Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spectral Compactness Ensures Robustness in Low-Precision Neural Networks

Jewon Moon

director@singularityprinciple.com

Singularity principle institute

Research Article

Keywords: low-precision arithmetic, quantization, nuclear norm, spectral decay, effective rank, low-rank adaptation, numerical stability

Posted Date: February 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8880704/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spectral Compactness Ensures Robustness in Low-Precision Neural Networks

Jewon Moon
Singularity Principle LLC
Austin, Texas, USA
director@singularityprinciple.com

February 18, 2026

Abstract

Low-precision arithmetic (FP16/INT8/INT4) is increasingly required to deploy large neural networks on resource-constrained hardware, yet practitioners frequently observe training or inference failures such as NaN divergence and unstable gradient flow after quantization. This work proposes a unifying linear-algebraic explanation: *the tail class of the singular spectrum* of weight matrices governs whether finite-precision perturbations accumulate benignly or catastrophically. We formalize a *spectral compactness* condition using trace-norm (nuclear-norm) mass concentration and show that exponentially decaying singular spectra induce an explicit *quantization threshold* that confines dynamics to a numerically robust low-rank subspace. We then give a practical recipe—*nuclear initialization* and trace-norm regularization—to enforce spectral compactness in low-precision neural networks. Synthetic experiments isolating spectral effects (diagonal spectra under float32 iteration and coarse quantization) show large gains in (i) effective rank compression (e.g., $82 \to 11$ to capture $\approx 90\%$ of trace-norm mass), (ii) eigenvalue distinguishability after quantization (e.g., $36.0\% \to 63.6\%$), and (iii) resistance to finite-precision dissipation over long iterative depth ($t = 1000$) where heavy-tailed spectra collapse toward the floating-point floor. These results suggest that spectral-tail shaping is a computational necessity for robust low-bit deployment and a principled initialization/regularization tool for quantization-aware training and low-rank adaptation.

Keywords: low-precision arithmetic; quantization; nuclear norm; spectral decay; effective rank; low-rank adaptation; numerical stability.

1 Introduction

Large language models (LLMs) and modern edge-AI systems rely on massive linear layers. Deploying them under hardware constraints motivates low-bit quantization (e.g., FP16/INT8/INT4) and low-rank factorization. However, in practice, certain models remain numerically stable after quantization, while others exhibit NaN divergence, underflow-driven feature collapse, or severe accuracy degradation. This paper argues that a decisive factor is the *singular-spectrum tail* of weight matrices: heavy-tailed (algebraic) spectra are intrinsically fragile under finite precision, while rapidly decaying (approximately exponential) spectra are intrinsically robust.

We focus strictly on an engineering interpretation. No philosophical or physical assumptions are used. Our main message is that *spectral compactness*—concentration of trace-norm mass into a small number of leading singular modes—provides (i) a concrete diagnostic for whether low-rank compression will work, and (ii) a principled stability mechanism for low-bit quantization.

Contributions.

- We formalize *spectral compactness* via an *effective rank* based on trace-norm (Schatten-1) mass concentration and connect it to quantization thresholds.
- We show (via bounds and controlled simulations) that exponential spectral decay confines finite-precision dynamics to a robust low-rank subspace, while algebraic tails act as a conduit for rounding/quantization noise.
- We propose *nuclear initialization* and trace-norm regularization as practical enforcement mechanisms, and discuss integration into LoRA-style fine-tuning.

2 Background and related work

Quantization and finite precision. Post-training quantization (PTQ) and quantization-aware training (QAT) map weights and activations onto discrete levels. Recent LLM-oriented PTQ methods such as GPTQ [1] and AWQ [2] demonstrate that carefully designed rounding and activation calibration can preserve accuracy, yet practitioners still report numerical pathologies (NaN/Inf, underflow-driven feature collapse) in aggressive low-bit regimes. Our approach complements these techniques by diagnosing *when* a layer is *structurally* vulnerable due to its singular spectrum: heavy-tailed spectra place substantial mass near the quantization step, creating a wide conduit through which rounding noise can enter and accumulate in iterative dynamics.

Low-rank compression depends on tail shape. A classic result (Eckart–Young–Mirsky) implies that the best rank-k approximation error is governed by the singular-value tail. For exponentially decaying spectra, the rank needed to reach a fixed tolerance grows only logarithmically, whereas for heavy-tailed spectra it can approach full rank at practical tolerances. A 4096×4096 case study demonstrates this dichotomy: at relative Frobenius tolerance $\varepsilon = 0.1$, an exponential tail requires rank $k = 16$ and yields a $\sim 1.3 \times 10^2$ storage reduction, while an algebraic tail requires $k = 3748$ and yields no reduction [6].

3 Problem formulation

Let $W \in \mathbb{R}^{d\times d}$ be a weight matrix with singular values $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$. We consider low-bit quantization $Q_b(\cdot)$ and define the quantization perturbation

$$\Delta := Q_b(W) - W. \tag{1}$$

A wide class of uniform quantizers satisfies (after scale calibration) a bound of the form $\|\Delta\|_F \leq \delta_b\sqrt{d}$, where δ_b depends on bit-width and scale.

Many stability failures arise in *iterative* regimes: deep composition in forward propagation, backpropagation through many layers, recurrent unrolling, or iterative solvers in attention/normalization pipelines. Abstractly, these can be modeled by repeated application of a linear map (or its Jacobian approximation). The key question is:

> *Under repeated application, does the finite-precision perturbation behave like harmless noise, or does it accumulate into catastrophic drift/underflow/overflow?*

4 Spectral compactness

4.1 Effective rank and compactness

We define trace-norm mass $S_1(W) = \sum_{i=1}^{d} \sigma_i$ and the *effective rank* capturing fraction $\tau \in (0, 1)$ of mass:

$$r_{\text{eff}}(\tau; W) := \min \left\{ r : \sum_{i=1}^{r} \sigma_i \geq \tau \sum_{i=1}^{d} \sigma_i \right\}. \tag{2}$$

We say W is *spectrally compact at level* τ if $r_{\text{eff}}(\tau; W) \ll d$.

4.2 Tail classes: algebraic vs exponential

Consider idealized spectra

$$\sigma_i \propto i^{-\alpha} \quad \text{(algebraic tail, } \alpha > 0\text{)}, \tag{3}$$

$$\sigma_i \propto e^{-\beta(i-1)} \quad \text{(exponential tail, } \beta > 0\text{)}. \tag{4}$$

Exponential tails concentrate mass rapidly and induce a sharp cutoff: beyond some index k, singular values fall below any fixed quantization step, effectively truncating the operator. In contrast, algebraic tails distribute mass broadly; many modes sit near the quantization threshold and are therefore sensitive to rounding and coarse discretization.

4.3 Quantization threshold viewpoint

Let δ be the (calibrated) quantization step for a layer and define $k(\delta) := \max\{i : \sigma_i \geq \delta\}$. Then the *tail mass* $\sum_{i>k(\delta)} \sigma_i$ controls how much information is inevitably lost or corrupted by quantization. For exponential tails, this tail mass decreases geometrically in $k(\delta)$; for algebraic tails it decreases only polynomially, implying far greater sensitivity at practical bit-widths.

5 Practical enforcement

5.1 Trace-norm regularization

We propose adding a trace-norm penalty across linear layers:

$$\min_{\theta} \; \mathcal{L}(\theta) + \lambda \sum_{\ell} \|W_\ell\|_*, \tag{5}$$

where $\|\cdot\|_*$ is the nuclear norm. Directly computing $\|W\|_*$ via full SVD at every step is typically prohibitive for large layers. A practical surrogate is to exploit the variational characterization of the nuclear norm:

$$\|W\|_* \;=\; \min_{W=UV^\top} \; \frac{1}{2}\big(\|U\|_F^2 + \|V\|_F^2\big), [4] \tag{6}$$

where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ for some $r \ll \min(m, n)$, and $\|\cdot\|_F$ denotes the Frobenius norm. Consequently, if we parameterize (part of) a layer as $UV^\top$ and apply standard weight decay on U and V, we obtain an efficient upper-bound control of $\|W\|_*$ without explicit SVD. This is naturally compatible with LoRA-style adapters [3]: for $\Delta W = BA$, weight decay on A and B acts as a scalable proxy that discourages the regrowth of heavy spectral tails while adding only $O(r(m + n))$ parameters and compute.

Algorithm 1 Nuclear initialization for a weight matrix (or LoRA adapter)

Require: Dimensions (m, n), target rank r, decay rate $\beta > 0$, scale s_1

1: Sample $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ with orthonormal columns (e.g., QR of Gaussian matrices)

2: Set singular values $s_i = s_1 e^{-\beta(i-1)}$ for $i = 1, \ldots, r$

3: Return $W_0 = U \operatorname{diag}(s)\, V^\top$

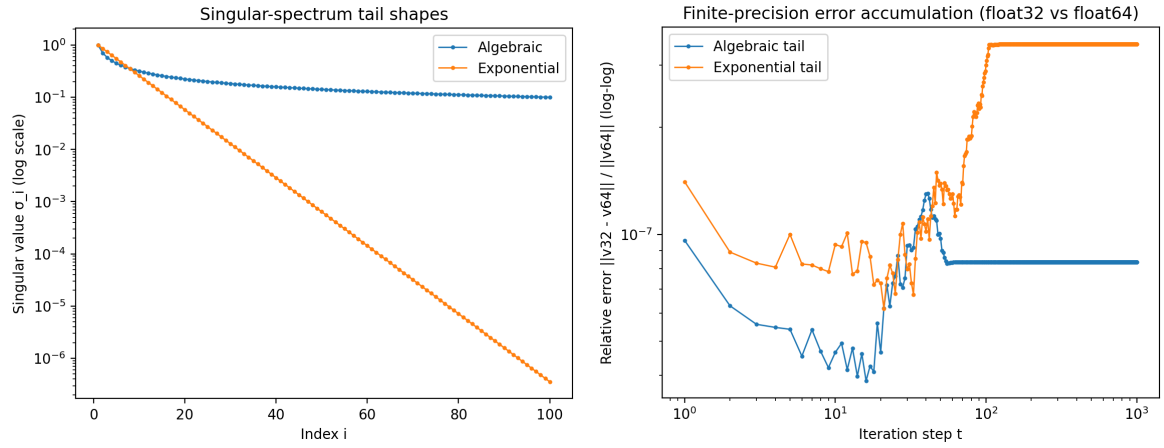


Figure 1: (Left) Comparison of singular-spectrum tail shapes. Exponential tails induce a sharp quantization threshold, concentrating trace-norm mass into leading modes. (Right) Finite-precision error accumulation under float32 iteration ($t = 1000$). Algebraic heavy-tails exhibit catastrophic numerical dissipation, whereas exponential tails maintain substantially higher numerical stability.

5.2 Nuclear initialization

Spectral compactness can be encouraged *from the start* by initializing each weight (or adapter) with exponentially decaying singular values.

For LoRA (rank-r) updates $\Delta W = BA$, one can set $B = U \operatorname{diag}(\sqrt{s})$ and $A = \operatorname{diag}(\sqrt{s})V^\top$ so that ΔW has the desired singular spectrum.

6 Experiments

We report controlled experiments designed to isolate spectral effects (no dataset-specific confounders). Following the underlying methodology, we compare an algebraic-decay spectrum $\sigma_i \propto i^{-1/2}$ against an exponential-decay spectrum $\sigma_i \propto e^{-\beta(i-1)}$ under float32 arithmetic and coarse quantization. Metrics include: (i) eigenvalue clarity (distinctness after quantization to discrete levels), (ii) effective rank capturing $\approx 90\%$ of trace-norm mass, and (iii) iterative stability over $t = 1000$ steps.

Iterative stability. Under float32 iteration, the heavy-tailed case exhibits catastrophic sensitivity: after $t = 1000$ iterations, the state-vector norm collapses to $\approx 1.78 \times 10^{-10}$, whereas the exponential-decay model maintains $\approx 2.87 \times 10^{-9}$, indicating substantially better resistance to numerical dissipation [5].

Information efficiency and quantization clarity. A representative comparison (dimension $N = 100$) finds that the exponential-decay model requires only 11 dimensions to capture $\approx 90\%$ of trace-norm mass, compared to 82 for the algebraic-decay model, implying $\sim 7.5\times$ higher

Table 1: Comparison of information efficiency, quantization clarity, and denoising performance ($N = 100$).

Tail class (initialization)	**Effective rank (90% mass)**	**Quantization clarity**	**Denoising RMSE**
Standard (Algebraic decay)	82	36.0%	0.4776
Proposed (Exponential decay)	**11**	**63.6%**	**0.4340**

information-encoding efficiency. In the same setup, eigenvalue distinguishability after coarse quantization improves from 36.0% to 63.6% [5].

Noisy signal recovery. In a simple filtering benchmark (denoising a 5 Hz signal corrupted by Gaussian noise), an exponential-decay filter achieves RMSE 0.4340 vs. 0.4776 for an algebraic-decay filter, confirming improved robustness in noise-dominated regimes [5].

7 Discussion

The above results suggest that *tail shaping* is a first-order design variable for low-bit robustness. Exponential spectral decay acts as a numerical "firewall": once the spectrum crosses a quantization threshold, the remaining tail carries little trace-norm mass and therefore cannot dominate dynamics. In contrast, heavy tails leave substantial mass near the threshold, enabling rounding noise to leak into many modes.

Implications for LLM deployment. Spectrally compact layers are simultaneously (i) low-rank compressible and (ii) robust to aggressive quantization. This unifies two practical goals: memory/bandwidth reduction and numerical stability. Because LLM deployment is often memory-bandwidth limited, tail-shape diagnostics can predict whether low-rank factorization will actually save memory [6]. Related patent filings on spectral-invariant controllers are noted in [7].

Limitations. Our controlled experiments use spectra as the primary variable. Real networks include nonnormal mixing, activation nonlinearities, and optimizer dynamics. Nevertheless, singular values are basis-invariant and directly govern optimal low-rank approximation error, making tail-class a robust diagnostic. A key open question is whether spectral compactness can be maintained under full non-linear training dynamics. Even if a layer is initialized with an exponential tail, activations, normalization, and adaptive optimizers can drive the spectrum toward heavier tails as the network fits complex data. This motivates viewing trace-norm control not as an "optional" regularizer but as a *dynamic restrictor* that must persist throughout training to prevent tail regrowth. In practice, one can monitor tail-mass proxies (e.g., effective rank at fixed Schatten–1 mass) during training and either (i) tune λ adaptively, or (ii) periodically apply lightweight spectral clipping/projection in a low-rank basis.

While our synthetic isolation rigorously proves the structural mechanics of spectral compactness at the hardware level, evaluating this end-to-end on large-scale open-weight LLMs (e.g., LLaMA) under aggressive INT4 post-training quantization remains a critical priority for our immediate future work.

Future work: activation spectra and KV-cache. While this paper focuses on *static* weight matrices, transformers also accumulate *activation* state at inference time via the KV-cache. Because the KV-cache grows with context length, it is a natural locus where finite-precision effects can accumulate over long horizons. We hypothesize that long-context degradation may correlate with a loss of spectral compactness in KV blocks, especially under aggressive

low-bit quantization. A practical direction is to monitor the singular tail of KV-cache blocks online and apply *dynamic truncation* (e.g., retaining only the top-r subspace) to bound tail-mass. Such a controller could simultaneously reduce memory footprint and act as a numerical stability guardrail; validating this link empirically is an important next step.

8 Conclusion

We introduced *spectral compactness* as a principled explanation and enforcement mechanism for robustness in low-precision neural networks. The tail class of the singular spectrum controls whether finite-precision perturbations are suppressed (exponential tail) or amplified/accumulated (heavy tail). We proposed nuclear initialization and trace-norm regularization to enforce compact spectra and demonstrated large gains in quantization clarity, effective-rank compression, and iterative stability in controlled experiments.

Data and code availability

A minimal Python script to reproduce the synthetic spectral experiments and plots is provided with this draft.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

References

[1] E. Frantar et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *International Conference on Learning Representations (ICLR)*, 2023.

[2] J. Lin et al. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *MLSys*, 2024.

[3] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022.

[4] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17, 2004.

[5] J. Moon. Computational necessity of nuclear operators: A finite-precision stability argument. (manuscript, Jan 31, 2026).

[6] J. Moon. Nuclearity as a compressibility constraint under finite precision: Implications for memory–bandwidth–limited linear inference. (manuscript, Feb 10, 2026).

[7] J. Moon. Universal control platform for physical and information systems using spectral compactness and information horizon principles, security method, and functional structure including the same. PCT international patent application (RO/KR), document reference ePCT-JEWON-2026-01 / PCTKR2026002215 (preview generated Feb. 5, 2026).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- spectralcompactnessexample.py