QALF: DEPLOYMENT ERROR AS A REGULARIZER FOR QUANTIZATION-AWARE LOW-RANK FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank factorization reduces a network's parameter count, while quantization reduces the bits used to store each parameter. Combining them usually means fitting the factors in floating point and rounding them afterwards. Factors that reconstruct a layer accurately can therefore perform poorly at the precision used for storage. We propose QALF, which accounts for quantization during fitting. Under an independent-noise model, we derive each factor's expected contribution to layer-output error in closed form. Adding these contributions to the input-weighted reconstruction loss gives a first-order approximation to the error after quantization. The quantization model determines the regularizer's form, and the target bit-width sets its weight, with no separately tuned trade-off coefficient. On ResNet-18/ImageNet, QALF outperforms dense GPTQ at comparable stored sizes up to the dense 3-bit budget. At 4-bit weights and 8-bit activations, it improves accuracy over the input-weighted reconstruction baseline by 1.8–2.8 percentage points with GPTQ and 4.1–4.7 with round-to-nearest, the rounding rule the objective models. It also exceeds CP-ADMM on its native grid, the closest quantization-aware factorization method, by 0.95–4.43 points at 4 bits. QALF remains more accurate than a tuned component-norm penalty combined with an optimized gauge, despite having higher kernel reconstruction error. The same derivation extends to matrix factorizations: on MLP down-projections in six language models from 410M to 8B parameters, QALF lowers 3-bit perplexity by 1.0–5.4 against SVD-LLM; its factors lose 71–87% less to rounding. For normalized CP factors, we establish conditions under which the regularizer prevents diverging components and the objective attains a minimum.

1 INTRODUCTION

Deploying a network under a memory budget requires choosing between fewer parameters, fewer bits per parameter, or some combination of the two. Low-rank factorization and quantization offer these complementary reductions, but they are usually applied in sequence: the factors are fitted in floating point and rounded afterwards. A factorization that reconstructs a layer accurately can then lose much of that accuracy during quantization. Fitting the factors for the precision at which they will be stored changes that trade-off: on ResNet-18, a factorization built this way is more accurate than dense GPTQ at comparable stored sizes up to the dense 3-bit budget (Figure 1a).

Canonical polyadic (CP) factorization makes the limitation of reconstruction loss clear. The three factors of each rank-one component can be rescaled without changing the reconstructed tensor. These equivalent representations have the same floating-point reconstruction error, but their per-channel ranges and rounding errors can differ. Reconstruction loss alone therefore cannot determine which representation will perform best after quantization.

We propose QALF (**Q**uantization-**A**ware **L**ow-rank **F**actorization), which accounts for quantization while fitting the factors. Under an explicit noise model, we derive the expected output error caused by quantizing each factor separately. We call this contribution its *deployment price*. It depends on the factor's quantization range, the remaining factors, and the input second moment. Adding the three contributions to the input-weighted reconstruction loss gives a first-order approximation to the error after all factors are quantized. The quantization model determines the regularizer's form, and the target bit-width sets its coefficient, with no separately tuned trade-off coefficient.

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

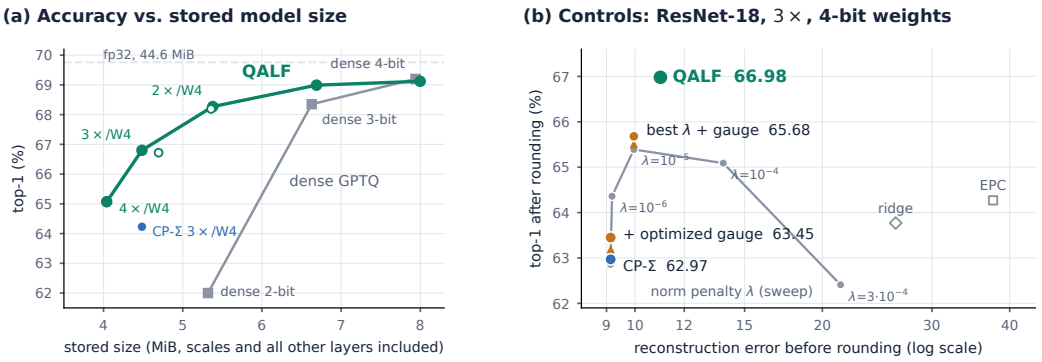

Figure 1: **Factorizing for the deployed precision.** (a) ResNet-18 top-1 against stored size (whole model: packed weights, quantization scales and every other layer). QALF uses 4-, 6- and 8-bit factors at 2×–4×; open markers are two further QALF configurations, 4×/W6 and 3×/W6. Dense GPTQ rounds the same layers at 2, 3 and 4 bits (per-channel affine grid without grouping, as for the QALF factors; seed 0). Factorizing is the better use of memory up to the dense 3-bit size; at 4 bits the two meet. (b) Controls at 3× and 4 bits, with identical rank, calibration data, rounding and BN recalibration. Optimizing the gauge moves CP-Σ vertically (same kernel). A component-norm penalty swept over λ peaks at 65.39%, and adding the optimized gauge gives 65.68%. QALF reconstructs the kernel less accurately than either control and deploys at 66.98%. Symmetric grid, seed 0; three-seed results on both grids are in Table 2.

Existing approaches address different parts of this problem. Data-aware factorizations use the input distribution to weight reconstruction errors (Wang et al., 2025; Kalle et al., 2025). Gauge optimization selects a rescaling of an already constructed approximation (Sao et al., 2026). Generic stabilizers trade component norms against reconstruction error at a level the user sets (Phan et al., 2019), while CP-ADMM constrains the factors to a quantization grid during fitting and measures reconstruction error in weight space (Cherniuk et al., 2024). QALF uses a model of quantization error at the layer output to guide the construction of the approximation. It can improve both the rescaling of a fixed kernel and the kernel itself, accepting a larger floating-point reconstruction error when this improves accuracy after rounding (Figure 1b).

The regularizer also addresses CP degeneracy. A best rank-R approximation need not exist: increasingly accurate approximations can contain large components that cancel one another (de Silva & Lim, 2008). Large factor ranges require coarse quantization steps, making these cancellations sensitive to rounding. For normalized CP factors, we establish conditions under which the regularizer prevents diverging components and the objective attains a minimum, where reconstruction alone may not.

Our contributions are:

- **A closed-form deployment price.** We derive each factor's expected contribution to output error under a quantization-noise model and use these contributions to construct an objective for CP and matrix factorizations. The target bit-width sets the regularizer's coefficient (§4–§5).
- **Improved accuracy at a given stored size.** On ResNet-18/ImageNet, QALF outperforms dense GPTQ at comparable stored sizes up to the dense 3-bit budget. At 4-bit weights and 8-bit activations, it exceeds the input-weighted reconstruction baseline by 4.1–4.7 percentage points under round-to-nearest and 1.8–2.8 under GPTQ. The gains increase with compression and become smaller at 8 bits. A tuned norm penalty combined with an optimized gauge remains below QALF on every seed (§6).
- **A theoretical guarantee and a matrix extension.** We establish conditions under which the normalized finite-bit CP objective attains a minimum (Theorem 2). Applying the same derivation to $W = LR$ lowers 3-bit perplexity by 1.0–5.4 against SVD-LLM on MLP down-projections in six language models from 410M to 8B parameters; its factors lose 71–87% less to rounding.

2 Related work

Data-aware low-rank decomposition. Weighting low-rank approximation by the input distribution goes back to Denton et al. (2014). ASVD and SVD-LLM use activation statistics and whitening to relate truncation to output error in language models (Yuan et al., 2025; Wang et al., 2025), and Kalle et al. (2025) develop alternating least squares for CP and Tucker-2 convolutions under an input-weighted norm. We keep this reconstruction objective and add a factorwise quantization price.

CP degeneracy and stabilization. Fixed-rank CP approximation need not admit a minimizer: large rank-one components can cancel while their sum converges (de Silva & Lim, 2008). EPC minimizes the component norms subject to a reconstruction-error bound (Phan et al., 2019); Phan et al. (2020) minimize sensitivity to isotropic factor perturbations for CNN compression; Zhou et al. (2021) penalize degeneracy in tensor regression. Our price also bounds component norms, with a form and a weight that follow from the deployed quantizer.

Post-training quantization. AdaRound and BRECQ reconstruct layer and block outputs (Nagel et al., 2020; Li et al., 2021); OBC and GPTQ use second-order information to compensate rounding errors (Frantar & Alistarh, 2022; Frantar et al., 2023); SmoothQuant, AWQ and OmniQuant rescale or reparameterize weights and activations before rounding (Xiao et al., 2023; Lin et al., 2024; Shao et al., 2024). QALF produces the factors, which are then quantized downstream; we evaluate RTN and GPTQ.

Low-rank factorization with quantization. The closest CP method, CP-ADMM (Cherniuk et al., 2024), constrains the factors to a quantization grid during alternating optimization; QALF keeps the factors continuous, optimizes them against a model of their rounding error, and leaves the rounding itself to a downstream quantizer. LPLR quantizes randomized low-rank matrix factors with error guarantees (Saha et al., 2023); CALDERA, LQER and QERA combine a quantized dense matrix with low-rank terms (Saha et al., 2024; Zhang et al., 2024; 2025); we represent the weight by its factors alone and optimize them against their rounding. For gauges, ASVD reports that splitting singular values between the factors improves quantized accuracy (Yuan et al., 2025, App. A.8), and, concurrently, Sao et al. (2026) derive the expected product error under uniform noise and optimize a diagonal gauge of fixed factors. Applied to CP, their diagonal-gauge problem corresponds to the restriction of ours to a single equivalence class.

3 Problem Setup and Deployment Contract

QALF optimizes continuous CP factors, while deployment uses their quantized versions. This section defines the two error measures that follow, reconstruction in floating point and output error after the factors are rounded, together with the scaling freedom that separates them.

3.1 Functional low-rank factorization

Let $\mathcal{W} \in \mathbb{R}^{T\times S\times D^2}$ denote a convolution kernel with T output channels, S input channels, and merged spatial dimension D^2. A rank-R canonical polyadic (CP) approximation is

$$\hat{\mathcal{W}} = [\![A, B, C]\!] := \sum_{r=1}^{R} a_r \circ b_r \circ c_r, \qquad A \in \mathbb{R}^{T\times R}, \quad B \in \mathbb{R}^{S\times R}, \quad C \in \mathbb{R}^{D^2\times R}. \tag{1}$$

The three factors B, C, A can be mapped in this order to a 1×1 input projection, a depthwise spatial convolution, and a 1×1 output projection (Lebedev et al., 2015).

We introduce the Σ-norm loss in equation 3. Let $X_u \in \mathbb{R}^{SD^2\times P}$ be the input feature map unfolded into P patches, we define the second moment Σ of the input

$$\Sigma := \mathbb{E}_X\left[X_u X_u^\top\right] \succeq 0 \tag{2}$$

summed over the patches of an image and averaged over calibration images. The expected squared difference between the dense convolution $f(\mathcal{W}; x)$ and the factorized layer $\hat{f}(\hat{\mathcal{W}}; x)$ is

$$\mathcal{L}_\Sigma(A, B, C) := \mathbb{E}_x\left[\left\|f(\mathcal{W}; x) - \hat{f}(\hat{\mathcal{W}}; x)\right\|_2^2\right] = \left\|\left(\mathcal{W} - \hat{\mathcal{W}}\right)_{(1)} \Sigma^{1/2}\right\|_F^2, \tag{3}$$

where $\mathcal{W}_{(1)} \in \mathbb{R}^{T\times SD^2}$ is the mode-1 unfolding (matricization) of the kernel with respect to the first mode. The unweighted Frobenius objective is recovered at $\Sigma = I$. Minimizing $\mathcal{L}_\Sigma$ without any deployment price defines the functional CP-Σ baseline used in §6.

3.2 The deployed target

For bit-width b, let Q_b^A, Q_b^B, and Q_b^C denote the quantizers applied to the three factors. Under the perturbation model introduced in §4.1, we are interested in the expectation of the functional error after all factors have been quantized.

$$\mathcal{R}_b(A,B,C) := \mathbb{E}_Q\left[\mathcal{L}_\Sigma\left(Q_b^A(A), Q_b^B(B), Q_b^C(C)\right)\right] \tag{4}$$

The deployment contract fixes the rank, bit-width, quantization granularity, clipping rule, and calibration distribution. In the convolutional implementation considered here, A is quantized per row, whereas B and C are quantized per column. These axes coincide with the output channels of the three factorized sub-layers.

3.3 Gauge freedom in CP Factorization

A CP factorization is determined only up to a rescaling of its components. For positive vectors $\alpha, \beta \in \mathbb{R}^R_{>0}$, all the following factorizations reconstruct the same tensor.

$$[\![A,B,C]\!] = \left[\!\left[AD_\gamma^{-1}, BD_\alpha, CD_\beta\right]\!\right] \qquad \gamma_r = \alpha_r\beta_r \,\forall r \in [\![1,R]\!] \tag{5}$$

Consequently, $\mathcal{L}_\Sigma$ is constant along the orbit generated by equation 5, but the per-channel ranges, clipping thresholds, and rounding errors are not. The hardware stores and rounds one representative of the orbit, so two factorizations of the same tensor can deploy with different errors. §4 derives a tractable approximation to equation 4; restricted to the orbit it selects a representative, and minimized freely it moves the approximation.

4 Deployment Error, QALF Objective, and Geometry

This section approximates the deployed risk equation 4. Each factor's perturbation is priced on its own, and the three prices are then combined with the floating-point reconstruction loss. On a canonical parameterization the same price removes the escape directions responsible for CP degeneracy.

4.1 Channelwise perturbation model

For each stored factor, we define its quantization error as

$$\delta A := Q_b^A(A) - A, \qquad \delta B := Q_b^B(B) - B, \qquad \delta C := Q_b^C(C) - C. \tag{6}$$

Consider a vector $x \in \mathbb{R}^n$ stored as one quantization channel. For the symmetric b-bit reference quantizer, let

$$q_b := 2^{b-1} - 1, \qquad \mu(b) := \frac{1}{12q_b^2}. \tag{7}$$

We denote by $w_{b,n}(x) \geq 0$ the *effective range* of the channel n: it is normalized so that $\mu(b)w_{b,n}(x)^2$ is the modeled mean-squared perturbation of each entry. §5.1 instantiates a variant of the absolute max clipping range so that the same quantity also accounts for overload error.

Assumption 1 (Independent channelwise perturbations)**.** *For every channel $x \in \mathbb{R}^n$, the perturbation $\delta x := Q_b(x) - x$ satisfies*

$$\delta x \sim \mathbb{U}\left[-\frac{w_{b,n}(x)}{2q_b}, \frac{w_{b,n}(x)}{2q_b}\right]; \qquad \mathbb{E}[\delta x] = 0, \qquad \mathbb{E}[\delta x\, \delta x^\top] = \mu(b)w_{b,n}(x)^2 I_n \tag{8}$$

The perturbation entries are independent, and perturbations belonging to different channels or factors are mutually independent.

The assumption serves as an analytical reference for round-to-nearest quantization. Clipping introduces biased overload errors, and error-compensating rounders correlate perturbations. The error calculation below needs no condition on $w_{b,n}$; those required for well-posedness are stated in §4.4.

4.2 SINGLE-FACTOR CLOSED-FORM ERROR

Consider the error produced when only the output factor A is quantized. Multilinearity gives the exact output-factor perturbation

$$[\![\delta A, B, C]\!]_{(1)} = \delta A Z^\top \tag{9}$$

where Z is the columnwise Khatri–Rao product of the other factors

$$Z := C \odot B = [c_1 \otimes b_1 \quad \cdots \quad c_R \otimes b_R] \in \mathbb{R}^{SD^2 \times R}, \qquad b_r := B_{:,r}, \quad c_r := C_{:,r} \tag{10}$$

Theorem 1 (Output-factor deployment price)**.** *Under Assumption 1, per-row quantization of A satisfies*

$$\begin{aligned} \mathcal{E}_Q(A) &:= \mathbb{E}\left[\left\|\delta A Z^\top \Sigma^{1/2}\right\|_F^2\right] & M_A &:= Z^\top \Sigma Z \succeq 0 \\ &= \mu(b)\, \underbrace{\rho_w(A)^2 \operatorname{tr}(M_A)}_{\Phi_A} & \rho_w(A)^2 &:= \sum_{t=1}^{T} w_{b,R}(A_{t,:})^2 \end{aligned} \tag{11}$$

The first factor in Φ_A describes the row quantization of A; the second measures how the remaining factors propagate their perturbations under the input metric.

For the interior factors, each column is one quantization channel. $s \in [\![1, S]\!]$ indicate a channel index and $d \in [\![1, D^2]\!]$ a spatial index. Let $\Sigma^{(s,s)} \in \mathbb{R}^{D^2 \times D^2}$ be the principal block obtained by fixing the input channel to s, and let $\Sigma^{[d,d]} \in \mathbb{R}^{S \times S}$ be the principal block obtained by fixing the spatial position to d. We define the quantities analogous to Φ_A.

$$\kappa_r^B := \sum_{s=1}^{S} c_r^\top \Sigma^{(s,s)} c_r, \qquad \Phi_B := \sum_{r=1}^{R} \|a_r\|_2^2 w_{b,S}(b_r)^2 \kappa_r^B, \tag{12}$$

$$\kappa_r^C := \sum_{d=1}^{D^2} b_r^\top \Sigma^{[d,d]} b_r, \qquad \Phi_C := \sum_{r=1}^{R} \|a_r\|_2^2 w_{b,D^2}(c_r)^2 \kappa_r^C. \tag{13}$$

Then the corresponding single-factor errors are

$$\mathcal{E}_Q(B) = \mu(b)\Phi_B, \qquad \mathcal{E}_Q(C) = \mu(b)\Phi_C \tag{14}$$

Independence removes covariance blocks that are off-diagonal in the perturbed mode. Consequently, Φ_B and Φ_C use only the indicated principal blocks of Σ, whereas Φ_A retains the full input second moment through $M_A = Z^\top \Sigma Z$.

4.3 THE QALF OBJECTIVE

Expanding the three-factor quantized product yields the floating-point reconstruction residual, three terms containing one factor perturbation, and terms containing two or three perturbations. Under Assumption 1, the cross terms between factor perturbations vanish in expectation, as do their cross terms with the reconstruction residual. Hence

$$\mathcal{R}_b(A, B, C) = \mathcal{L}_\Sigma(A, B, C) + \mu(b)\big(\Phi_A + \Phi_B + \Phi_C\big) + \mathcal{R}_{\geq 2}, \tag{15}$$

where $\mathcal{R}_{\geq 2}$ collects interactions containing at least two factor perturbations and is of second order in the noise variances under the model. The Quantization-Aware Low-Rank Factorization objective is

$$\boxed{\mathcal{J}_b(A, B, C) := \mathcal{L}_\Sigma(A, B, C) + \mu(b)\big(\Phi_A + \Phi_B + \Phi_C\big)} \tag{16}$$

Each single-factor price is exact under Assumption 1; their sum with the reconstruction loss is a first-order approximation to the expected error once all three factors are quantized. The trade-off between reconstruction and quantization is fixed by the target bit-width and the quantizer model, so no external coefficient is introduced. As precision increases, $\mu(b) \to 0$ and equation 16 approaches the functional CP-Σ objective. All the proofs are detailed in Appendix A.

4.4 SELECTION ON THE SCALING ORBIT AND WELL-POSEDNESS

The analytical quantization error controls two freedoms that reconstruction leaves open. It selects which representative of a fixed CP product is stored, and it prevents the factor norms from escaping along the degenerate sequences of classical CP approximation.

Canonical slice. The full parameter space contains the noncompact internal scaling $b_r \mapsto \eta_r b_r$, $c_r \mapsto \eta_r^{-1} c_r$, along which the product and the three prices remain unchanged. We state the result on the normalized domain

$$\mathcal{S} := \mathbb{R}^{T\times R} \times \left(\mathbb{S}^{S-1}\right)^R \times \left(\mathbb{S}^{D^2-1}\right)^R, \tag{17}$$

so that $\|b_r\|_2 = \|c_r\|_2 = 1$ and component scale is carried by a_r.

Theorem 2 (Well-posed finite-bit objective)**.** *Assume $\Sigma \succeq \varepsilon I$ for some $\varepsilon > 0$ and that the scale functional is continuous, degree-one homogeneous, and positive away from the origin. For every integer $b \geq 2$, $\mathcal{J}_b$ is coercive in A on $\mathcal{S}$ and attains a minimum. Along a canonical CP-degenerate sequence with $\max_r \|a_r\|_2 \to \infty$, the objective diverges (proof in Appendix A.6).*

To see the coercive mechanism, let $m_r = (c_r \otimes b_r)^\top \Sigma (c_r \otimes b_r)$. On the canonical slice, $m_r \geq \varepsilon$ and thus $\mathrm{tr}(M_A) = \sum_r m_r \geq R\varepsilon$. Positive homogeneity and compactness of the unit sphere give a constant $c_w > 0$ such that $w_{b,R}(a) \geq c_w \|a\|_2$. Therefore

$$\mathcal{J}_b(A, B, C) \geq \mu(b)\, c_w^2 R\varepsilon\, \|A\|_F^2, \tag{18}$$

which diverges with $\|A\|_F$. Compactness of the normalized factors and continuity then yield attainment. The unregularized problem can fail to attain its infimum, for some formats and ranks on a set of tensors of positive measure (de Silva & Lim, 2008). On the unquotiented parameter space the internal B/C scaling remains a flat direction, and the theorem claims no coercivity there.

Selection along the orbit. Under the transformation equation 5, Φ_B and Φ_C are invariant because their scale factors cancel componentwise; only Φ_A varies. Restricted to one orbit, the QALF objective therefore selects the stored representative of a fixed kernel by minimizing Φ_A over the diagonal gauge γ. For the absmax scale this is a convex second-order-cone program (Appendix A.5); for the clipped scale of §5.1 we optimize it in log-coordinates. This restricted problem corresponds to the concurrent gauge selection of Sao et al. (2026) applied to CP; minimizing the full objective also moves the kernel, and §6.2 separates the two effects.

4.5 TWO-FACTOR MAPS

The same bookkeeping applies to a matrix factorization $W \approx LR$ with both factors quantized per row; Appendix A.3 treats general chains $W = F_m \cdots F_1$. With ℓ_k the k-th column of L,

$$\mathbb{E}[\mathcal{E}_L^2] = \mu(b)\rho_w(L)^2\, \mathrm{tr}(R\Sigma R^\top), \qquad \mathbb{E}[\mathcal{E}_R^2] = \mu(b)\, \mathrm{tr}(\Sigma) \sum_k \|\ell_k\|_2^2 w_{b,n_i}(R_{k,:})^2, \tag{19}$$

the $m = 2$ case of the CP prices (Appendix C). We use it for language models in §6.2.

5 PRACTICAL QALF

5.1 CLIPPED-ℓ_2 EFFECTIVE RANGE

We instantiate the effective range introduced in §4.1 by matching its modeled variance to a clipped-quantizer distortion:

$$\mu(b) w_{b,n}(a)^2 := \min_{c\geq 0} \Big[\underbrace{\frac{c^2}{12 q_b^2}}_{\text{granular noise}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} (|a_i| - c)_+^2}_{\text{clipping error}} \Big]. \tag{20}$$

The first term is the uniform-noise variance associated with step size c/q_b, while the second is the mean squared error caused by saturation outside $[-c, c]$. Minimizing over c therefore selects the clipping threshold that balances quantization resolution against overload error. Overload errors are biased, so this term does not satisfy the zero-mean part of Assumption 1. We use it as an effective variance and check its calibration against realized rounding in §6.2.

5.2 SOLVER

We optimize QALF by alternating updates to the three factors. QALF and CP-Σ use the same initialization within each comparison: a seeded Gaussian draw for Table 1, and a data-free ridge-regularized CP-ALS solution in the ablation setting; the QALF entries marked † start from CP-Σ instead. Each sweep updates B, then C, normalizes their columns to unit norm while transferring their scales into A, and finally updates A. Each block update minimizes the full objective equation 16 over that factor by L-BFGS with a strong-Wolfe line search, at most 15 iterations; the clipping threshold is recomputed exactly at every evaluation by sorting, and the envelope gradient of Appendix B.1 avoids differentiating through it. Sweeps stop when the relative reconstruction error changes by less than 10^{-8}, or after $\leq$1000 sweeps. Normalization returns the factors to $\mathcal{S}$ and preserves the kernel, Φ_B and Φ_C, but can change Φ_A (Appendix B.3), so the procedure approximates block descent on $\mathcal{S}$ under a finite budget, and we make no convergence claim. CP-Σ uses the same normalization but solves each block exactly, by least squares for A and MINRES for B and C. The coefficient $\mu(b)$ is fixed by the target bit-width. Our experiments use the RTN coupling $\mathrm{tr}(M_A)$, which does not model GPTQ's error compensation; Appendix B.2 derives a GPTQ-aware alternative.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTING

Models and factorization. Our primary evaluation uses ResNet-18 on the ImageNet-1K validation set (50k images). Each full convolution except the first is replaced by a rank-R CP chain; 1×1 downsample convolutions are not factorized. R is set per layer to reach a compression factor of the factorized weights ($2\times$ halves their size). We use 2048 training images to estimate Σ and for BN recalibration after quantization.

Methods and deployment. QALF optimizes equation 16 with the per-channel symmetric clipped-ℓ_2 range of §5.1. CP-Σ minimizes the same input-weighted reconstruction equation 3 without the price. CP-ADMM (Cherniuk et al., 2024) constrains the factors to its native per-tensor symmetric grid during factorization, starting from EPC-stabilized CP; CP-EPC is that starting point rounded like the other methods. QALF, CP-EPC and CP-Σ factors are rounded by RTN or GPTQ (Frantar et al., 2023) on the per-channel affine grid; QALF's RTN uses the harder symmetric grid its price models. All quantizers choose their clipping threshold by an MSE search over $[0.2, 1.2]\times$ the channel maximum (200 candidates), as in CP-ADMM. WbA8 denotes b-bit weights in the factorized layers and 8-bit activations, calibrated on 500 training images; the first, last and downsample layers are kept at 8 bits. After rounding, BN statistics are recalibrated and folded into the dequantization scales and biases. The ablation setting (Table 2, the dense comparison, and the GoogLeNet and ResNet-50 comparisons) quantizes the weights only; there A is rounded by GPTQ and B, C by per-channel OBQ, identically for every method.

6.2 RESULTS

Does QALF improve the compression trade-off? In the ablation setting QALF at $2\times$ with 4-, 6- and 8-bit factors reaches 68.3%, 69.0% and 69.1%, against 62.0%, 68.4% and 69.2% for dense GPTQ at 2, 3 and 4 bits, at whole-model sizes within 2% of each other (Figure 1a). QALF is more accurate at the dense 2- and 3-bit budgets and comparable at 4. In the main WbA8 setting dense GPTQ reaches 57.10% and 67.74% at W2A8 and W3A8 against 68.27% for QALF at $2\times$/W4A8 (Table 1), which is both more accurate than the dense 3-bit model and smaller.

External baselines. Against CP-EPC, the stabilized CP that CP-ADMM starts from, rounded the same way, QALF gains $+1.14$, $+2.09$ and $+3.81$ points at 4 bits. Against CP-ADMM on its native grid it gains $+0.95$, $+2.33$ and $+4.43$ at 4 bits, and $+0.14$, $+0.87$ and $+2.12$ at 8 bits; against the stronger of its two grids at each ratio, $+0.95$, $+1.65$ and $+3.56$ (Appendix D).

Is the price the reason? CP-Σ and QALF share the initialization, the input weighting and the deployment; their objectives differ only by $\mu(b)(\Phi_A+\Phi_B+\Phi_C)$, although their solvers differ (§5.2). Under RTN, the rounding rule the price models, QALF gains $+4.1$, $+4.5$ and $+4.7$ points at $2\times$, $3\times$

Table 1: ResNet-18 top-1 accuracy (%) on ImageNet-1K after weight quantization of the factorized layers, with 8-bit activations. Columns give the compression factor of the factorized layers. CP-ADMM constrains its factors to its native per-tensor symmetric grid (a per-channel affine variant is in Appendix D); every other method is rounded on the per-channel affine grid, except QALF's RTN (symmetric). CP-EPC is CP stabilized by EPC (Phan et al., 2019). Mean ± sd over 3 seeds; a seed fixes the calibration subsets used for Σ, activation calibration, GPTQ and BN recalibration, and is shared across methods. Uncompressed FP32 ResNet-18 reaches 69.76 %. Full-precision accuracies at 2× are in Appendix D. Best mean per column in bold. † Initialized from CP-Σ.

		W4A8			W8A8		
Factorization	Rounding	2×	3×	4×	2×	3×	4×
CP-ADMM	native grid	67.32 ±0.09	64.37 ±0.20	60.18 ±0.54	68.54 ±0.08	66.64 ±0.19	63.77 ±0.19
CP-EPC	RTN	65.37 ±0.17	61.94 ±0.29	57.80 ±0.32	68.42 ±0.04	66.47 ±0.14	63.43 ±0.19
	GPTQ	67.13 ±0.10	64.61 ±0.06	60.80 ±0.14	68.40 ±0.04	66.52 ±0.13	63.44 ±0.17
CP-Σ	RTN	63.87 ±0.12	61.78 ±0.49	59.33 ±0.45	68.27 ±0.02	66.61 ±0.09	64.42 ±0.08
	GPTQ	66.45 ±0.13	64.56 ±0.14	61.83 ±0.13	68.31 ±0.03	66.62 ±0.12	64.48 ±0.02
QALF	RTN	67.95 ±0.09	66.31 ±0.08	64.04 ±0.18	68.66† ±0.05	67.49† ±0.04	65.84† ±0.09
	GPTQ	**68.27** ±0.04	**66.70** ±0.07	**64.61** ±0.06	**68.68**† ±0.05	**67.51**† ±0.08	**65.89**† ±0.06

Table 2: Controls at 3×, 4-bit weights, identical deployment (ablation setting, weights only; mean ± sd over 3 seeds). The penalty coefficient is an oracle choice: the best of the sweep on the evaluation set, in the control's favour.

grid	CP-Σ	+ gauge	+ tuned penalty	+ penalty + gauge	QALF
symmetric	62.95 ±0.18	63.41 ±0.22	65.65 ±0.23	65.94 ±0.23	**66.98** ±0.07
affine	64.26 ±0.19	64.62 ±0.10	66.32 ±0.29	66.43 ±0.14	**66.93** ±0.12

and 4× (W4A8, Table 1); under GPTQ, which compensates part of the rounding error, it still gains +1.82, +2.14 and +2.78; both grow with compression. The input weighting alone does not account for this: QALF with $\Sigma = I$ beats data-aware CP-Σ with GPTQ by 1.4 and 0.8 points at 2× and 2.5× (Appendix D). An L-BFGS solver on the unpriced objective does not recover the gain and is numerically unstable: CP-Σ reaches 61.7% in full precision at 2× and collapses after 4-bit rounding (0.5% with GPTQ). With exact block solves, more sweeps make it worse after rounding: from 10 to 100 sweeps, full precision rises from 68.59 to 69.03 while W4A8 accuracy with RTN falls from 63.81 to 48.20 (Table 10). We report the oracle-best 10 sweeps, chosen on the evaluation set in the baseline's favour; the behaviour is consistent with the component growth of §4.4. The initialization changes little either: from CP-ADMM's factors instead of the ridge solve, QALF reaches 67.09% against 66.98% at 3× and 4 bits (ablation setting, seed 0). At higher precision the price shrinks ($\mu(8) = \mu(4)/330$) and so does the advantage: +0.37, +0.89 and +1.41 at W8A8, and a tie at 2× (69.10 each, ablation setting) when both are built for 8 bits.

Gauge and stabilization. We test rescaling and generic stabilization at 3× and 4 bits (ablation setting, three seeds, identical deployment). Optimizing the gauge of CP-Σ leaves the kernel unchanged and gains 0.4 points. The stabilization control adds $\lambda \operatorname{tr}(M_A)\|A\|_F^2$ to CP-Σ: it keeps the data-dependent coupling of Φ_A but replaces the quantizer's range with a Frobenius norm and $\mu(b)$ with a tunable weight, swept over $[10^{-8}, 10^{-3}]$ and selected on the evaluation set in the control's favour. Adding the optimized gauge brings it closer, and QALF is still ahead on every seed (Table 2), by 1.04 points on the symmetric grid that the price models and by 0.51 on the affine grid of Table 1, whose zero-point absorbs part of the range imbalance the price corrects. EPC and ridge-regularized CP stay further behind (Figure 1b). QALF reconstructs the kernel worse than both the gauge-optimized and the tuned-penalty control. The best λ moves with precision, 10^{-5} at 4 bits and 10^{-7} at 8 bits (Appendix D), where $\mu(b)$ follows the bit-width. Running the gauge optimizer on QALF factors lowers Φ_A by 13% (31% on CP-Σ) and leaves their accuracy unchanged, against the 0.4 points it adds to CP-Σ.

The strength of the regularizer. The factors produced by QALF depend on the target bit-width b. Deployed at 4, 6 and 8 bits, the 4-bit build reaches 68.25, 68.64 and 68.67 after GPTQ. The

Table 3: Language models, 3-bit weights; mean ± sd over 3 seeds. WikiText-2 perplexity before / after rounding both factors. The MLP down-projection of every block is factorized at 80% of its parameters and rounded by GPTQ (act-order, group 128, 128 calibration sequences). Damage removed $= 1 - \Delta_{\mathrm{QALF}}/\Delta_{\text{SVD-LLM}}$, with Δ the perplexity added by rounding, per seed and averaged; on the two Llama models the baseline's GPTQ is unstable across seeds (Appendix D).

	SVD-LLM		QALF		damage
Model	fp	W3	fp	W3	removed
Pythia-410M	20.57 ±0.03	25.99 ±0.11	21.24 ±0.02	**22.50** ±0.03	76.8%
Qwen2.5-0.5B	15.89 ±0.02	20.09 ±0.10	16.97 ±0.09	**18.16** ±0.09	71.6%
Pythia-1B	14.70 ±0.01	16.52 ±0.03	15.00 ±0.00	**15.53** ±0.01	71.0%
Qwen2.5-1.5B	11.07 ±0.02	13.64 ±0.17	11.75 ±0.01	**12.34** ±0.01	76.8%
Llama-3.2-3B	9.68 ±0.02	16.56 ±4.43	10.00 ±0.02	**11.15** ±0.18	79.3%
Llama-3.1-8B	8.00 ±0.01	13.90 ±3.89	8.35 ±0.01	**8.97** ±0.06	86.5%

converse direction is costly: a 6-bit build deployed at 4 bits loses 1.0 point to the matched 4-bit build at 2× (67.24 vs. 68.25 after GPTQ), and 1.4 and 1.5 points at 3× and 4× in the ablation setting. Scaling the price by $\eta = 0.5$ or 2 costs 0.19 and 0.25 points (3×, 4 bits, seed 0): the derived strength is the best of the three.

Does the error model explain the gain, and does it carry over? Figure 3 (Appendix D) tracks the objective, its clipped surrogate and the realized RTN loss along one layer's optimization; the surrogate stays above the realized loss throughout. Rounding each factor alone at the final iterate, the realized-to-predicted ratio over the 16 factorized layers has median 0.86 for Φ_A, 0.83 for Φ_B and 0.57 for Φ_C; in the ablation setting, predicted and realized damage have rank correlation 0.99. Φ_C overpredicts because a spatial channel has $D^2 = 9$ entries, few enough that the deployed MSE clipping search can fit a step to them. On the 3× QALF factors the interaction terms dropped from equation 16 amount to 0.38% of the price at 4 bits, and at most 0.45% on any single layer (Appendix A.4).

Other architectures. On GoogLeNet at 3× W4 (ablation setting, symmetric per-channel grid), QALF reaches 67.49% against 66.02% for CP-Σ (+1.47) and 66.42% for CP-ADMM (+1.07). On ResNet-50 at 3× W4, factorizing its sixteen 3×3 convolutions, QALF reaches 75.34% against 75.00% for CP-Σ.

Language models. Table 3 applies the two-factor price of §4.5 to the MLP down-projections of six language models from three families, in its groupwise form at 3-bit weights (Appendix C.2). QALF refines the SVD-LLM factors at the same rank, so the baseline is its own starting point: its full-precision perplexity is 0.3 to 1.1 higher, and its perplexity after rounding is 1.0 to 5.4 lower; its factors lose 71–87% less to rounding. These input covariances are close to rank one, so the fit enforces $\Sigma \succeq \varepsilon I$ (Theorem 2) with an untuned $\varepsilon = \lambda_1(\Sigma)/n$ applied to the fit alone; the rounder uses the true Σ. On the two models tested without it QALF still removes 71–75% (Appendix D).

7 Limitations and Conclusion

Limitations. The price assumes independent, zero-mean channelwise noise, and the clipped-ℓ_2 range treats overload error as variance, which overprices the nine-entry spatial factor. It prices round-to-nearest, while GPTQ correlates perturbations; a matched coupling is calibrated for A alone (Appendix B.2). It models the symmetric grid, whose range is positive on nonzero channels as Theorem 2 requires; on the affine grid the margin over the best tuned control halves (Table 2). The evaluation covers three CNNs, and one projection per language-model block, which tests the transfer of the derivation rather than end-to-end compression.

Conclusion. QALF fits a low-rank factorization for the precision at which it will be stored. Its regularizer follows from the deployed quantizer and the bit-width, and moves both the stored rescaling and the approximation itself: the factors it produces reconstruct the kernel less accurately than the tuned controls and deploy more accurately. For normalized CP factors we establish conditions under which the objective attains a minimum.

AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the manuscript text, for assistance in writing and debugging experiment and code analysis, and for literature search. All derivations were carried out and checked by the authors; all code was reviewed and its outputs verified by the authors; all experimental numbers reported were produced by the authors' own code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every closed form is proved in Appendix A under the single stochastic assumption stated in §4.1. The objective and its effective ranges are specified in §5 and Appendix B, including the derived constants $\mu(b)$ and $\kappa(b, n)$; the reconstruction–quantization trade-off has no tuned coefficient. The experimental protocol (models, evaluation set, calibration set size, rank rule, baselines, deployment quantizers, BN recalibration and seeds) is in §6.1. Source code for the QALF solver, the CP-Σ, CP-EPC and CP-ADMM baselines, the gauge optimizer, the deployment quantizers and the experiment drivers is included as supplementary material; the language-model code will be released with the camera-ready version.

REFERENCES

Daria Cherniuk, Stanislav Abukhovich, Anh-Huy Phan, Ivan Oseledets, Andrzej Cichocki, and Julia Gusak. Quantization Aware Factorization for Deep Neural Network Compression. *Journal of Artificial Intelligence Research*, 81:973–988, December 2024. ISSN 1076-9757. doi: 10.1613/jair.1.16167.

Vin de Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem, April 2008.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Elias Frantar and Dan Alistarh. Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4475–4488. Curran Associates, Inc., 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, March 2023.

Alper Kalle, Théo Rudkiewicz, Mohamed Ouerfelli, and Mohamed Tamaazousti. Distribution-Aware Tensor Decomposition for Compression of Convolutional Neural Networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition, April 2015.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction. In *International Conference on Learning Representations*, 2021.

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, May 2024.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or Down? Adaptive Rounding for Post-Training Quantization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7197–7206. PMLR, November 2020.

Anh-Huy Phan, Petr Tichavský, and Andrzej Cichocki. Error Preserving Correction: A Method for CP Decomposition at a Target Error Bound. *IEEE Transactions on Signal Processing*, 67(5): 1175–1190, March 2019. ISSN 1941-0476. doi: 10.1109/TSP.2018.2887192.

Anh-Huy Phan, Konstantin Sobolev, Konstantin Sozykin, Dmitry Ermilov, Julia Gusak, Petr Tichavsky, Valeriy Glukhov, Ivan Oseledets, and Andrzej Cichocki. Stable Low-rank Tensor Decomposition for Compression of Convolutional Neural Network, August 2020.

Rajarshi Saha, Varun Srivastava, and Mert Pilanci. Matrix Compression via Randomized Low Rank and Low Precision Factorization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18828–18872. Curran Associates, Inc., 2023. doi: 10.52202/075280-0826.

Rajarshi Saha, Naomi Sagan, Varun Srivastava, Andrea J. Goldsmith, and Mert Pilanci. Compressing Large Language Models using Low Rank and Low Precision Decomposition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 88981–89018. Curran Associates, Inc., 2024. doi: 10.52202/079017-2823.

Piyush Sao, Narasinga Miniskar, Pedro Valero-Lara, Keita Teranishi, and Sudip Seal. Contraction-Gauge Preconditioning for Quantized Matrix Multiplication, July 2026.

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. *International Conference on Learning Representations*, 2024:45472–45496, May 2024.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression. *International Conference on Learning Representations*, 2025:19299–19319, May 2025.

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 *Proceedings of the 40th International Conference on Machine Learning*, pp. 38087–38099. PMLR, July 2023.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models, August 2025.

Cheng Zhang, Jianyi Cheng, George Anthony Constantinides, and Yiren Zhao. LQER: Low-Rank Quantization Error Reconstruction for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 58763–58779. PMLR, July 2024.

Cheng Zhang, Jeffrey T. H. Wong, Can Xiao, George Constantinides, and Yiren Zhao. QERA: An Analytical Framework for Quantization Error Reconstruction. *International Conference on Learning Representations*, 2025:12531–12560, May 2025.

Ya Zhou, Raymond K. W. Wong, and Kejun He. Tensor Linear Regression: Degeneracy and Solution. *IEEE Access*, 9:7775–7788, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3049494.

A Full Proofs

Throughout this section the continuous factors and the input second moment Σ are fixed, and expectations are over the perturbation model. Let $b \geq 2$, $q_b = 2^{b-1} - 1$, and $\mu(b) = 1/(12q_b^2)$. We use the channelwise second moments and independence in Assumption 1 of the main text. We write $a_r = A_{:,r}$, $b_r = B_{:,r}$, and $c_r = C_{:,r}$.

A.1 Output-factor error

With $\delta A = Q_b^A(A) - A$ the quantization noise introduced by the quantization of A. Multilinearity gives the exact change in the reconstructed kernel when only A is perturbed:

$$[\![\delta A, B, C]\!]_{(1)} = \delta A Z^\top, \qquad Z = C \odot B \tag{21}$$

The corresponding squared output perturbation is

$$E_A^2 := \|\delta A Z^\top \Sigma^{1/2}\|_F^2 = \mathrm{tr}(\delta A^\top \delta A\, M_A), \qquad M_A = Z^\top \Sigma Z \tag{22}$$

Per-row quantization of A assigns the same modeled variance to all entries in a row. Hence

$$\mathbb{E}[(\delta A)_{t,r}(\delta A)_{t',r'}] = \mathbf{1}_{t=t'}\mathbf{1}_{r=r'}\mu(b) w_{b,R}(A_{t,:})^2, \tag{23}$$

$$\mathbb{E}[\delta A^\top \delta A] = \mu(b)\rho_w(A)^2 I_R, \qquad \rho_w(A)^2 := \sum_{t=1}^{T} w_{b,R}(A_{t,:})^2 \tag{24}$$

Taking expectations in equation 22 proves

$$\mathbb{E}[E_A^2] = \mu(b)\rho_w(A)^2\,\mathrm{tr}(M_A) = \mu(b)\Phi_A \tag{25}$$

The granularity determines the contraction. If independent errors instead have variances $v_{t,r}$ that vary within each row, then

$$\mathbb{E}[E_A^2] = \sum_{r=1}^{R}\left(\sum_{t=1}^{T} v_{t,r}\right)(M_A)_{rr}. \tag{26}$$

The scalar-times-trace expression follows from tying the variance across columns within each row; it is not independent of the quantizer granularity.

A.2 Interior-factor errors

For $\delta B = Q_b^B(B) - B$,

$$E_B^2 := \|A V^\top \Sigma^{1/2}\|_F^2 = \mathrm{tr}(A^\top A\, V^\top \Sigma V), \qquad V = C \odot \delta B \tag{27}$$

We index the unfolded input by (d, s) consistently with $c_r \otimes b_r$. We can then define the principal submatrices

$$(\Sigma^{(s,s)})_{d,d'} := \Sigma_{(d,s),(d',s)}, \qquad (\Sigma^{[d,d]})_{s,s'} := \Sigma_{(d,s),(d,s')}. \tag{28}$$

Since $V_{(d,s),r} = C_{d,r}(\delta B)_{s,r}$ and each column of B is one channel,

$$\mathbb{E}[(\delta B)_{s,r}(\delta B)_{s',r'}] = \mathbf{1}_{s=s'}\mathbf{1}_{r=r'}\mu(b) w_{b,S}(b_r)^2, \tag{29}$$

$$\mathbb{E}[(V^\top \Sigma V)_{r,r'}] = \mathbf{1}_{r=r'}\mu(b) w_{b,S}(b_r)^2 \sum_{s=1}^{S} c_r^\top \Sigma^{(s,s)} c_r. \tag{30}$$

Consequently, with

$$\kappa_r^B := \sum_{s=1}^{S} c_r^\top \Sigma^{(s,s)} c_r, \qquad \kappa_r^C := \sum_{d=1}^{D^2} b_r^\top \Sigma^{[d,d]} b_r, \tag{31}$$

the two interior-factor errors are

$$\mathbb{E}[E_B^2] = \mu(b)\sum_{r=1}^{R} \|a_r\|_2^2 w_{b,S}(b_r)^2 \kappa_r^B = \mu(b)\Phi_B, \tag{32}$$

$$\mathbb{E}[E_C^2] = \mu(b)\sum_{r=1}^{R} \|a_r\|_2^2 w_{b,D^2}(c_r)^2 \kappa_r^C = \mu(b)\Phi_C \tag{33}$$

The second calculation exchanges the roles of the spatial and channel indices. Independence eliminates covariance entries that are off-diagonal in the perturbed index. Thus Φ_B and Φ_C retain the indicated principal submatrices, whereas Φ_A uses the full input second moment through M_A. In general, the component-dependent weights in these sums prevent separation into a scalar function of the perturbed factor times a function of the other factors.

A.3 FACTORIZED LINEAR MAPS

Let $F_k \in \mathbb{R}^{d_k \times d_{k-1}}$ and consider the linear map $F_m \cdots F_1$. Defining

$$L_k := F_m \cdots F_{k+1}, \qquad P_k := F_{k-1} \cdots F_1, \tag{34}$$

with empty products equal to the appropriate identity. Perturbing only F_k by δF_k changes the map exactly by $L_k \delta F_k P_k$. Set $G = L_k^\top L_k$ and $M = P_k \Sigma P_k^\top$. Then

$$E_k^2 = \|L_k \delta F_k P_k \Sigma^{1/2}\|_F^2 = \operatorname{tr}(\delta F_k^\top G \delta F_k M). \tag{35}$$

Under independent per-row errors,

$$\mathbb{E}[(\delta F_k^\top G \delta F_k)_{j,j'}] = \sum_{i,i'} G_{i,i'} \mathbb{E}[(\delta F_k)_{i,j} (\delta F_k)_{i',j'}] \tag{36}$$

$$= \mathbf{1}_{j=j'} \mu(b) \sum_i \|(L_k)_{:,i}\|_2^2 w_{b,d_{k-1}}((F_k)_{i,:})^2. \tag{37}$$

It follows that

$$\mathbb{E}[E_k^2] = \mu(b) \left(\sum_i \|(L_k)_{:,i}\|_2^2 w_{b,d_{k-1}}((F_k)_{i,:})^2 \right) \operatorname{tr}(P_k \Sigma P_k^\top) \tag{38}$$

For the output factor, $L_m = I$, so its noise scale depends only on its own rows. For the input factor, $P_1 = I$, so the input metric enters its single-factor price only through $\operatorname{tr}(\Sigma)$. Interior CP factors are structured factors of a Khatri–Rao product; their formulas in Appendix A.2 cannot be obtained by treating their entries as independent entries of an unstructured chain matrix.

A.4 JOINT QUANTIZATION AND THE QALF APPROXIMATION

For a nonempty subset $I \subseteq \{A, B, C\}$, let Δ_I denote the CP tensor obtained by replacing exactly the factors indexed by I with their perturbations. Multilinearity yields

$$\hat{\mathcal{W}}_Q = [\![A + \delta A, B + \delta B, C + \delta C]\!] = \hat{\mathcal{W}} + \sum_{\varnothing \neq I \subseteq \{A,B,C\}} \Delta_I. \tag{39}$$

Define $\langle U, V \rangle_\Sigma := \operatorname{tr}(U_{(1)} \Sigma V_{(1)}^\top)$ and $\|U\|_\Sigma^2 := \langle U, U \rangle_\Sigma$. If $I \neq J$, at least one noise factor occurs in only one of the two terms. Conditioning on every other factor leaves an expression linear in that independent, zero-mean perturbation. Therefore

$$\mathbb{E}\langle \Delta_I, \Delta_J \rangle_\Sigma = 0 \quad (I \neq J), \qquad \mathbb{E}\langle \mathcal{W} - \hat{\mathcal{W}}, \Delta_I \rangle_\Sigma = 0. \tag{40}$$

Expanding the squared loss gives the exact identity

$$\mathcal{R}_b = \mathcal{L}_\Sigma + \mu(b)(\Phi_A + \Phi_B + \Phi_C) + R_{\geq 2}, \tag{41}$$

$$R_{\geq 2} := \mathbb{E}\|\Delta_{\{A,B\}}\|_\Sigma^2 + \mathbb{E}\|\Delta_{\{A,C\}}\|_\Sigma^2 + \mathbb{E}\|\Delta_{\{B,C\}}\|_\Sigma^2 + \mathbb{E}\|\Delta_{\{A,B,C\}}\|_\Sigma^2 \geq 0. \tag{42}$$

The objective $\mathcal{J}_b$ retains the reconstruction and single-perturbation terms.

For completeness, the omitted terms can also be evaluated explicitly. Write $u_r := w_{b,S}(b_r)^2$, $v_r := w_{b,D^2}(c_r)^2$, and $\rho^2 := \rho_w(A)^2$. Independence gives

$$\begin{aligned} R_{\geq 2} = {} & \mu(b)^2 \rho^2 \sum_{r=1}^R (u_r \kappa_r^B + v_r \kappa_r^C) \\ & + \mu(b)^2 \operatorname{tr}(\Sigma) \sum_{r=1}^R \|a_r\|_2^2 u_r v_r \\ & + \mu(b)^3 \rho^2 \operatorname{tr}(\Sigma) \sum_{r=1}^R u_r v_r. \end{aligned} \tag{43}$$

The double-noise terms are quadratic and the triple-noise term is cubic in the factorwise variances. This order statement does not by itself bound their size relative to the retained terms at a given precision. In particular, the effective ranges can depend on b. The nonnegative remainder is a property of the stochastic model; it is not an upper-bound guarantee for deterministic clipped RTN or GPTQ.

A.5 CONVEX PROGRAM FOR THE ABSMAX ORBIT

Consider the positive diagonal rescaling

$$A' = AD_\gamma^{-1}, \quad B' = BD_\alpha, \quad C' = CD_\beta, \quad \gamma_r = \alpha_r\beta_r, \tag{44}$$

where $D_\lambda = \mathrm{diag}(\lambda)$. Let $m_r = (c_r \otimes b_r)^\top \Sigma (c_r \otimes b_r)$ and $v_r = \gamma_r^2$. For the absmax scale,

$$\Phi_A(v) = F(v)G(v), \qquad F(v) := \sum_{t=1}^{T} \max_r \frac{A_{t,r}^2}{v_r}, \quad G(v) := \sum_{r=1}^{R} m_r v_r. \tag{45}$$

For $\tau > 0$, $F(\tau v) = F(v)/\tau$ and $G(\tau v) = \tau G(v)$. Whenever $G(v) > 0$, normalizing to $G(v) = 1$ leaves the product unchanged. Each reciprocal term is convex on $v_r > 0$; maxima and sums preserve convexity. The epigraph problem is therefore

$$\begin{aligned} \underset{v>0,\ s\geq 0}{\text{minimize}} \quad & \sum_{t=1}^{T} s_t \\ \text{subject to} \quad & s_t v_r \geq A_{t,r}^2 \qquad (t = 1, \ldots, T;\ r = 1, \ldots, R), \\ & \sum_{r=1}^{R} m_r v_r = 1. \end{aligned} \tag{46}$$

Each product constraint has the rotated second-order-cone representation

$$(s_t, v_r, \sqrt{2}A_{t,r}) \in \{(x, y, z) : x, y \geq 0,\ 2xy \geq z^2\}.$$

If $\Sigma \succ 0$ and zero CP components have been removed, then $m_r > 0$ and every column of A is nonzero. The constraint bounds v_r above, while the objective diverges as any $v_r \downarrow 0$. A minimizer is thus attained with positive v. Degenerate zero components should be removed before applying this formulation.

The clipped scale remains homogeneous, so common-scale invariance still holds. The absmax epigraph derivation does not establish convexity for the clipped orbit objective; optimizing that objective in log-coordinates is a separate numerical procedure.

A.6 WELL-POSEDNESS ON THE CANONICAL SLICE

Assume $\Sigma \succeq \varepsilon I$ for some $\varepsilon > 0$, and that each scale functional is continuous, degree-one homogeneous, and positive away from zero. For a fixed finite $b \geq 2$, work on

$$\mathcal{S} = \mathbb{R}^{T\times R} \times (\mathbb{S}^{S-1})^R \times (\mathbb{S}^{D^2-1})^R. \tag{47}$$

Here $\|b_r\|_2 = \|c_r\|_2 = 1$. Accordingly,

$$m_r \geq \varepsilon \|c_r \otimes b_r\|_2^2 = \varepsilon, \qquad \mathrm{tr}(M_A) \geq R\varepsilon. \tag{48}$$

Continuity and positivity on the compact unit sphere give

$$c_w := \min_{\|x\|_2=1} w_{b,R}(x) > 0, \qquad w_{b,R}(x) \geq c_w \|x\|_2. \tag{49}$$

Since the loss and all penalties are nonnegative,

$$\mathcal{J}_b \geq \mu(b)\Phi_A \geq \mu(b) c_w^2 R\varepsilon \|A\|_F^2. \tag{50}$$

Thus $\mathcal{J}_b$ is uniformly coercive in A on $\mathcal{S}$. A minimizing sequence has bounded A, and its other factors lie in compact products of spheres. A convergent subsequence therefore exists, and continuity gives attainment of the minimum. Along any canonical sequence with $\max_r \|a_r\|_2 \to \infty$, the objective tends to infinity.

This result concerns the normalized domain and the independent-noise objective. It is neither a convergence theorem for the numerical solver nor a proof of coercivity on the unrestricted factor space or for the GPTQ-aware surrogate. Internal inverse rescaling of B and C remains a flat direction

in the unrestricted space. Moreover, normalizing both factors and moving their scales into A can change Φ_A; see Appendix B.3.

For a fixed bit-independent scale, $\mu(b) \to 0$ recovers $\mathcal{L}_\Sigma$ pointwise. The same limit holds for the clipped scale in Appendix B.1, because its squared effective range is bounded by $\|x\|_\infty^2$ for each fixed channel. Pointwise convergence does not imply convergence of minimizers, or equality with unregularized ALS at any finite precision. Classical CP approximation can fail to attain a best fixed-rank approximation in the settings identified by de Silva & Lim (2008); nonattainment is not asserted for every rank, dimension, and target tensor.

A.7 Selection along a fixed CP orbit

Under equation 44, the reconstruction is unchanged. Degree-one homogeneity of the channel scale and quadratic homogeneity of κ_r^B imply

$$\|a'_r\|_2^2 w_{b,S}(b'_r)^2 \kappa_r^{B\prime} = \gamma_r^{-2}\alpha_r^2\beta_r^2\|a_r\|_2^2 w_{b,S}(b_r)^2\kappa_r^B = \|a_r\|_2^2 w_{b,S}(b_r)^2\kappa_r^B. \tag{51}$$

Thus Φ_B is invariant; the same argument applies to Φ_C. Only Φ_A can vary, and on the orbit

$$\mathcal{J}_b|_{\text{orbit}} = \text{constant} + \mu(b)\Phi_A(\gamma). \tag{52}$$

For a bit-independent scale such as absmax, multiplying by the positive coefficient $\mu(b)$ does not change the set of orbit minimizers. For the clipped scale, $w_{b,n}$ itself depends on b, and the preferred representative can change with precision. This statement is about the first-order QALF objective; it does not identify the optimizer of every higher-order or deterministic quantization risk.

B Effective ranges and optimization details

B.1 Clipped effective range

For a channel $a \in \mathbb{R}^n$, define

$$D_b(a,c) := \mu(b)c^2 + \frac{1}{n}\sum_{i=1}^n (|a_i| - c)_+^2, \qquad c \geq 0, \tag{53}$$

$$\mu(b)w_{b,n}(a)^2 := \min_{c\geq 0} D_b(a,c). \tag{54}$$

The first term models granular noise for the step c/q_b; the second is the mean squared overload error. This is a continuous distortion surrogate. It is not the exact deterministic MSE of rounding each channel entry, and its optimum need not coincide with that of a finite deployment clipping search. The effective range $w_{b,n}(a)$ includes overload distortion and is generally larger than the minimizing clipping threshold $c^\star$.

Let $u_{(1)} \geq \cdots \geq u_{(n)}$ be the sorted absolute entries and set $u_{(n+1)} = 0$. On an interval $u_{(k+1)} \leq c \leq u_{(k)}$, only the first k entries can contribute to the clipping term, giving

$$D_{b,k}(c) = \mu(b)c^2 + \frac{1}{n}\sum_{i=1}^k (u_{(i)} - c)^2. \tag{55}$$

With $S_k = \sum_{i=1}^k u_{(i)}$, its stationary point is

$$c_k = \frac{S_k}{k + n\mu(b)} = \frac{S_k}{k + n/(12q_b^2)}. \tag{56}$$

For a nonzero channel, choose a candidate satisfying $u_{(k+1)} \leq c_k \leq u_{(k)}$. The objective is strictly convex in c, so the minimizing threshold is unique, including when multiple active-set descriptions coincide at a boundary. If $a = 0$, take $c^\star = 0$. Sorting costs $O(n \log n)$ and the prefix-sum scan costs $O(n)$.

Equivalently, with $\kappa(b,n) = 1/(n\mu(b)) = 12q_b^2/n$,

$$w_{b,n}(a)^2 = \min_{c\geq 0}\left[c^2 + \kappa(b,n)\sum_i (|a_i| - c)_+^2\right]. \tag{57}$$

The function inside the minimum is jointly convex in (a, c); its partial minimum is therefore convex in a. Uniqueness of the minimizing threshold gives the envelope gradient

$$\frac{\partial w_{b,n}(a)^2}{\partial a_i} = 2\kappa(b,n)\,\mathrm{sgn}(a_i)\,(|a_i| - c^\star)_+\,, \tag{58}$$

with derivative zero at $a_i = 0$. Thus the squared effective range can be differentiated without differentiating the sorting operation.

Replacing (a, c) by $(\alpha a, |\alpha| c)$ multiplies the objective by $|\alpha|^2$, so

$$c^\star(\alpha a) = |\alpha| c^\star(a), \qquad w_{b,n}(\alpha a) = |\alpha| w_{b,n}(a). \tag{59}$$

The squared effective range is continuous, finite, and zero only at $a = 0$: a zero minimum would require both $c = 0$ and every $a_i = 0$. Finally, evaluating the objective at $c = \|a\|_\infty$ gives

$$w_{b,n}(a)^2 \le \|a\|_\infty^2 \le \|a\|_2^2. \tag{60}$$

These properties verify the scale assumptions used in the canonical well-posedness proof. The coefficient $\kappa(b, n)$ is fixed by the precision and channel length, rather than independently tuned.

B.2 A LOCAL COMPENSATION COEFFICIENT

For a positive-definite curvature matrix H_δ, consider a perturbation vector z whose i-th coordinate is constrained to the rounding error e. A Lagrange-multiplier calculation gives

$$\min_{z:z_i=e} z^\top H_\delta z = \frac{e^2}{[H_\delta^{-1}]_{ii}}, \qquad z^\star = \frac{e H_\delta^{-1} e_i}{[H_\delta^{-1}]_{ii}}, \tag{61}$$

where e_i is the coordinate basis vector. This motivates

$$H_\delta = H + \delta\,\mathrm{diag}(H), \qquad c_{\mathrm{obq}}(H_\delta) = \sum_i \frac{1}{[H_\delta^{-1}]_{ii}}. \tag{62}$$

The coordinate inequality $[H_\delta^{-1}]_{ii}(H_\delta)_{ii} \ge 1$ implies

$$c_{\mathrm{obq}}(H_\delta) \le \mathrm{tr}(H_\delta), \tag{63}$$

with equality for diagonal positive-definite curvature. The definition requires positive definiteness; diagonal damping alone does not repair a zero diagonal entry automatically.

This is a local compensation calculation in the spirit of OBQ/GPTQ (Frantar & Alistarh, 2022; Frantar et al., 2023). Summing costs computed with a single full inverse does not account for every changing active set, error, and curvature in a sequential rounder. Accordingly, $\Phi_A^{\mathrm{GPTQ}} = \rho_w(A)^2 c_{\mathrm{obq}}(M_{A,\delta})$ is a surrogate, not the expectation proved for independent RTN noise. The comparison above is with $\mathrm{tr}(H_\delta)$, not automatically with $\mathrm{tr}(H)$.

The exact GPTQ coupling. GPTQ freezes each column once it is rounded, and its loss is exactly $\sum_j e_j^2/U_{jj}^2$ with $U^\top U = H_\delta^{-1}$ (upper Cholesky factor, in the rounder's column order). Under the channelwise noise model the expected GPTQ error of A is therefore $\mu(b)\rho_w(A)^2 c_{\mathrm{seq}}(M_{A,\delta})$ with $c_{\mathrm{seq}} = \sum_j 1/U_{jj}^2$, and $c_{\mathrm{obq}} \le c_{\mathrm{seq}} \le \mathrm{tr}$. Measured on A over the 16 factorized layers of ResNet-18, the realized-to-predicted ratio is 1.00 (range 0.84–1.09) for c_{seq}, against 2.02 for c_{obq} and 0.31 for the trace. Pricing the rounder that is actually deployed pays off before BN recalibration: at $4\times$, 4 bits, a GPTQ-aware coupling in the solver raises accuracy from 56.44% to 58.35% (+1.9). Recalibration absorbs the difference (64.77% against 64.83%), which is why the experiments use the RTN coupling.

B.3 NORMALIZATION AND A FINITE-BUDGET BLOCK PROCEDURE

For nonzero interior columns, define $s_r^B = \|b_r\|_2$ and $s_r^C = \|c_r\|_2$. The transformation

$$a_r \leftarrow a_r s_r^B s_r^C, \qquad b_r \leftarrow b_r / s_r^B, \qquad c_r \leftarrow c_r / s_r^C \tag{64}$$

preserves the reconstructed kernel and returns the factors to $\mathcal{S}$. It also preserves Φ_B and Φ_C, but generally changes Φ_A. It is therefore a representation selection step, rather than an objective-neutral normalization. By contrast, $b_r \mapsto \eta_r b_r$, $c_r \mapsto \eta_r^{-1} c_r$, with a_r unchanged, preserves all three prices. The theorem on the canonical slice must be distinguished from unconstrained block optimization followed by equation 64.

Given an initialization, a sweep budget, and a selected RTN or GPTQ-aware objective, a finite-budget procedure consists of estimating Σ, cycling over the factor blocks, applying the specified normalization, and then deploying the quantized factors. If normalization is used, objective trajectories should be evaluated at a consistent point after that operation. No convergence claim follows from the iteration budget alone. The reconstruction–quantization coefficient is determined by b; this does not eliminate numerical choices such as initialization, damping, iteration caps, or the treatment of zero components.

B.4 Contraction cost

Let $d = SD^2$. Forming ΣZ with a dense input second moment costs $O(d^2 R)$. If $\Sigma = U^\top U/N$, where the rows of $U \in \mathbb{R}^{N \times d}$ are calibration patches, then

$$M_A = (UZ)^\top (UZ)/N, \qquad \Sigma Z = U^\top (UZ)/N. \tag{65}$$

The leading patch contractions cost $O(NdR)$, with an additional $O(NR^2)$ to form the Gram matrix. End to end, a 300-sweep QALF build of the 16 factorized layers of ResNet-18 at $3\times$ takes about 45 minutes on one RTX 3070 GPU in our implementation; a CP-Σ build takes about 5 minutes.

C Two-factor specialization and matrix gauges

Let $W \approx LR$, where $L \in \mathbb{R}^{n_o \times r}$ and $R \in \mathbb{R}^{r \times n_i}$. In this section R denotes the right matrix factor, and r is the rank. Set $M_L = R\Sigma R^\top$ and write $\ell_k = L_{:,k}$. The functional reconstruction loss is $\|(W - LR)\Sigma^{1/2}\|_F^2$.

C.1 Per-row deployment prices

Specializing equation 38 to two factors gives

$$\mathbb{E}[E_L^2] = \mu(b)\rho_w(L)^2 \operatorname{tr}(M_L), \qquad \rho_w(L)^2 = \sum_{t=1}^{n_o} w_{b,r}(L_{t,:})^2, \tag{66}$$

$$\mathbb{E}[E_R^2] = \mu(b) \operatorname{tr}(\Sigma) \sum_{k=1}^{r} \|\ell_k\|_2^2 w_{b,n_i}(R_{k,:})^2. \tag{67}$$

Write these as $\mu(b)\Phi_L$ and $\mu(b)\Phi_R$, respectively. The input metric enters the output-factor price through $R\Sigma R^\top$, whereas the right-factor price depends on it only through its trace under this particular independent per-row model. This does not preclude other noise models or grouping rules from retaining more directional information. When both factors are quantized, the first-order objective includes both prices. As in Appendix A.4, the exact modeled risk additionally contains the nonnegative double-noise term $\mathbb{E}\|\delta L \delta R \Sigma^{1/2}\|_F^2$.

For a two-sided metric with an output-side weight $\Sigma_g \succeq 0$ (an output second moment, or any positive semidefinite weighting of the layer output), the same calculation also gives

$$\mathbb{E}\|\Sigma_g^{1/2} \delta L R \Sigma^{1/2}\|_F^2 = \mu(b) \operatorname{tr}(M_L) \sum_{t=1}^{n_o} (\Sigma_g)_{tt} w_{b,r}(L_{t,:})^2. \tag{68}$$

Independence of the output-factor perturbations removes the off-diagonal entries of Σ_g from this expectation.

C.2 Groupwise quantization

Let $\mathcal{G}$ partition the rank indices of each row of L, and let $\mathcal{H}$ partition the input coordinates of each row of R. With independent, constant-variance errors within each row–group channel, the

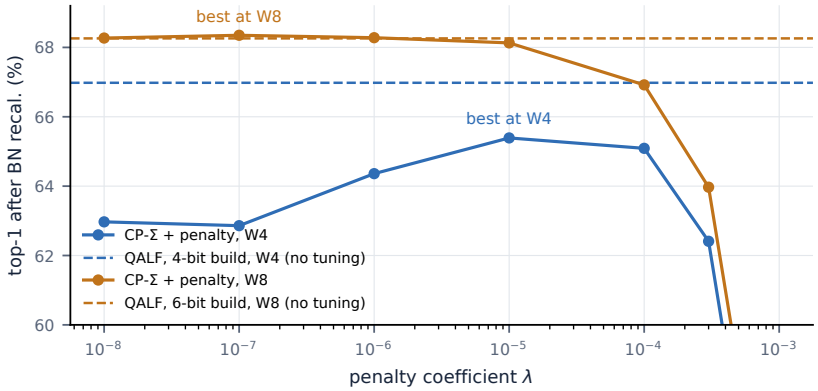

Figure 2: CP-Σ with the data-scaled norm penalty $\lambda \operatorname{tr}(M_A)\|A\|_F^2$, deployed at 4 and 8 bits (ResNet-18, 3×, symmetric grid, seed 0, identical deployment). The best λ differs between the two bit-widths; dashed lines are QALF, with no coefficient to tune (its 4-bit build at W4 and its 6-bit build at W8).

corresponding prices are

$$\mathbb{E}[E_L^2] = \mu(b) \sum_{t=1}^{n_o} \sum_{g \in \mathcal{G}} w_{b,|g|} (L_{t,g})^2 \sum_{k \in g} (M_L)_{kk}, \tag{69}$$

$$\mathbb{E}[E_R^2] = \mu(b) \sum_{k=1}^{r} \|\ell_k\|_2^2 \sum_{h \in \mathcal{H}} w_{b,|h|} (R_{k,h})^2 \sum_{j \in h} \Sigma_{jj}. \tag{70}$$

These follow by using the group-dependent entrywise variances in the trace expansion, as in equation 26. A groupwise deployment should therefore use groupwise prices; the full-row formulas do not apply to it.

C.3 Whitened reconstruction and compensation

Assume $\Sigma \succ 0$, and let $W\Sigma^{1/2} = U\Lambda V^\top$ be its singular value decomposition. A rank-r weighted reconstruction optimum is

$$L = U_r \Lambda_r^a, \qquad R = \Lambda_r^{1-a} V_r^\top \Sigma^{-1/2}. \tag{71}$$

Indeed, this construction truncates the SVD in the variable $W\Sigma^{1/2}$. Its output-factor coupling is diagonal:

$$M_L = R\Sigma R^\top = \Lambda_r^{2(1-a)}. \tag{72}$$

For fixed quantization grids, the resulting output-factor rounding objective separates across coordinates. Nearest-level rounding minimizes each coordinate term, and off-diagonal compensation has no effect for that curvature. This statement concerns the output-factor subproblem; it does not imply optimal quantization of the whole network, nor does it make the diagonal gauge irrelevant when grid ranges depend on the factors.

D Additional results

Table 4 reports, at 2× compression and a single seed, the full-precision accuracy of each factorization next to its deployed accuracy. Table 5 deploys the QALF 4-bit build at 4, 6 and 8 bits. Figure 2 shows the full sweep of the norm-penalty control of §6.2: its best coefficient moves with the bit-width, while QALF's weight is fixed by it.

Table 4: ResNet-18 accuracy percentage at $2\times$ compression. Each entry is factorized FP32 / quantized top-1; the FP32 value belongs to the same method and target bit-width as the deployed value. For CP-Σ only the quantized results are shown; its FP32 accuracy is 68.59. Grids: pca = per-channel affine, pcs = per-channel symmetric. * Initialized from CP-Σ.

Factorization $\times 2$	deployment	W4A8	W6A8	W8A8
CP-Σ	RTN pca	63.81	68.10	68.26
	GPTQ pca	66.45	68.18	68.35
QALF	RTN pcs	68.75 / 68.05	68.84 / 68.41	68.98^* / 68.69^*
	GPTQ pca	68.25	68.50	68.74^*

Table 5: Robustness to the selected bit-width.

Method	FP32	Rounding	W4A8	W6A8	W8A8
QALF	68.75	RTN	68.05	68.57	68.74
		GPTQ	68.25	68.64	68.67

The covariance floor in language models. Table 6 separates the price from the floor $\Sigma + \varepsilon I$, $\varepsilon = \lambda_1(\Sigma)/n$, used in the solve of Table 3. The floor changes the full-precision cost — down on Pythia-410M, up on Qwen2.5-0.5B — and barely changes how much rounding damage the price removes. The gain comes from the price. On Llama-3.2-3B and Llama-3.1-8B the baseline's GPTQ is itself unstable across calibration seeds: first-token outliers dominate its Hessian in the early blocks, so one seed of each model is far worse than the other two (per-seed damage removed 70.2/77.6/90.3% and 83.3/94.5/81.7%). Across the models of Table 3, the largest standard deviation of QALF's rounded perplexity over the three seeds is 0.18.

Table 6: The covariance floor, 3-bit weights, same protocol and seeds as Table 3: WikiText-2 perplexity before / after rounding both factors (mean over 3 seeds; $\pm$ is the sd after rounding).

	SVD-LLM	QALF, with floor	QALF, no floor
Pythia-410M	20.57 / 25.99	21.24 / 22.50 $\pm$0.03	21.68 / 23.02 $\pm$0.04
Qwen2.5-0.5B	15.89 / 20.09	16.97 / 18.16 $\pm$0.09	16.53 / 17.75 $\pm$0.01
damage removed	–	76.8% / 71.6%	75.3% / 71.0%

Data-free QALF. Table 7 replaces Σ by the identity in QALF. At 4 bits the data-free price still beats data-aware CP-Σ; at 6 bits, where $\mu(b)$ is 20 times smaller, the input weighting matters more and CP-Σ is ahead.

GoogLeNet with 8-bit activations. Table 8 reports QALF on GoogLeNet under the protocol of Table 1, across compression factors and bit-widths.

CP-ADMM on a per-channel affine grid. Table 9 constrains CP-ADMM to the per-channel affine grid instead of its native per-tensor grid. At 8 bits the runs are unstable across seeds (standard deviation up to 1.27) and at $2\times$ the 8-bit result falls below the 4-bit one, so the main table reports the method on the grid its authors use. At 4 bits the variant is stable and stronger than the native grid at $3\times$ and $4\times$ (65.05 and 61.05 against 64.37 and 60.18): against the better of the two variants at each ratio, QALF gains 0.95, 1.65 and 3.56 points rather than 0.95, 2.33 and 4.43.

CP-Σ and the number of sweeps. Table 10 runs CP-Σ longer. Full-precision accuracy keeps improving while the 4-bit deployment collapses, consistent with the component growth discussed in §4.4.

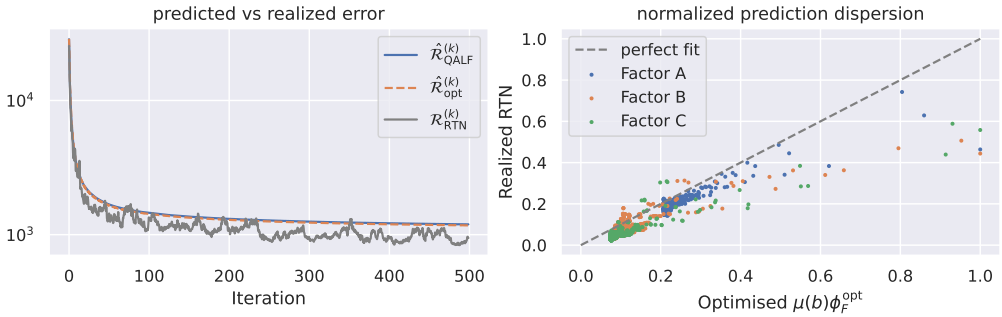

Figure 3: Relationship between the modeled, optimized, and realized deployment objectives for ResNet18's `layer1.0.conv1` layer under 4 bit per-channel symmetric target.

Table 7: Data-free QALF ($\Sigma = I$) against data-aware QALF and CP-Σ (ResNet-18, top-1 %, 8-bit activations; RTN on the per-channel symmetric grid for QALF, GPTQ on the per-channel affine grid).

		W4A8		W6A8	
		2×	2.5×	2×	2.5×
QALF	RTN	68.05	67.06	68.41	68.31
	GPTQ	68.25	67.50	68.50	68.23
QALF data-free	RTN	67.09	65.37	67.27	66.38
	GPTQ	67.81	66.54	67.68	66.88
CP-Σ	RTN	63.81	62.92	68.10	67.41
	GPTQ	66.45	65.71	68.18	67.38

Table 8: GoogLeNet, QALF, top-1 % after weight quantization of the factorized layers (8-bit activations; RTN per-channel symmetric, GPTQ per-channel affine).

	W4A8			W6A8		
	2×	2.5×	3×	2×	2.5×	3×
RTN	68.34	67.82	67.38	68.78	68.54	68.00
GPTQ	68.36	67.92	67.51	68.80	68.38	68.05

Table 9: CP-ADMM constrained to the per-channel affine grid (mean ± sd over 3 seeds, top-1 %, 8-bit activations).

	W4A8			W8A8		
	2×	3×	4×	2×	3×	4×
CP-ADMM (pca)	67.13 ±0.68	65.05 ±0.34	61.05 ±0.11	66.91 ±1.27	–	62.42 ±0.24

Table 10: CP-Σ at 2× against the number of alternating sweeps (top-1 %, RTN on the per-channel affine grid).

sweeps	10	20	30	50	100
FP32	68.59	68.77	68.89	68.94	69.03
W8A8	68.26	68.54	68.51	68.44	68.17
W4A8	63.81	62.90	61.27	56.67	48.20