reproducibility Statement) contains a frozen configuration file for every run, and its `REPRODUCE.md` gives the commands for the five experiment families below. Table 3 lists where the results of each family are reported. As in Section 5, WB denotes weight-only quantization at B bits.

Table 3: Experiment families and their reported results.

Section	Content	Reported in
B.2	Global scale sensitivity and ablations	Figures 1, 2, 5, 6, 7, 8, and 9
B.3	Scale selection and zero-shot accuracy	Tables 1, 2, 6, 7, and 8
B.4	Preprocessing and effective rank	Figure 10 and Table 9
B.5	Gaussian landscapes and finite-width convergence	Figures 11 and 12
B.6	Local reconstruction sensitivity and empirical landscapes	Figures 3, 13, 14, and 15

B.1.1 Models

We quantize five publicly available decoder-only models (Table 4). Each is loaded in fp16 at a pinned Hugging Face revision with a context length of 2,048 tokens. OPT-125M is converted locally from the pinned PyTorch checkpoint to safetensors. Every linear layer inside the transformer blocks is quantized: the attention projections `q`, `k`, `v`, `o`, and the MLP projections (`gate`, `up` and `down` for Llama and Qwen; `fc1` and `fc2` for OPT). The quantization of each of this module exactly matches the quantization of $\vartheta^{(\ell)}$ in the sense of Section 3, calibrated on its own input activations. The token embeddings and the language-model head stay in fp16. The short names in Table 4 are those used in the figures.

Table 4: Models, short names used in the figures, and pinned checkpoint revisions.

Model	Short name	Hugging Face repository	Revision
OPT-125M	OPT-125M	`facebook/opt-125m`	`27dcfa74d334`
Llama-3.2-1B-Instruct	Llama-1B	`meta-llama/Llama-3.2-1B-Instruct`	`9213176726f5`
Llama-3.2-3B-Instruct	Llama-3B	`meta-llama/Llama-3.2-3B-Instruct`	`0cb88a4f764b`
Llama-3.1-8B-Instruct	Llama-8B	`meta-llama/Llama-3.1-8B-Instruct`	`0e9e39f249a1`
Qwen3-8B	Qwen3-8B	`Qwen/Qwen3-8B`	`b968826d9c46`

B.1.2 Quantization setup

Grid. All experiments are weight-only; activations remain in fp16. A WB weight is quantized with the round-to-nearest map Π_{s,M_B} of (1), i.e. onto the symmetric uniform grid $s\{-M_B, \dots, M_B\}$ with $M_B = 2^{B-1} - 1$ levels per side and $2^B - 1$ levels in total, with clipping to $\pm M_B s$. At $B = 2$ this is the ternary grid $\{-s, 0, s\}$. We use $B \in \{2, 3, 4, 6, 8\}$. One fp16 scale is stored per quantization group, and there are no zero-points. Per-channel W3 on Llama-3.1-8B therefore costs 3.003 bits per weight.

Granularity. *Per-channel* quantization uses one scale per output row, which gives the row-wise problems studied in Section 4. *Grouped* quantization uses one scale per contiguous group of $g \in \{512, 256, 128, 64\}$ input columns. With activation ordering, the groups are formed after the column permutation.

GPTQ. Layers are quantized sequentially, block by block. Within a block the order is $\{\mathtt{q}, \mathtt{k}, \mathtt{v}\} \to \mathtt{o} \to \{\mathtt{up}, \mathtt{gate}\} \to \mathtt{down}$. Each block is calibrated on the outputs of the already-quantized preceding blocks ("carry" propagation), so the activations $X^{\ell-1}$ in (2) come from the partially quantized network. The Hessian proxy is

$$H = \frac{2}{N_c} \sum_{n=1}^{N_c} X_n X_n^\top,$$

where $X_n \in \mathbb{R}^{d \times 2048}$ holds the layer inputs for calibration window n. Up to the constant factor, which cancels in the normalized loss of Section 4 and in every scale rule below, this is the Gram matrix $XX^\top$ of the theory. H is accumulated in fp32 over $N_c = 128$ windows with a forward batch of 4 windows. Before inverting H we add $\lambda \,\mathrm{Id}_d$ with $\lambda = 0.01 \cdot \mathrm{mean}(\mathrm{diag}\, H)$. Columns are processed in order of decreasing H_{ii} (activation ordering) in lazy batches of 128 columns, using the inverse-Cholesky update of the reference implementation (Frantar et al., 2022). Columns with $H_{ii} = 0$ are set to zero. For grouped quantization the scale of a group is selected on the error-compensated weights at the point where the algorithm enters the group.

Other engines. *ResComp* (Li et al., 2026) runs GPTAQ-style (Li et al., 2025) paired compensation plus a residual correction toward the full-precision weights, with $\alpha_{\text{GPTAQ}} = \alpha_{\text{ResComp}} = 0.25$. It uses the stability mode of the reference implementation: `org` at W2 and `allw` otherwise. *QRoNoS* (Zhang et al., 2026) re-solves the full layer at the first column and uses "reset" propagation: each block is calibrated on full-precision inputs. Both engines use the damping, activation ordering and block size above; for grouped rows the compensation block is $\min(128, g)$. Because these engines use paired inputs, they additionally admit the rule RTNH-cross, which searches s over the grid of Appendix B.1.3 to minimize $\|WX_{\text{fp}} - \Pi_{s,M_B}(W)\, X_q\|_F^2$, where X_{fp} and X_q are the full-precision and quantized-prefix inputs. We run it for $B \le 3$ only.

Hadamard incoherence processing (HIP). HIP (Appendix B.4) is applied independently to the input dimension of every matrix, $W \mapsto WR$ with $R = \mathrm{diag}(\xi)\, Q$, where Q is a normalized Hadamard matrix ($Q^\top Q = \mathrm{Id}_d$) from the QuIP# implementation (Tseng et al., 2024) at a pinned commit, and $\xi \in \{-1, 1\}^d$ are Rademacher signs drawn from a seed and the matrix name. The Hessian is transformed as $H \mapsto R^\top H R$, which leaves its spectrum, and hence $r_{\text{eff}}(H)$, unchanged (Appendix B.4). At inference the layer input is rotated online by $R^\top$, so the network function is unchanged before quantization. Unless stated otherwise the rotation seed is 0.

B.1.3 Scale-selection rules

Let w denote the (error-compensated) weights of one row or group, and let $s^e = \max_q |w_q|/M_B$ be its min-max scale (Section 3). Every rule outputs one scale s per row or group (Table 5). Min-max and Analytic require no search; Shrink-2.4 and WMSE search without activation data; Proxy-static and RTNH use the Hessian.

Table 5: Scale-selection rules. $e(s) = \Pi_{s,M_B}(w) - w$ is the rounding error at scale s, and H_{gg} is the undamped Hessian block of the row or group. The search window for η depends on the experiment (Appendix B.1.3).

Rule	Scale	Objective
Min-max	s^e	none
Shrink-2.4	$p\, s^e$, $p \in \{1.00, 0.99, \ldots, 0.21\}$	$\sum_q \lvert e_q(p\, s^e)\rvert^{2.4}$
WMSE	$s^e e^\eta$, grid search	$\sum_q e_q(s)^2$
Proxy-static (ours)	$s^e e^\eta$, grid search	$\sum_q e_q(s)^2 / U_{qq}^2$ (GPTQ step loss)
RTNH	$s^e e^\eta$, grid search	$e(s)^\top H_{gg}\, e(s)$
Analytic	$s^*(M_B)\, \hat{\sigma}$	none; $\hat{\sigma}$ is the RMS of w

Min-max and Shrink-2.4. Min-max is the symmetric min-max choice s^e of Section 3. Shrink-2.4 is the clipping search of the GPTQ reference quantizer, applied to our symmetric $(2^B - 1)$-level grid: it scores the 80 shrink fractions $p_i = 1 - i/100$, $i = 0, \ldots, 79$, of the min-max range by the $\ell^{2.4}$ rounding error.

Analytic scale. $s^*(M_B)$ is the unique minimizer of the limiting objective $\mathcal{E}^\infty(M_B, \cdot)$ (Proposition 1), obtained as the zero of $\partial_s \mathcal{E}^\infty(M_B, \cdot)$ in the closed form (28) by bisection. This gives $s^*(M_B) = 1.2240,\ 0.6508,\ 0.3534,\ 0.1055,\ 0.0309$ for $B = 2, 3, 4, 6, 8$ ($M_B = 1, 3, 7, 31, 127$). $\hat{\sigma}$ is computed on the current compensated weights and excludes input columns with $H_{ii} = 0$.

Proxy-static. Let U be the upper Cholesky factor of the inverse damped Hessian in activation order, and let $e_q(s) = \Pi_{s,M_B}(w_q) - w_q$. For each row or group, Proxy-static selects

$$s_{\mathrm{PS}} \in \arg\min_{s \in \mathcal{S}} \sum_q \frac{e_q(s)^2}{U_{qq}^2},$$

where $\mathcal{S}$ is the candidate grid described below. Candidate scores use a fixed snapshot of the weights at group entry, including compensation from previous groups. The objective uses GPTQ's per-column loss weights without simulating candidate-dependent error feedback. After scale selection, quantization proceeds with the usual GPTQ updates. Algorithm 2 summarizes the procedure. We motivate proxy-static by analysing the GPTQ. Our selector is motivated by the loss incurred at each GPTQ quantization step. Let A denote the positive-definite, damped Gram matrix in activation order, and consider a quadratic loss $\mathcal{L}(z) = z^\top A z$. At step q, let $F_q = \{q, \ldots, d\}$ denote the remaining free coordinates, and let $w^{(q)}$ be their current compensated weights. Fixing coordinate q to its quantized value introduces the residual

$$\epsilon_q(s) = w_q^{(q)} - Q_s(w_q^{(q)}).$$

The optimal compensating update solves

$$\min_{\Delta \in \mathbb{R}^{|F_q|}} \Delta^\top A_{F_q F_q} \Delta \quad \text{subject to} \quad \Delta_q = -\epsilon_q(s).$$

where $A_{F_q F_q}$ is the principal submatrix of A indexed by F_q. Writing e_q for the corresponding coordinate vector, the solution and minimum are

$$\Delta^\star = -\frac{\epsilon_q(s)}{[A_{F_q F_q}^{-1}]_{qq}} A_{F_q F_q}^{-1} e_q, \qquad \Delta\mathcal{L}_q(s) = \frac{\epsilon_q(s)^2}{[A_{F_q F_q}^{-1}]_{qq}}.$$

Here the increment is measured from the conditional quadratic minimizer given the previously fixed coordinates.

If $A^{-1} = U^\top U$ with U upper triangular, then

$$[A_{F_q F_q}^{-1}]_{qq} = U_{qq}^2.$$

Thus GPTQ's step loss is $\epsilon_q(s)^2 / U_{qq}^2$. Evaluating these losses for every candidate scale would require recomputing the candidate-dependent compensation trajectory.

Proxy-static instead freezes the weights at group entry. For a row or group g with snapshot v, it selects

$$s_{\mathrm{PS}} \in \arg\min_{s \in \mathcal{S}} \sum_{q \in g} \frac{(v_q - Q_s(v_q))^2}{U_{qq}^2}.$$

This retains GPTQ's per-column loss weights while approximating the sequential residuals by static rounding residuals. The factor U is already available from GPTQ, so candidate scoring requires no additional factorization or candidate-specific compensation pass. After scale selection, the usual GPTQ updates are applied. Algorithm 2 summarizes the procedure.

B.1.4 Compute resources and software

Every run uses a single NVIDIA H100 (94 GB) or H200 (141 GB) GPU.

B.2 Global scale-sensitivity

In this section we add further experiments to understand how different scale-selection rules impact the performance of PTQ of LLMS. This section presents additional experiments examining how scale-selection rules affect the performance of quantized LLMs.

Algorithm 2 GPTQ with Proxy-static scale selection

Require: Weights $W \in \mathbb{R}^{m\times d}$, input Gram H, bit-width $B \geq 2$, group width G, candidate count C, interval $[\eta_{\min}, \eta_{\max}]$, damping fraction λ
Ensure: Quantized weights $\widehat{W}$ and group scales $\{s^\star_{rg}\}$

$K \leftarrow 2^{B-1} - 1$
Define $Q_s(x) = s\ \mathrm{clip}(\mathrm{round}(x/s), -K, K)$
$\mathcal{D} \leftarrow \{j : H_{jj} = 0\}$
$W_{:,\mathcal{D}} \leftarrow 0$; $H_{jj} \leftarrow 1$ for $j \in \mathcal{D}$
Let P sort $\mathrm{diag}(H)$ in descending order
$\widetilde{W} \leftarrow WP^\top$; $\widetilde{H} \leftarrow PHP^\top$
$\widetilde{H} \leftarrow \widetilde{H} + \lambda\ \mathrm{mean}(\mathrm{diag}(\widetilde{H}))I$
Compute upper triangular U satisfying $\widetilde{H}^{-1} = U^\top U$
$(\eta_k)_{k=1}^{C} \leftarrow \mathrm{linspace}(\eta_{\min}, \eta_{\max}, C)$
Replace the η_k closest to zero by 0
for each contiguous group g in activation order **do**
    $V \leftarrow \mathrm{copy}(\widetilde{W}_{:,g})$ ▷ freeze compensated group-entry weights
    for each output row r **do**
        $a \leftarrow \max_{j\in g} |V_{rj}|$
        if $a = 0$ **then**
            $s^\star_{rg} \leftarrow 1$ ▷ any positive scale quantizes zero exactly
        else
            $s^{(0)}_{rg} \leftarrow a/K$
            for $k = 1, \ldots, C$ **do**
                $s_{rgk} \leftarrow s^{(0)}_{rg} \exp(\eta_k)$
                $L_{rgk} \leftarrow \sum_{j\in g} \dfrac{(V_{rj} - Q_{s_{rgk}}(V_{rj}))^2}{U^2_{jj}}$
            end for
            Select $k^\star$ minimizing L_{rgk}, breaking numerical ties toward smallest $|\eta_k|$
            $s^\star_{rg} \leftarrow s_{rgk^\star}$
        end if
    end for
    for columns $j \in g$ in activation order **do**
        $(q_j)_r \leftarrow Q_{s^\star_{rg}}(\widetilde{W}_{rj})$ for every row r
        $e_j \leftarrow (\widetilde{W}_{:,j} - q_j)/U_{jj}$
        $\widetilde{W}_{:,j:d} \leftarrow \widetilde{W}_{:,j:d} - e_j U_{j,j:d}$
        $\widehat{W}^{\mathrm{ord}}_{:,j} \leftarrow q_j$
    end for
end for
$\widehat{W} \leftarrow \widehat{W}^{\mathrm{ord}} P$
return $\widehat{W}$, $\{s^\star_{rg}\}$, and the group assignment induced by P

B.2.1 Sensitivity to scale selection with HIP preprocessing

HIP preprocessing appears to produce a steeper decrease in NLL spread with bit-width, qualitatively similar to the effect of reducing the group size to 128 in GPTQ without HIP. However, as discussed in Remark 3, this observation does not directly imply lower curvature of the GPTQ loss landscape. The spread also depends on how far apart the scales selected by the different rules are. HIP may increase agreement among these choices: suppressing extreme weight coordinates can reduce the influence of outliers on min-max scaling, while a more Gaussian weight distribution may improve agreement between WMSE-based scales and the analytic prediction. This provides a possible additional explanation for the reduced variation in end-to-end performance.

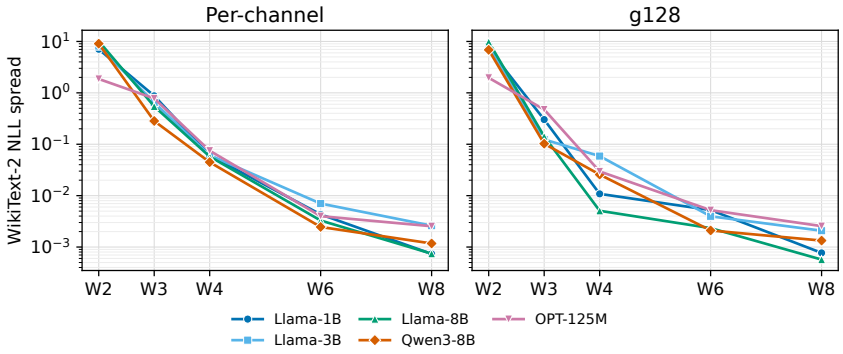


Figure 5: WikiText-2 NLL spread across scale-selection rules for GPTQ with HIP preprocessing, plotted against bit-width for different models. Left: per-channel quantization. Right: group size 128. The spread is the difference between the maximum and minimum NLL across Min-max, Shrink-2.4, WMSE, Proxy-static, and RTNH. The vertical axis uses a logarithmic scale.

B.2.2 Sensitivity to scale selection with GPTQ extensions

To assess whether the trend in Figure 1 extends beyond GPTQ, we repeat the scale-selection comparison with ResComp (Li et al., 2026) and QRoNoS (Zhang et al., 2026). We evaluate both methods at W2, W4, and W8 on Llama-3.2-1B and Llama-3.1-8B, with and without HIP preprocessing.

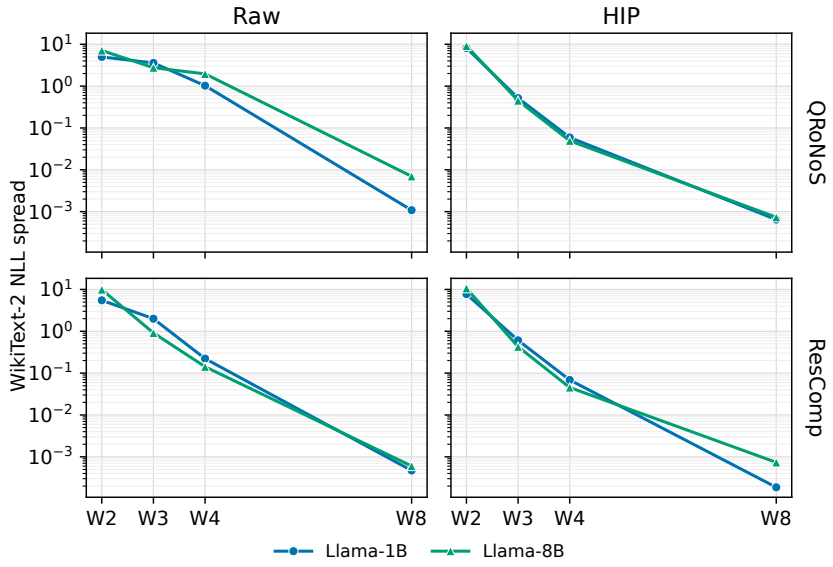


Figure 6: WikiText-2 NLL spread across scale-selection rules for ResComp and QRoNoS, with and without HIP preprocessing, as a function of bit-width. The spread is the difference between the maximum and minimum NLL across the tested scale-selection rules.

Both methods exhibit a decrease in NLL spread with increasing bit-width, qualitatively matching the trend observed for GPTQ. HIP preprocessing is associated with an even steeper decrease. These results suggest that the reduced sensitivity to scale selection at higher precision extends beyond GPTQ to related PTQ methods in the tested settings.

B.2.3 SCALE SENSITIVITY WITHOUT MIN-MAX

To understand whether the bit-width dependence of NLL spread is mainly due to the poor performance of min-max at low precision, we repeat the comparison using only the search based rules Shrink-2.4, WMSE, Proxy-static, and RTNH.

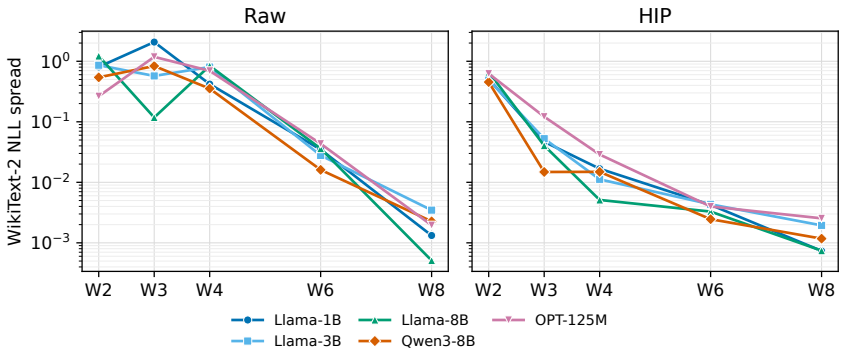


Figure 7: WikiText-2 NLL spread across Shrink-2.4, WMSE, Proxy-static, and RTNH for per-channel GPTQ without HIP (left) and with HIP (right). The spread is the maximum minus minimum NLL across these four rules; min-max is excluded. The vertical axis uses a logarithmic scale.

Figure 7 shows that the broad reduction in NLL spread with increasing precision persists after excluding min-max. Without HIP, the dependence is nonmonotonic between W2 and W4, followed by a pronounced decrease at higher precisions. With HIP, the decrease is more consistent across the tested bit-widths, with a particularly large reduction between W2 and W3. Thus, the greater practical importance of scale selection at low precision persists among the search-based rules and can not be explained by the bad performance of min-max. Performing a scale search alone does not ensure comparable end-to-end performance: the criterion used to select the scale remains important.

B.2.4 END-TO-END SENSITIVITY TO CONTROLLED SCALE PERTURBATIONS

As discussed in Remark 3 the scale-sensitivity plots based on different scale rules do not account for the possibility that discrepancy of the scale rules themselves change with the precision. We therefore add a controlled global sensitivity analysis at a fixed relative perturbation amplitude. We perturb the scales returned by the WMSE search and measure the resulting change in end-to-end performance in NLL. For each bit-width B and channel c, we set

$$s_{B,c}^{(r,\pm)} = s_{B,c}^{0}(1 \pm \delta\epsilon_{r,c}), \qquad \delta = 0.05,$$

where $\epsilon_{r,c} \in \{-1, +1\}$ are independent random signs. We use five random directions, matched across bit-widths within each model, and evaluate both signs of each direction. Each perturbed scale configuration is applied in a fresh sequential GPTQ run from the original floating-point weights, recomputing downstream activations, Gram matrices, and error compensation.

Let $\Delta_{B,r,\pm}$ denote the change in NLL relative to the corresponding unperturbed baseline. We summarize the magnitude of the response by

$$C_B(\delta) = \frac{1}{2R} \sum_{r=1}^{R} \left(|\Delta_{B,r,+}| + |\Delta_{B,r,-}| \right), \qquad R = 5.$$

This statistic captures both improvements and degradations without cancellation between opposite perturbation arms. We interpret $C_B(\delta)$ as finite-perturbation scale sensitivity, not directly as curvature. Away from a stationary point its leading contribution is the directional NLL gradient; near a stationary point, second-order effects can dominate. The absolute values retain improvements as well as degradations, so this statistic measures response magnitude rather than the signed penalty from perturbing the scales. We evaluate per-channel GPTQ without HIP at W2, W3, W4, W6, and W8 on five models, using one calibration draw and evaluating on WikiText-2 and C4.

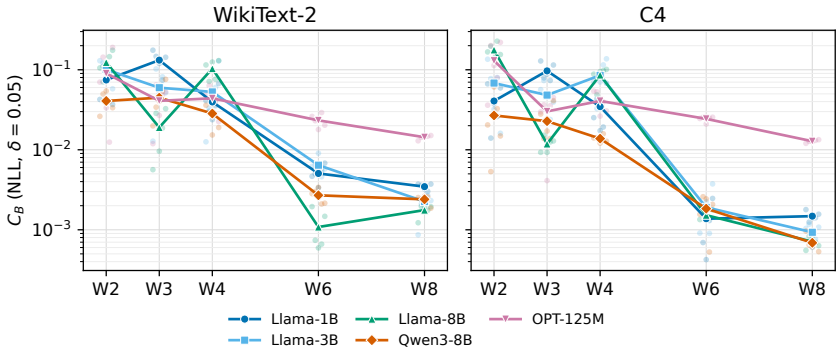


Figure 8: End-to-end sensitivity to controlled scale perturbations. Each channel's WMSE-selected scale is increased or decreased by 5%, followed by a complete sequential GPTQ run. Each point reports the mean absolute NLL change relative to the unperturbed WMSE baseline, averaged over ten perturbed runs corresponding to five random sign patterns and their opposites. Experiments use per-channel GPTQ and one calibration draw; random directions are matched across bit-widths within each model.

Figure 8 shows substantially larger responses at W2–W4 than at W6–W8. Across the Llama and Qwen models and the two evaluation corpora, the reported W4-to-W6 reduction is approximately 8–96-fold. OPT-125M exhibits a weaker reduction and remains more sensitive at high precision. The dependence on bit-width is not monotonic: for example, Llama-3.1-8B is more sensitive at W4 than at W3 on WikiText-2.

These results further support greater practical sensitivity to relative scale perturbations at low precision in the tested setting. They complement the local reconstruction-loss analysis, while showing that its regular bit-width dependence need not neccessarily translate into a monotonic end-to-end response. The interpretation also depends on baseline quality: some W2 models are severely degraded, and several W4 baselines perform worse than their W3 counterparts.

B.2.5 Calibration-Set Size

Our main experiments use 128 calibration sequences of 2048 tokens each. To assess whether our observations depend on this choice, we vary the number of calibration sequences from 16 to 256 while keeping the sequence length fixed. We compare Min-max, Analytic, and Proxy-static on Llama-1B and OPT-125M at W2 and W3, with and without HIP.

Figure 9 shows that calibration-set size affects NLL, sometimes nonmonotonically, but substantial differences between scale-selection rules persist across the tested sizes. In particular, increasing the calibration set does not consistently close the performance gaps between rules. Thus, the practical importance of scale selection at low precision is not specific to our default calibration-set size.

B.3 Zero-shot evaluation

We complement perplexity evaluation with six zero-shot tasks: PIQA, ARC-Easy, ARC-Challenge, HellaSwag, WinoGrande, and BoolQ. All quantized results use per-channel weight-only quantization, calibration seed 0, and 128 C4 sequences of 2048 tokens each. The symmetric grid has $K_B = 2^{B-1} - 1$, so W2 is ternary. *Raw* denotes no preprocessing; HIP uses rotation seed 0. We report `acc_norm` for PIQA, ARC-Easy, ARC-Challenge, and HellaSwag, and `acc` for WinoGrande and BoolQ. Accuracies are percentages, while meanis computed as their unweighted average on the six tasks. WikiText-2 and C4 columns report perplexity. Lower perplexity and higher accuracy are better. Bold entries identify column-wise best reported values within each matched model, precision, preprocessing, engine, and campaign panel, including ties at the displayed precision. The FP16 reference is not bolded. These are single-calibration-seed measurements; differences of a few tenths of an accuracy point do not establish resolved rankings.

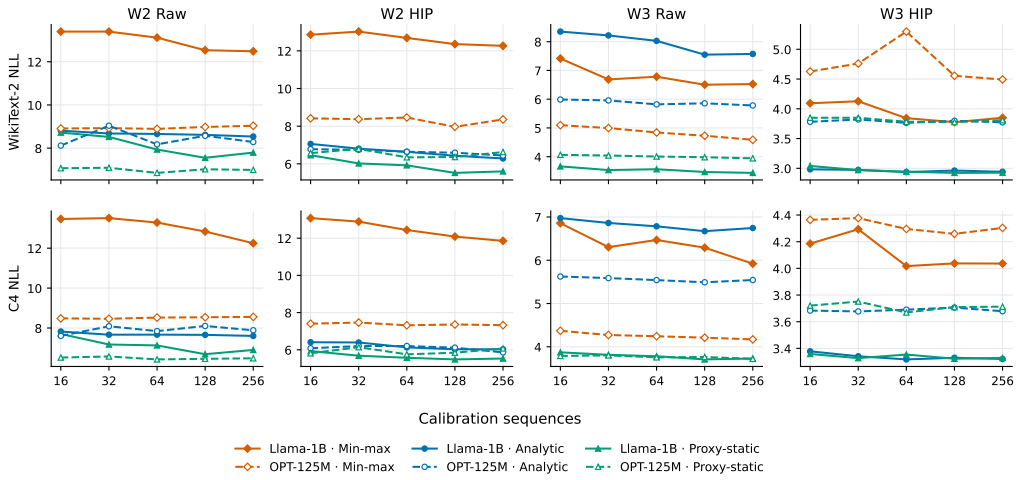


Figure 9: Effect of calibration-set size on WikiText-2 NLL (top) and C4 NLL (bottom). Each plot compares Min-max, Analytic, and Proxy-static for Llama-1B and OPT-125M at W2 and W3, with and without HIP. Each calibration sequence contains 2048 tokens.

W3 scale selection. Table 6 compares four scale-selection rules for GPTQ. With HIP, the search-free analytic rule achieves mean accuracies of 72.08%, 63.75%, and 72.36% on Llama-3.1-8B, Llama-3.2-3B, and Qwen3-8B, respectively, comparable to the searched rules and substantially above min-max. Without HIP, the analytic rule fails severely on all three models. Proxy-static achieves the lowest perplexity on both corpora and the highest six-task mean in each raw W3 comparison. Thus, the relative performance of the selectors depends strongly on preprocessing.

W2 scale selection. Table 7 shows substantial degradation in several W2 configurations, especially without HIP and with min-max. The proxy-static results use the search interval $\eta \in [-1, 1]$. Proxy-static achieves lower perplexity on both corpora than Shrink-2.4 in all six model–preprocessing comparisons, but higher six-task mean accuracy in only three. For example, on Llama-3.1-8B with HIP, proxy-static achieves WikiText-2/C4 perplexity of 31.38/43.81, compared with 39.16/53.99 for Shrink-2.4, while their mean accuracies are 48.50% and 48.69%, respectively. Scale selection therefore affects both metrics, but their rankings need not coincide.

Transfer across quantization engines. Table 8 compares Shrink-2.4 and proxy-static for GPTQ, ResComp, and QRoNoS at W2 with HIP preprocessing. This focused campaign uses 128 log-spaced proxy candidates over $\eta \in [-1.0, 0.1]$. Its search settings differ from the main GPTQ campaign, so overlapping configurations are reported separately. Proxy-static lowers perplexity on both corpora in all six model-engine comparisons and achieves higher mean accuracy in five, although some accuracy differences are small. We omit the corresponding results without HIP because all tested configurations exhibit severe degradation. These findings support proxy-static as a practical scale selector beyond GPTQ, while showing that perplexity improvements do not always translate into accuracy gains.

B.4 Preprocessing and effective rank

HIP Consider a fixed nonzero weight row $w \in \mathbb{R}^d$ and a uniformly random orthogonal matrix R. A random rotation distributes wR uniformly on the sphere of radius $\|w\|_2$. Equivalently,

$$wR \overset{d}{=} \|w\|_2 \frac{g}{\|g\|_2}, \qquad g \sim \mathcal{N}(0, \mathrm{Id}_d).$$

Since $\|g\|_2/\sqrt{d}$ concentrates around one, individual coordinates are approximately Gaussian in high dimensions, with standard deviation $\|w\|_2/\sqrt{d}$, the row's RMS. Applying the corresponding inverse

Table 6: W3 per-channel GPTQ with a symmetric seven-level grid $\{ks : k = -3, \ldots, 3\}$. The proxy-static search uses $\eta \in [-1, 1]$. Panels (a) and (b) show raw and HIP results. FP16 references are excluded from bolding, which identifies the best quantized rule within each model and panel. Lower perplexity and higher accuracy are better.

Model	Rule	Wiki2 PPL	C4 PPL	PIQA	ARC-E	ARC-C	HellaSwag	WinoGrande	BoolQ	Mean
(a) Without preprocessing										
Llama-3.1-8B	FP16 reference	7.213	11.388	80.96	79.59	55.12	79.26	74.03	84.10	75.51
	Shrink-2.4	14.05	17.18	70.46	55.30	36.86	70.76	66.61	79.48	63.25
	Proxy-static	**11.76**	**16.33**	**72.91**	**63.51**	**41.21**	**72.09**	**68.51**	**79.85**	**66.34**
	Analytic	9.3×10^5	6.43×10^5	51.36	25.42	26.79	26.73	50.75	62.17	40.54
	Min-max	41.47	45.51	62.35	40.78	28.24	47.20	51.54	52.42	47.09
Llama-3.2-3B	FP16 reference	11.048	16.486	75.52	67.85	46.16	70.44	67.32	78.62	67.65
	Shrink-2.4	42.48	47.93	69.91	53.75	34.22	59.46	55.96	66.85	56.69
	Proxy-static	**21.16**	**27.55**	**70.29**	**58.46**	**35.92**	**60.93**	**61.96**	**70.95**	**59.75**
	Analytic	2310.17	1027.36	51.52	26.68	25.68	26.68	52.96	39.94	37.24
	Min-max	120.61	73.61	56.20	33.16	26.11	41.56	52.72	52.84	43.77
Qwen3-8B	FP16 reference	9.715	15.363	77.75	80.93	56.48	74.92	67.64	86.61	74.05
	Shrink-2.4	37.95	33.39	66.49	43.39	31.57	50.36	55.56	75.23	53.77
	Proxy-static	**18.51**	**21.39**	**73.23**	**60.14**	**40.61**	**62.68**	**64.40**	**81.83**	**63.82**
	Analytic	3.13×10^{10}	5.99×10^{10}	52.12	25.72	27.22	26.55	49.17	37.83	36.43
	Min-max	22.39	26.90	63.22	41.46	26.54	53.57	52.64	58.65	49.35
(b) HIP preprocessing										
Llama-3.1-8B	FP16 reference	7.213	11.388	80.96	79.59	55.12	79.26	74.03	84.10	75.51
	Shrink-2.4	**9.09**	**14.45**	78.45	74.41	**50.77**	**74.00**	71.67	81.74	71.84
	Proxy-static	9.23	14.62	78.13	75.67	48.81	73.87	**73.32**	81.65	71.91
	Analytic	9.21	14.60	**78.56**	**76.14**	50.60	73.65	71.11	**82.45**	**72.08**
	Min-max	15.40	22.77	70.46	54.38	33.19	62.02	59.51	62.11	56.94
Llama-3.2-3B	FP16 reference	11.048	16.486	75.52	67.85	46.16	70.44	67.32	78.62	67.65
	Shrink-2.4	14.96	20.21	71.76	60.40	38.31	64.33	64.40	75.90	62.52
	Proxy-static	15.01	20.56	**72.52**	62.63	38.05	63.76	64.09	75.66	62.79
	Analytic	**14.59**	**20.06**	72.09	**63.68**	**39.59**	**64.64**	**65.04**	**77.46**	**63.75**
	Min-max	30.21	33.48	59.19	37.25	27.13	54.46	55.49	55.66	48.20
Qwen3-8B	FP16 reference	9.715	15.363	77.75	80.93	56.48	74.92	67.64	86.61	74.05
	Shrink-2.4	10.76	16.78	75.68	73.82	48.55	70.65	68.11	85.26	70.35
	Proxy-static	**10.68**	**16.70**	**76.88**	73.82	49.49	**70.91**	69.30	85.50	70.98
	Analytic	10.87	16.93	76.71	**77.82**	**52.47**	70.90	**70.32**	**85.93**	**72.36**
	Min-max	15.20	20.99	69.80	48.32	32.76	61.49	56.59	68.04	56.17

transformation to the activations preserves the original linear map exactly:

$$\widetilde{W} = WR, \qquad \widetilde{X} = R^\top X, \qquad \widetilde{W}\widetilde{X} = WX.$$

Although the full-precision output is unchanged, coordinatewise quantization in the rotated basis leads to different errors.

For fixed activations X, the orthogonal rotation gives $\widetilde{H} = R^\top H R$. This changes the coordinate system but preserves the Hessian's eigenvalues. Since $H = XX^\top$ is positive semidefinite, its trace is the sum of its eigenvalues and its operator norm is the largest eigenvalue. Both are therefore unchanged, so

$$r_{\text{eff}}(\widetilde{H}) = \frac{\operatorname{tr}(\widetilde{H})}{\|\widetilde{H}\|_{\text{op}}} = \frac{\operatorname{tr}(H)}{\|H\|_{\text{op}}} = r_{\text{eff}}(H).$$

Thus, rotation alone cannot improve the effective rank; changes in upstream quantization may still affect it by changing X. Rotation therefore acts only on the weight distribution, not on the conditioning of the objective. QuIP (Chee et al., 2023) uses exactly such structured random orthogonal preprocessing to reduce weight outliers and obtain incoherence guarantees, substantially improving quantization performance.

Our HIP transform uses efficient randomized Hadamard mixing based on QuIP# Tseng et al. (2024) rather than a full uniformly random rotation. Here, each transformed coordinate is a normalized random signed sum of the original weights, motivating a Gaussian approximation through a central-limit argument when no individual contribution dominates.

Table 7: W2–ternary per-channel GPTQ with grid $\{-s, 0, s\}$. Panels (a) and (b) show raw and HIP results. FP16 references are excluded from bolding, which identifies the best quantized rule within each model and panel. Lower perplexity and higher accuracy are better.

Model	Rule	Wiki2 PPL	C4 PPL	PIQA	ARC-E	ARC-C	HellaSwag	WinoGrande	BoolQ	Mean
(a) Without preprocessing										
Llama-3.1-8B	FP16 reference	7.213	11.388	80.96	79.59	55.12	79.26	74.03	84.10	75.51
	Shrink-2.4	158.59	110.10	53.37	28.58	24.57	**34.08**	50.12	**62.54**	**42.21**
	Proxy-static	**157.51**	**105.48**	**53.50**	**29.00**	23.70	31.30	50.70	55.30	40.60
	Analytic	2.18×10^5	2.37×10^5	51.14	24.12	**27.30**	26.09	51.14	56.61	39.40
	Min-max	7.87×10^5	6.73×10^5	50.76	24.62	26.19	26.66	**51.22**	48.13	37.93
Llama-3.2-3B	FP16 reference	11.048	16.486	75.52	67.85	46.16	70.44	67.32	78.62	67.65
	Shrink-2.4	981.90	639.95	52.23	23.86	26.28	26.93	50.20	53.79	38.88
	Proxy-static	**442.39**	**315.76**	**52.30**	**25.80**	25.80	**27.70**	50.50	**54.40**	**39.40**
	Analytic	9291.88	3934.12	51.25	25.13	25.68	26.47	50.91	38.04	36.25
	Min-max	2.59×10^5	2.47×10^5	51.41	24.16	**27.56**	26.40	**52.25**	47.46	38.21
Qwen3-8B	FP16 reference	9.715	15.363	77.75	80.93	56.48	74.92	67.64	86.61	74.05
	Shrink-2.4	135.89	85.64	57.02	**32.03**	24.23	33.43	**52.25**	**62.11**	**43.51**
	Proxy-static	**71.14**	**54.13**	**57.70**	31.10	24.70	**36.40**	50.60	56.60	42.80
	Analytic	2.16×10^{11}	3.58×10^{11}	52.12	26.30	**26.79**	26.27	48.46	37.83	36.30
	Min-max	3.93×10^5	6.45×10^4	50.27	24.58	26.28	26.12	49.17	48.93	37.56
(b) HIP preprocessing										
Llama-3.1-8B	FP16 reference	7.213	11.388	80.96	79.59	55.12	79.26	74.03	84.10	75.51
	Shrink-2.4	39.16	53.99	**62.68**	39.86	24.40	**42.90**	54.93	**67.37**	**48.69**
	Proxy-static	**31.38**	**43.81**	62.00	**39.90**	25.10	42.00	**56.50**	65.20	48.50
	Analytic	119.45	75.27	58.43	36.78	23.63	38.12	52.57	58.53	44.68
	Min-max	6.82×10^5	6.6×10^5	52.45	25.25	**26.02**	26.75	48.15	52.11	38.45
Llama-3.2-3B	FP16 reference	11.048	16.486	75.52	67.85	46.16	70.44	67.32	78.62	67.65
	Shrink-2.4	123.56	123.84	56.80	**34.93**	23.98	**34.70**	51.93	60.58	43.82
	Proxy-static	**94.12**	**95.11**	**57.30**	34.30	24.10	34.50	52.20	**62.30**	**44.10**
	Analytic	179.82	196.67	57.24	31.40	23.29	31.42	**52.33**	56.73	42.07
	Min-max	3.43×10^5	2.65×10^5	52.23	26.05	**25.34**	26.24	50.12	46.39	37.73
Qwen3-8B	FP16 reference	9.715	15.363	77.75	80.93	56.48	74.92	67.64	86.61	74.05
	Shrink-2.4	32.46	36.32	68.39	52.82	33.11	48.17	59.83	68.99	55.22
	Proxy-static	**22.65**	**29.75**	**68.80**	**57.60**	**34.60**	**51.20**	**62.90**	**74.40**	**58.20**
	Analytic	46.98	45.13	65.07	50.34	29.44	45.04	57.93	67.13	52.49
	Min-max	5.06×10^4	1.91×10^4	49.46	24.24	26.96	25.55	49.88	43.21	36.55

Whitening. Whitening follows the same template as HIP, a change of basis T that preserves the linear map, but drops the orthogonality constraint and thereby changes the geometry of the reconstruction objective. Let T be an invertible matrix and set

$$\widetilde{W} = WT, \qquad \widetilde{X} = T^{-1}X, \qquad \widetilde{H} = T^{-1}HT^{-\top}.$$

Unlike orthogonal rotations, such transformations alter the Hessian spectrum and hence its effective rank. For positive definite H, choosing T with $TT^\top = H$, e.g. $T = H^{1/2}$, whitens the activations: $\widetilde{H} = I$, $r_{\text{eff}}(\widetilde{H}) = d$, and the reconstruction objective reduces to the plain weight error $\|\widetilde{W} - \widetilde{Q}\|_F^2$, for which coordinatewise rounding is optimal. The cost is that the whitening moves the anisotropy from the Hessian into the weights: the rows of $\widetilde{W} = WH^{1/2}$ have covariance proportional to H rather than being isotropic, and the quantization grid, mapped back to the original basis, is no longer a cube lattice but its image under $H^{-1/2}$. Coordinatewise quantization in the whitened basis therefore optimizes an isotropic objective over a different set of representable weights.

Diagonal rescaling. Between rotation and full whitening sits the per-channel rescaling used by SmoothQuant (Xiao et al., 2023) and AWQ (Lin et al., 2024). Let $T = D$ be diagonal with positive entries, so that

$$\widetilde{W} = WD, \qquad \widetilde{X} = D^{-1}X, \qquad \widetilde{H} = D^{-1}HD^{-1}.$$

The natural choice $D = \mathrm{diag}(H)^{1/2}$ divides each input channel by its root-mean-square activation, turning $\widetilde{H}$ into the correlation matrix of the input channels. The resulting matrix has $\mathrm{tr}(\widetilde{H}) = d$,

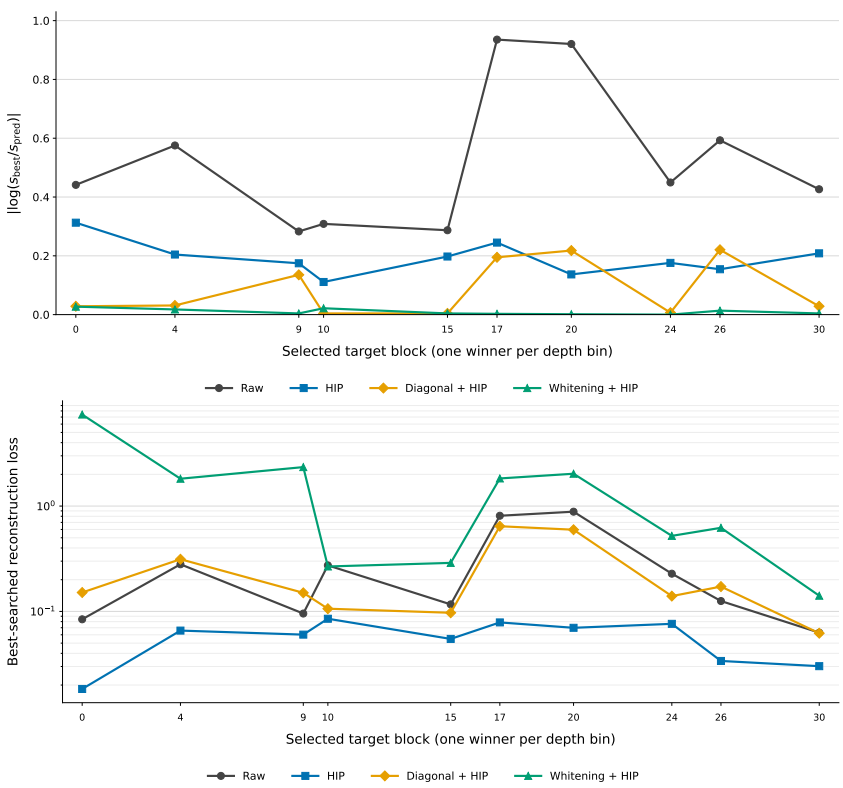


Figure 10: Target-local preprocessing on ten difficult W2 GPTQ rows selected across the 32 blocks of Llama-3.1-8B. Top: absolute log ratio between the best-searched scale and the Gaussian analytic scale. Bottom: best-searched local reconstruction loss on a logarithmic axis. Each point represents one matrix–row pair selected from a prespecified depth bin; horizontal positions indicate the actual block indices. All four conditions share the same unrotated, quantized upstream stream and pre-target data. Transforms are applied only to the selected target, and losses use a common normalization based on the raw row and Gram. Minima are the best evaluated values, not certified global optima.

while the off-diagonal entries are the correlation coefficients $H_{ij}/\sqrt{H_{ii}H_{jj}}$. Consequently

$$r_{\text{eff}}(\widetilde{H}) = \frac{d}{\|\widetilde{H}\|_{\text{op}}},$$

which equals d exactly when the channels are uncorrelated, in which case diagonal rescaling coincides with whitening.

Matched target-local comparison. To distinguish scale agreement from reconstruction quality, we examine ten difficult W2 targets from Llama-3.1-8B, selecting one matrix-row pair from each of ten fixed depth bins. In each of the bins, we select the matrix with the largest ratio of its summed row-wise best-searched GPTQ reconstruction error to signal energy. Within that matrix, we select the output row with the largest normalized best-searched GPTQ loss.

For each target, the raw, HIP, diagonal-scaling+HIP, and whitening+HIP conditions share the same pre-target weights and activation Hessian. Preprocessing is applied only to the selected target, followed by local GPTQ evaluation across candidate scales. Figure 10 and Table 9 show that whitening brings the best-searched scale closer to the Gaussian analytic prediction in these rows, but generally increases the best-searched reconstruction loss relative to HIP alone. Improved scale agreement therefore need not translate into improved reconstruction quality.

B.5 Quantization error landscapes

B.5.1 Gaussian landscapes and convergence

We illustrate the finite-width behavior of the normalized RTN loss using W3 symmetric quantization with levels $\{-3, \dots, 3\}$. For $z \sim \mathcal{N}(0, I_d)$, we consider the identity Gram and random orthogonal projectors of rank k, yielding

$$L_d(s) = \frac{\|z - Q_s(z)\|_2^2}{d}, \qquad L_{d,k}(s) = \frac{\|U^\top (z - Q_s(z))\|_2^2}{k},$$

respectively, where $U^\top U = I_k$ and U is independent of z. To understand if the effective rank condition (H2) of Theorem 1 is necessary we choose $k \in \{d, \lceil\sqrt{d}\rceil, \lceil\log d\rceil\}$, and $constant = 8$.

Figure 11 compares individual landscapes at increasing widths with the Gaussian reference. While $k = d, \lceil\sqrt{d}$ converge at decreasing rates towards the gaussian landscape, $k = \lceil\log d\rceil$ and $k = 8$ keep their unregular form even at high d..

This matches the theoretical prediction, as $k = d, \lceil\sqrt{d}$ satisfy the Theorem 1 asymptotic rank condition, whereas the logarithmic and fixed-rank constructions probe regimes outside that sufficient condition. These curves illustrate individual realizations; the uniform error plot in Figure 12 summarizes variability across 20 repetitions and validates Theorem 1.

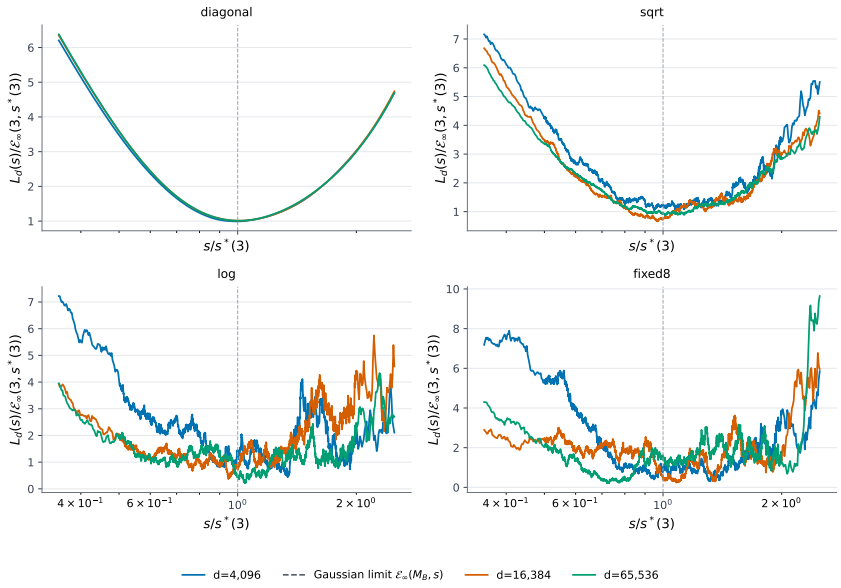


Figure 11: Finite-width W3 RTN landscapes for four effective-rank constructions. Each panel overlays $d \in \{4096, 16384, 65536\}$ and the Gaussian reference (dashed). Curves show one realization per configuration (repetition 0). The horizontal coordinate is $s/s^\star(3)$, and all losses are divided by $\mathcal{E}_\infty(3, s^\star(3))$. Only $0.35 \le s/s^\star(3) \le 2.5$ is displayed.

B.6 Local Reconstruction Sensitivity with HIP

We repeat the local reconstruction sensitivity analysis of Section 5.3 with HIP preprocessing. Figure 13, the HIP counterpart of Figure 3, shows a similar approximately exponential decrease in median sensitivity with increasing bit-width.

To compare sensitivity magnitudes directly, Figure 14 overlays the results with and without HIP at $\delta = 5\%$. For each model and bit-width, we also report the median paired ratio $C_{\text{sym}}^{\text{HIP}}/C_{\text{sym}}^{\text{noHIP}}$, matching calibration seed, matrix, and row ID. Only pairs with positive, finite sensitivity in both conditions are included. Each condition is evaluated around its own best-searched scale.

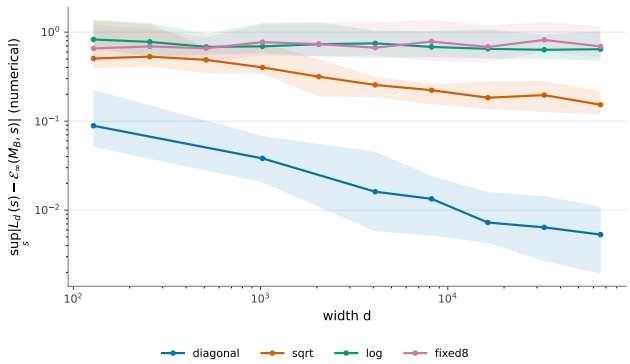


Figure 12: **Uniform-error estimates for W3 Gaussian quantization landscapes.** Median maximum error over the evaluated scale grid $\mathcal{S}$, $\max_{s\in\mathcal{S}} |L_d(s) - \mathcal{E}_\infty(M_3, s)|$, as a function of width d. Shaded regions show the 10th–90th percentiles across 20 repetitions. Curves correspond to effective ranks $k = d$, $\lceil\sqrt{d}\rceil$, $\lceil\log d\rceil$, and 8. Errors decrease in the identity and square-root-rank settings, while the logarithmic- and fixed-rank settings show no comparable decrease over the tested widths..

An illustrative reconstruction landscape. Figure 15 compares the Gaussian reference with empirical RTN and GPTQ landscapes for one fixed row of Llama-3.1-8B at W2, W3, and W8, with and without HIP. The RTN landscapes resemble the Gaussian reference in overall shape, although their minimizing scales can differ. The basin insets illustrate how HIP brings the analytic scale closer to the empirical minimizer in this example.

GPTQ compensation also changes the shape of the local basin. However, this single-row comparison does not establish a general flattening effect. Moreover, the landscapes are displayed after division by the bit-dependent Gaussian minimum, whereas our sensitivity statistic excludes this additional normalization. Cross-precision sensitivity comparisons therefore rely on C_{sym} under the common loss normalization, rather than on the visual sharpness of these curves.

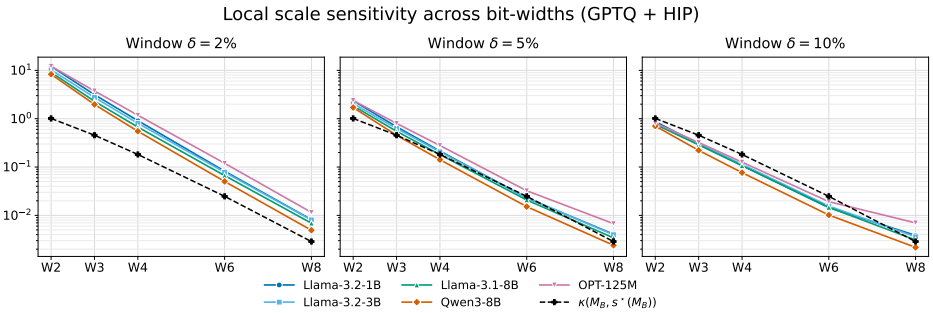


Figure 13: Local GPTQ scale sensitivity after HIP preprocessing. Colored curves show the median symmetric finite-window sensitivity C_{sym} around each row's best-searched scale at the tested precisions W2, W3, W4, W6, and W8. Panels use relative scale perturbations $\delta \in \{2\%, 5\%, 10\%\}$. The black dashed curve shows the Gaussian scale-invariant differential curvature κ, which is identical across panels. Medians pool the selected rows across all studied matrices and three calibration seeds for each model and bit-width.

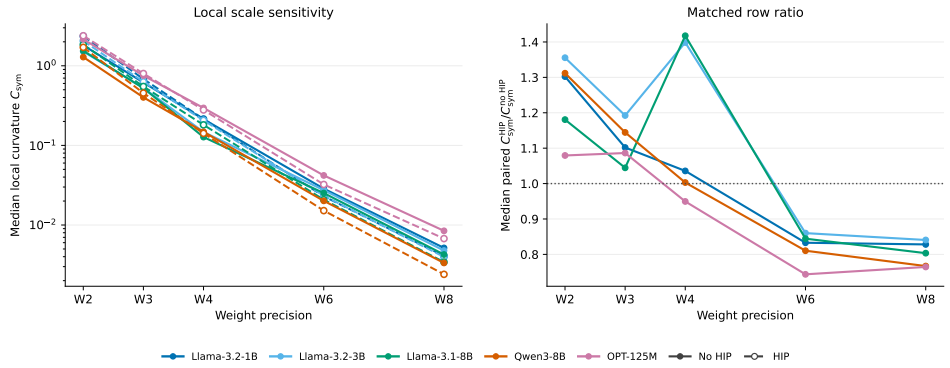


Figure 14: Effect of HIP on local GPTQ scale sensitivity at $\delta = 5\%$. Left: median symmetric finite-window sensitivity C_{sym} by model and bit-width, with filled markers for no HIP and open markers for HIP. Right: median of the paired HIP-to-no-HIP sensitivity ratios, matching model, bit-width, calibration seed, matrix, and row ID. Only pairs with positive, finite sensitivity in both conditions are included. This statistic is the median of individual ratios, rather than the ratio of the pooled medians.

Table 8: W2–ternary per-channel cross-engine comparison with HIP and grid $\{-s, 0, s\}$. Bold values compare selectors within each model and engine, excluding FP16 references. These measurements are separate from the main GPTQ campaign. Lower perplexity and higher accuracy are better.

Engine	Rule	Wiki2 PPL	C4 PPL	PIQA	ARC-E	ARC-C	HellaSwag	WinoGrande	BoolQ	Mean
(a) Llama-3.1-8B										
FP16 reference		7.213	11.388	80.96	79.59	55.12	79.26	74.03	84.10	75.51
GPTQ	Shrink-2.4	60.17	60.98	60.83	37.58	23.63	41.15	52.17	61.25	46.10
	Proxy-static	**33.71**	**45.22**	**63.00**	**40.45**	**25.51**	**43.82**	**55.33**	**66.76**	**49.14**
ResComp	Shrink-2.4	38.59	43.39	**62.79**	36.11	23.72	42.37	55.72	**65.90**	47.77
	Proxy-static	**29.69**	**35.45**	61.59	**36.28**	**23.81**	**44.93**	**55.96**	64.25	**47.80**
QRoNoS	Shrink-2.4	61.07	59.03	**66.54**	**45.71**	**27.22**	45.12	56.67	65.38	**51.11**
	Proxy-static	**38.17**	**43.70**	56.58	32.32	22.35	**46.87**	**57.38**	**67.49**	47.17
(b) Qwen3-8B										
FP16 reference		9.715	15.363	77.75	80.93	56.48	74.92	67.64	86.61	74.05
GPTQ	Shrink-2.4	34.16	38.39	67.63	54.50	31.74	48.65	**59.43**	**70.40**	55.39
	Proxy-static	**24.41**	**32.99**	**68.72**	**55.47**	**32.68**	**51.59**	58.80	67.00	**55.71**
ResComp	Shrink-2.4	105.93	101.35	56.91	32.62	22.70	34.46	**53.43**	**62.91**	43.84
	Proxy-static	**90.31**	**88.28**	**59.09**	**32.66**	**23.63**	**36.52**	52.33	60.98	**44.20**
QRoNoS	Shrink-2.4	76.01	57.94	67.03	48.32	28.24	44.60	58.09	63.36	51.61
	Proxy-static	**34.25**	**32.72**	**68.93**	**52.44**	**30.03**	**50.89**	**61.17**	**69.17**	**55.44**

Preprocessing	Median scale gap $\lvert\log(s_{\mathrm{search}}/s_{\mathrm{an}})\rvert$	Gap lower than HIP	Median best-loss ratio relative to HIP	Loss lower than HIP
None	0.445	0/10	3.46	0/10
HIP	0.187	—	1.00	—
Diagonal + HIP	0.0298	8/10	3.63	0/10
Whitening + HIP	0.00406	10/10	20.8	0/10

Table 9: Target-local W2 GPTQ preprocessing on ten selected difficult rows of Llama-3.1-8B-Instruct. Counts compare each target against HIP alone. Lower is better in both numeric columns.

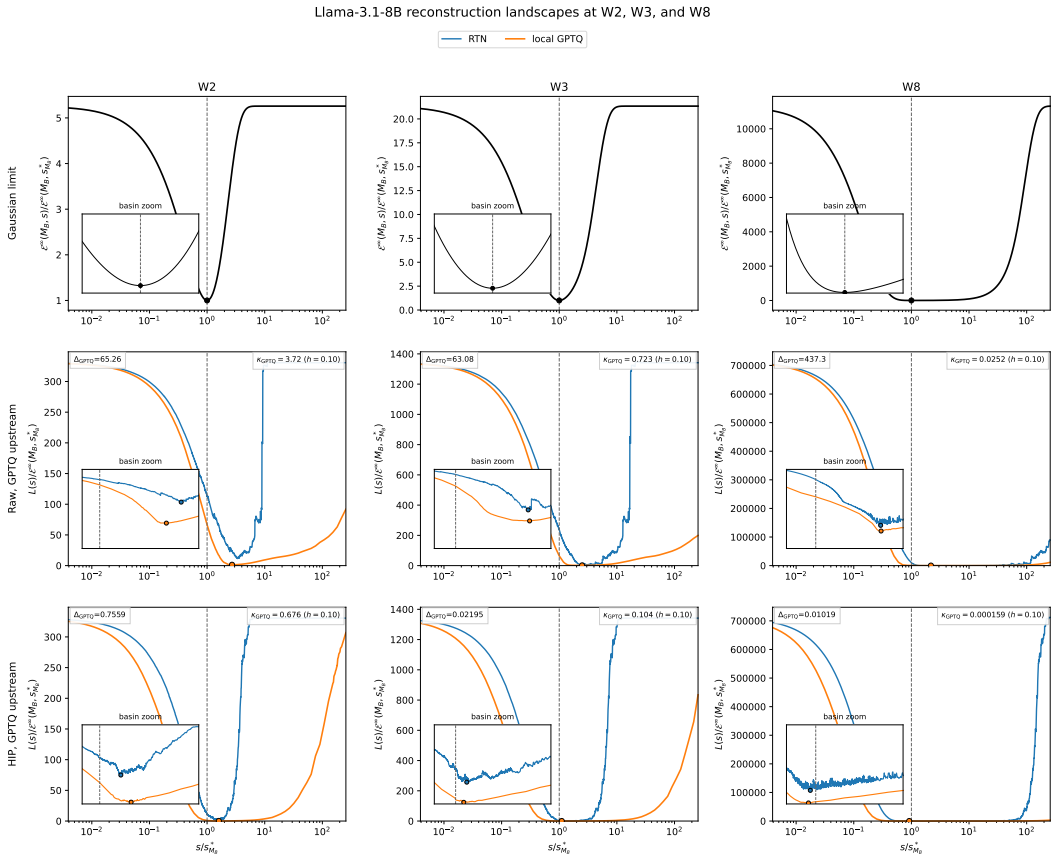


Figure 15: Scale-dependent reconstruction landscapes for Llama-3.1-8B, block 0, `k_proj`, output row 150. Columns show W2, W3, and W8; rows show the Gaussian reference and empirical RTN and GPTQ losses without and with HIP. Scales and losses are normalized by the Gaussian optimal scale and minimum loss, respectively. Dashed vertical lines mark the analytic scale; markers identify best-searched scales. Insets show the local basins. Empirical losses use inputs from the upstream GPTQ-quantized network at the corresponding precision. This figure illustrates one fixed row rather than an aggregate over the representative cohort.