

When Does Compensating Inherited Quantization Error Help? Mean-Square Guarantees for Layer-Wise Quantization and Their Limits

Tianhao Guo*

12 October 2026

Abstract

Layer-wise post-training quantization methods let each layer compensate for the quantization error inherited from earlier layers. We ask under what assumptions such compensation guarantees a network output error no larger than without it. For linear networks of any depth, with the required calibration Gram matrices invertible, if the rounding error of each layer has zero conditional mean and a deterministic second-moment map common to both schemes, as under subtractive dither on an unbounded grid, then compensation with strengths in $[0, 1]$ fixed in advance gives a mean-square output error no larger than no compensation, each scheme being followed along its own trajectory. The exact gain separates error removed directly from reduced excitation of later rounding noise. Outside this setting the comparison can fail. With rectified linear activations it fails for two independent reasons: noise changes activation states, and, even when no state changes, the correction moves error from samples on which a unit is inactive to samples on which it is active. With stable activations, commutators between the calibration projection and the activation patterns bound the excess. With a bounded codebook, even unique exact layer-wise minimizers with no overload can give a larger output error. For fixed rounding errors it already fails for two scalar linear layers, refuting a published universal guarantee. An exact residual identity gives the necessary and sufficient condition on a common layer input. All these comparisons hold the calibration input fixed. On two 8B models at 4 bits, fixed half-strength compensation beat uncompensated rounding in every model, text and draw tested, yet no rule for choosing the strength beat a layer-wise rule consistently, and lowering the full-network loss on the data used to choose did not reliably lower it on new text.

1 Introduction

Layer-wise post-training quantization processes the linear maps of a network one at a time, choosing for each a quantized weight matrix that reproduces the layer output on calibration data [2, 3, 4, 5]. Because every layer is treated separately, the error made in early layers reaches later layers through their inputs. Several recent methods change the layer objective so that a layer matches the output of the *original* network on the *original* input, while being fed the quantized input; the layer then compensates for the inherited error. The same objective appears under the names quantization error propagation [1], asymmetric calibration [6] and activation drift correction [17], and in [7, 16].

*Tianhao Guo is with the College of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China (e-mail: tianhao_guo@sxu.edu.cn). ORCID: 0000-0002-1317-868X.
Code: <https://github.com/tellsgo/quantization-compensation>.

These methods share the motivation of reducing errors inherited from earlier quantized layers. This raises a stronger question: under what assumptions does layer-wise compensation guarantee that the network output error is no larger than without compensation? The answer is not obvious. Each layer solves a least-squares problem, and the solution is closer to the target than the uncorrected one *before* rounding. After rounding this need not be so, and a comparison on a common input says nothing yet about two schemes that, from the second layer on, see different inputs. That is the question of this paper. The comparison requires us to follow both the inherited error and the activation that multiplies the next rounding error. Reducing the first is not enough; the second must be controlled along the two trajectories as well. Throughout, the calibration input is fixed and the randomness is that of rounding; whether an improvement on calibration data carries over to other inputs is a separate question, on which Section 9 reports experiments.

Related analyses exist. SPFQ [14] gives high-probability full-network error bounds for a stochastic path-following quantizer. The analysis of OPTQ and Qronos [8] gives layer-wise bounds for both algorithms and compares their error decompositions and upper bounds. These results do not establish an ordering of the realized total errors of a compensated and an uncompensated scheme.

We establish a network-level comparison under explicit noise assumptions, give the gain exactly, and determine where the comparison fails. The fixed-error guarantee of [1] is one such failure. Its mechanism is simple: the inherited error and the new rounding error can cancel, and removing the first leaves the second without its counterpart. Whether compensation helps is therefore a question about the correlation between the two errors, and the positive results below are obtained by controlling it. We use the formulation of [1] to study a family of compensated layer objectives parameterized by the strength. The related methods motivate the analysis; their quantizers, solvers and regularization choices require separate treatment. QEP denotes the compensated scheme and BASE the uncompensated one in the model studied here.

Contributions.

1. **A comparison along two trajectories, and the gain it yields** (Section 4). For linear networks of any depth, with rectangular layers allowed, Theorem 8 gives mean-square output error no larger with compensation than without it, for strengths in $[0, 1]$ fixed in advance, under the required Gram condition, zero conditional mean and a deterministic second-moment map common to both schemes. The proof carries two moments through the layers, that of the error and that of the quantized activation; the second governs how much the next rounding error is amplified. Proposition 9 separates the gain into error removed directly and reduced excitation of later noise. These compare compensation with no compensation. Ordering two non-zero strength schedules requires additional row-space conditions (Corollary 10); the closed form of Example 11 gives a ratio of expected errors in the specified random-teacher model, not a prediction for every fixed network.
2. **Two obstructions for ReLU networks, and their geometric cost** (Section 5). The comparison fails under the same dither when noise changes activation states (Example 12), and also when no state changes (Example 14). In the latter case the correction mixes calibration samples across activation patterns. Proposition 13 characterizes when this mixing is harmless for all inherited errors. Under stable activations, Theorem 15 bounds the excess by a term controlled by commutators of the calibration projection with the sample groups; Corollary 16 gives its small-error order.
3. **The local condition and the limits of the network comparison** (Sections 3, 6 and 7). An exact residual identity gives the necessary and sufficient condition for compensation to help

on a common input. Its expectation covers subtractive dither with fixed inter-column feedback, without giving a network guarantee for GPTQ or QuIP. Fixed perturbations can instead cancel inherited error, invalidating the published universal guarantee and strength monotonicity. For an actual four-level codebook, an open family has a larger output error despite unique exact layer-wise minimizers and interior targets, with no overload (Theorem 22). Synthetic computations illustrate the positive results, and a scoped Lean formalization provides a complementary check (Sections 8 and 10). Experiments on two small and two 8B models address the question the theorems leave open, whether gains on calibration data carry over to new inputs (Section 9).

The paper goes from one layer to the network: the per-layer identity and its expectation (Section 3), linear networks (Section 4), ReLU networks (Section 5), then fixed rounding errors and the published guarantee (Section 6) and actual quantizers (Section 7).

Scope. The tools are standard: orthogonal projections, dithered rounding and order relations between second-moment matrices. The contribution is the comparison they yield and its limits. We propose no new quantizer and do not dispute reported empirical gains. We have not established stable activations in trained networks. The experiments of Section 9 do not satisfy our hypotheses and are not tests of them. The projection contraction, continuous optimum and each scheme’s own bound in [1] remain valid; Section 6 isolates the comparisons that fail.

2 Setting

Consider a feedforward network, with the notation of [1, Appendix B.3],

$$X_1 = X, \quad X_{l+1} = \sigma_l(W_l X_l), \quad l = 1, \dots, L,$$

where each σ_l is γ_l -Lipschitz in the Frobenius norm and $\sigma_l(0) = 0$. A quantized forward pass uses weights $\widehat{W}_l$: $\widehat{X}_1 = X$, $\widehat{X}_{l+1} = \sigma_l(\widehat{W}_l \widehat{X}_l)$. Write $\delta_l = X_l - \widehat{X}_l$ and $\widehat{H}_l = \widehat{X}_l \widehat{X}_l^\top$. We require $\widehat{H}_l$ to be invertible only at layers of positive strength: a layer with $\alpha_l = 0$ is not corrected, its target is W_l , and no Gram matrix is inverted there. The weights may be rectangular; invertibility of quantized weights is an additional assumption of the published fixed-error guarantee examined in Section 6, not of the positive results.

The matrices E_l represent rounding errors. Sections 4 and 5 impose conditional moment assumptions on random errors; Section 6 instead fixes the same arbitrary matrices in both schemes. The uncorrected scheme uses $\widehat{W}_l^{\text{BASE}} = W_l + E_l$. QEP uses

$$W_l^*(\alpha_l) = W_l + \alpha_l W_l \delta_l \widehat{X}_l^\top \widehat{H}_l^{-1}, \quad \widehat{W}_l^{\text{QEP}} = W_l^*(\alpha_l) + E_l, \quad \alpha_l \in [0, 1].$$

For $M \in \{\text{BASE}, \text{QEP}\}$ define the residuals and the bound

$$R_l^M = W_l X_l - \widehat{W}_l^M \widehat{X}_l^M, \quad U^M = \sum_{k=1}^L \left(\prod_{s=k+1}^L \gamma_s \|W_s\|_2 \right) \gamma_k \|R_k^M\|_F.$$

Summary of the results. Several different comparisons arise, and their answers differ according to whether they are read exactly, to first order in a fixed shared perturbation, or in mean square for errors with zero conditional mean and a common second-moment map, of which subtractive dither is the main instance. Table 1 records what this paper establishes; “open” means not settled here.

Statement	Exact, shared E	First order, shared E	Mean square, zero conditional mean
$U^{\text{QEP}} \leq U^{\text{BASE}}$	fails (Thm. 17)	fails (Thm. 17)	fails for some error laws allowed by Thm. 8 (Sec. 4); open for uniform dither
Output error of QEP $\leq$ that of BASE	fails (Thm. 17); for deterministic code-books see Thm. 22	fails (Thm. 17)	linear networks: holds (Thm. 8). ReLU: fails in general (Ex. 12, 14); holds in the commuting case of Thm. 15, otherwise up to the excess bounded there
$\ \delta_{L+1}^{\text{QEP}}\ _F \leq U^{\text{BASE}}$	fails (Ex. 27)	holds, in the repaired form of App. B	holds for linear networks (Thm. 8) and in the commuting case of Thm. 15; otherwise open
Output error non-increasing in α	fails (Thm. 19)	fails (Thm. 19)	holds for two linear scalar layers (Prop. 7); fails between non-zero schedules for sign-valued errors (Sec. 4)
Layer residual on a common input, QEP $\leq$ BASE	for $\alpha > 0$, iff $2c \leq (2 - \alpha)d$ (Cor. 2)	same condition	holds for fixed-step dither on an unbounded grid (Thm. 4); can fail under overload (Sec. 3.3)

Table 1: What holds and what fails. The first four rows compare the two schemes on their own trajectories; the last row is a per-layer comparison on a common input. In the last column the bounds are compared through $\mathbb{E}(U^M)^2$, and entries established only for sign-valued errors say so. The network comparisons require the common deterministic second-moment map as well as zero conditional mean, and the additional hypotheses of the cited results.

3 One layer: when the correction helps

We begin with a single layer on a common input. An exact identity decides whether the correction helps for a given rounding error; taking expectations gives the conditions under which it helps on average, and subtractive dither supplies them. The network-level results of Sections 4 and 5 build on this section.

3.1 An exact identity

Fix one layer and one quantized input $\hat{X}$, and drop the layer index. Write

$$A = W\delta, \quad B = E\hat{X}, \quad P = \hat{X}^\top \hat{H}^{-1} \hat{X}, \quad R_\alpha = A(I - \alpha P) - B.$$

Thus R_0 is the BASE residual on this input and R_α the QEP residual with strength α . Let

$$d = \|AP\|_F^2, \quad c = \langle AP, B \rangle_F.$$

Proposition 1 (Residual identity). *Let A, B be real matrices of the same shape and P a symmetric idempotent matrix. For all real α, β ,*

$$\|R_\alpha\|_F^2 - \|R_0\|_F^2 = -\alpha(2 - \alpha)d + 2\alpha c, \tag{1}$$

$$\|R_\beta\|_F^2 - \|R_\alpha\|_F^2 = (\beta - \alpha)[(\alpha + \beta - 2)d + 2c]. \tag{2}$$

Proof. $R_\alpha = R_0 - \alpha AP$, so $\|R_\alpha\|_F^2 = \|R_0\|_F^2 - 2\alpha\langle R_0, AP \rangle_F + \alpha^2 d$. Because P is a symmetric idempotent, $\langle A, AP \rangle_F = \|AP\|_F^2 = d$, hence $\langle R_0, AP \rangle_F = d - c$. This gives (1); (2) follows by subtraction. $\square$

Corollary 2. *With the notation above:*

- (i) For $\alpha > 0$: $\|R_\alpha\|_F \leq \|R_0\|_F$ if and only if $2c \leq (2 - \alpha)d$.
- (ii) $\alpha \mapsto \|R_\alpha\|_F$ is non-increasing on $[0, 1]$ if and only if $c \leq 0$.
- (iii) $\|R_\alpha\|_F \leq \|R_0\|_F$ for all $\alpha \in [0, 1]$ if and only if $2c \leq d$.
- (iv) If $d > 0$, the minimizer of $\|R_\alpha\|_F$ over $[0, 1]$ is $\alpha^* = \text{clip}(1 - c/d, 0, 1)$.

Proof. (i) is (1) divided by $\alpha > 0$. For (ii), if $c \leq 0$ the bracket in (2) is non-positive whenever $\alpha + \beta \leq 2$. Conversely, if $c > 0$ then $AP \neq 0$, so $d > 0$; take $h = \min(1, c/d)$, and then $\|R_1\|_F^2 - \|R_{1-h}\|_F^2 = h(2c - hd) > 0$. (iii) follows from (i) since $(2 - \alpha)d \geq d$ on $[0, 1]$, with equality at $\alpha = 1$. (iv): $\|R_\alpha\|_F^2$ is a convex quadratic in α with vertex $1 - c/d$. $\square$

When $2c \leq d$ but $c > 0$, full correction is no worse than no correction, yet the residual first decreases and then increases as $\alpha \rightarrow 1$. “No worse at $\alpha = 1$ ” and “monotone in α ” are different properties.

Geometric reading. At $\alpha = 1$ the condition $2c \leq d$ is equivalent to $\|B - AP\|_F \geq \|B\|_F$. For $B = E\hat{X}$ we have $BP = B$, so B and AP lie in the same subspace, and the component $A(I - P)$ is common to both residuals. The correction is harmful exactly when the new perturbation term B is closer to the inherited term AP than to the origin, that is, when the two would largely cancel. In the counterexample of Theorem 17 in Section 6, $AP = -\varepsilon$ and $B = -\varepsilon - \varepsilon^2$.

Remark 3 (Scope of α^*). The formula for α^* holds B fixed. In an actual quantizer B depends on α through the rounding of $W^*(\alpha)$ (Section 7), so α^* is not by itself a rule for choosing the propagation strength.

3.2 In expectation

Theorem 4 (Per-layer mean-square statement). *Let A, B, P be random, with P almost surely a symmetric idempotent and A, B square integrable. Put $D = \mathbb{E}\|AP\|_F^2$ and $C = \mathbb{E}\langle AP, B \rangle_F$. For every fixed α ,*

$$\mathbb{E}\|R_\alpha\|_F^2 - \mathbb{E}\|R_0\|_F^2 = -\alpha(2 - \alpha)D + 2\alpha C.$$

In particular, if $\mathbb{E}[B \mid A, P] = 0$ then $C = 0$ and $\mathbb{E}\|R_\alpha\|_F^2 = \mathbb{E}\|R_0\|_F^2 - \alpha(2 - \alpha)D \leq \mathbb{E}\|R_0\|_F^2$ for $\alpha \in [0, 2]$, with $\mathbb{E}\|R_\alpha\|_F^2$ non-increasing on $[0, 1]$.

Proof. Take expectations in (1). Under the conditional-mean assumption, $C = \mathbb{E}\langle AP, \mathbb{E}[B \mid A, P] \rangle_F = 0$. $\square$

Remark 5 (A zero marginal mean is not enough). Let all quantities be scalars, $P = 1$ and $A = B = \pm 1$ with equal probability. Then $\mathbb{E}B = 0$, but $R_0 = 0$ and $R_\alpha = -\alpha A$, so $\mathbb{E}R_\alpha^2 = \alpha^2 > 0 = \mathbb{E}R_0^2$ for every $\alpha \neq 0$.

Remark 6 (The perturbation may depend on α). If the perturbation is B_α and $\mathbb{E}[B_\alpha \mid A, P] = 0$ for each α , the same computation gives

$$\mathbb{E}\|R_\alpha\|_F^2 - \mathbb{E}\|R_0\|_F^2 = -\alpha(2 - \alpha)D + \mathbb{E}\|B_\alpha\|_F^2 - \mathbb{E}\|B_0\|_F^2.$$

Zero conditional mean alone therefore does not suffice. Improvement holds exactly when $\mathbb{E}\|B_\alpha\|_F^2 - \mathbb{E}\|B_0\|_F^2 \leq \alpha(2 - \alpha)D$; a perturbation energy that does not grow with the correction is a sufficient condition.

3.3 Subtractive dither supplies both conditions

Scope. The statements in this subsection concern rounding to an unbounded uniform grid with a step size fixed in advance. They apply to a bounded codebook only when no target falls outside its range; when targets do fall outside, the per-layer statement can fail, as shown at the end of the subsection. Section 9 reports how often corrected weights leave the range of the original row on two small models; we have no estimate beyond those, and we do not attribute the unfavourable 2-bit cases reported in [1] to overload.

Let Q_Δ denote rounding to the nearest point of $\Delta\mathbb{Z}$. The subtractively dithered quantizer maps t to $Q_\Delta(t + u) - u$, where u is uniform on $[-\Delta/2, \Delta/2)$, independent of t , and known to the decoder. Its error $Q_\Delta(t + u) - u - t$ is uniform on $[-\Delta/2, \Delta/2)$ and independent of t [19, 20]; the vector analogue for lattices is in [21]. This holds for an unbounded grid; a bounded codebook requires in addition that no overload occurs.

Suppose each entry of $W_l^*(\alpha_l)$ is quantized in this way, with mutually independent fresh dithers that are independent of all earlier layers, and a step size that is fixed in advance and does not depend on the correction strength. Then E_l has independent entries, uniform with variance $\sigma^2 = \Delta^2/12$, and is independent of $(W_l\delta_l, \hat{X}_l)$, which are functions of earlier dither only. Hence $\mathbb{E}[E_l\hat{X}_l \mid W_l\delta_l, \hat{X}_l] = 0$, and $\mathbb{E}\|E_l\hat{X}_l\|_F^2 = \sigma^2 n_l \mathbb{E}\|\hat{X}_l\|_F^2$ does not depend on α_l , where n_l is the number of rows of W_l . Both conditions of Theorem 4 and Remark 6 hold, and on a common input

$$\mathbb{E}\|R_l^{\text{QEP}}\|_F^2 = \mathbb{E}\|\tilde{R}_l^{\text{BASE}}\|_F^2 - \alpha_l(2 - \alpha_l) \mathbb{E}\|W_l\delta_l P_l\|_F^2.$$

This is a statement about one layer on a common input. It does not by itself give a network-level comparison, because the two schemes produce different inputs for later layers. Section 4 treats the network level; we first compute a two-layer example exactly.

Proposition 7 (Two-layer linear network with dither). *Let $L = 2$, all matrices 1×1 , identity activations, teacher weights W_1, W_2 , input x , and subtractively dithered quantization with error variance σ^2 in both layers, with $\hat{X}_2 \neq 0$ almost surely. With α_1 arbitrary and $\alpha_2 = \alpha$, the mean-square output error of QEP is*

$$x^2\sigma^2[(1 - \alpha)^2W_2^2 + W_1^2 + \sigma^2].$$

It is non-increasing in α on $[0, 1]$, and strictly decreasing when $xW_2 \neq 0$; $\alpha = 0$ gives the uncorrected scheme, and $\alpha = 1$ removes the contribution $x^2\sigma^2W_2^2$ of the first-layer error.

Proof. Let r_1, r_2 be the quantization errors. They are independent, zero mean, with variance σ^2 , and r_2 is independent of r_1 whatever the second-layer target is. The first-layer correction vanishes, so $\hat{X}_2 = (W_1 + r_1)x$. The second-layer target is $W_2(1 - \alpha r_1/(W_1 + r_1))$, hence the output is $W_2W_1x + (1 - \alpha)W_2r_1x + r_2(W_1 + r_1)x$. The output error is $(1 - \alpha)W_2r_1x + r_2(W_1 + r_1)x$; its two terms are uncorrelated and $\mathbb{E}[r_2^2(W_1 + r_1)^2] = \sigma^2(W_1^2 + \sigma^2)$. $\square$

In a Monte Carlo check with $\Delta = 1$, two parameter settings, five values of α and 3×10^5 samples each, the estimates differ from this formula by at most 0.0017. This is a sanity check, not a proof.

Rounding with feedback. Methods such as GPTQ and QuIP round the entries of a weight row in sequence, adding to each target a fixed linear combination of the errors already made: $z_j = w_j + \sum_{i < j} (w_i - q_i) U_{ij}$, with U strictly upper triangular and determined by $\hat{X}$ [4, 5]. Suppose each q_j is obtained from z_j by subtractive dither on an unbounded grid, with fresh independent dither and fixed step sizes. Each innovation $n_j = q_j - z_j$ is then uniform and independent of the past, including its own target z_j , and the row error $e = q - w$ satisfies $e(I + U) = n$, that is, $e = n(I + U)^{-1}$. Hence $\mathbb{E}[E\hat{X} \mid W\delta, \hat{X}] = 0$, and $\mathbb{E}\|E\hat{X}\|_F^2$ does not depend on α provided U , the column order and the step sizes are not re-estimated when α changes. The per-layer statement therefore extends to dithered rounding with fixed feedback; the entries of E are now correlated, which the statement does not require them not to be. It does not extend to ordinary stochastic rounding, whose error variance depends on the position of the target within its cell.

Bounded codebooks. For a bounded uniform codebook $\{a, a + \Delta, \dots, b\}$ with nearest-point saturation, the dithered reconstruction of a target t is distributed as $\text{clip}(t, a, b) + V$ with V uniform on $[-\Delta/2, \Delta/2)$. Overload thus introduces a bias $\text{clip}(t, a, b) - t$ and leaves the variance unchanged, and the bias can break the per-layer statement. Take the codebook $\{0, 1, 2, 3\}$ with $\Delta = 1$ and

$$Z = \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{2} \end{pmatrix}, \quad W = \begin{pmatrix} 4 & 2 \\ 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} 4 & 1 \\ 0 & 1 \end{pmatrix}, \quad X = W^{-1}TZ,$$

so that $\hat{X} = Z$, $P = I$ and $W^*(\alpha) = W + \alpha(T - W)$. With entrywise dithered rounding,

$$\mathbb{E}\|WX - \widehat{W}(\alpha)Z\|_F^2 = \frac{5}{8} - \frac{\alpha}{2} + \frac{5}{4}\alpha^2,$$

which equals $\frac{5}{8}$ at $\alpha = 0$ and $\frac{11}{8}$ at $\alpha = 1$; the minimum is at $\alpha = \frac{1}{5}$. The entry 4 is clipped to 3 at every α , and correcting the neighbouring entry removes a term that partly cancelled this bias through the correlated inputs. Overload of the uncorrected targets is not needed: with the same T , $W = \begin{pmatrix} 2 & 2 \\ 0 & 1 \end{pmatrix}$, $Z = \begin{pmatrix} 1 & 0 \\ 2 & 1/2 \end{pmatrix}$ and $X = W^{-1}TZ$, every entry of W lies in the codebook, only the corrected entry $2 + 2\alpha$ overloads, for $\alpha > \frac{1}{2}$, and the loss $\frac{7}{8} + \frac{1}{4}(1 - \alpha)^2 + \max(0, 2\alpha - 1)^2$ falls to $\frac{15}{16}$ at $\alpha = \frac{1}{2}$ and then rises to $\frac{15}{8}$ at $\alpha = 1$, above its value $\frac{9}{8}$ at $\alpha = 0$. These examples use entrywise rounding with a non-diagonal input Gram matrix; they do not contradict Remark 21, which concerns exact minimization over the codebook.

Cost. The reconstruction $Q_\Delta(t + u) - u$ is not a codebook point. The decoder must regenerate the dither, for instance from a seed, and subtract it. There is also an accuracy cost. On an unbounded grid the dithered error has mean square $\Delta^2/12$ whatever the input. For inputs that vary smoothly on the scale of Δ , deterministic rounding has the same mean-square error; on a coarse grid, deterministic rounding of concentrated inputs can have a smaller one. We have not measured either cost. The diagnostic of Section 9 includes dithered rounding of single layers; we have not evaluated dithered propagation end to end on any trained model; Section 8 does so on small synthetic networks.

4 Linear networks

The per-layer statements compare the two schemes on a common input. In a network the two schemes produce different inputs from the second layer on, and each must be followed along its own trajectory. For linear networks the mean-square comparison holds for the whole network.

Theorem 8 (Linear networks). *Let all activations be the identity and let the strengths $\alpha_l \in [0, 1]$ be fixed in advance. Let $\mathcal{F}_l$ denote the history before layer l is quantized. Suppose that, for both schemes and every l ,*

$$\mathbb{E}[E_l \mid \mathcal{F}_l] = 0, \quad \mathbb{E}[E_l T E_l^\top \mid \mathcal{F}_l] = \Psi_l(T)$$

for every bounded $\mathcal{F}_l$ -measurable positive semidefinite T , where Ψ_l is a deterministic positive linear map, the same for both schemes. The network, the input X_1 and the depth are fixed. Suppose also that, for every l with $\alpha_l > 0$, the Gram matrix $\hat{H}_l$ along the QEP trajectory is almost surely invertible. Write $C_l^M = \mathbb{E}[\delta_l^M (\delta_l^M)^\top]$ and $G_l^M = \mathbb{E}[\hat{X}_l^M (\hat{X}_l^M)^\top]$. Then these moments are finite, and for every l ,

$$C_l^{\text{QEP}} \preceq C_l^{\text{BASE}}, \quad G_l^{\text{QEP}} \preceq G_l^{\text{BASE}},$$

and in particular $\mathbb{E}\|\delta_{L+1}^{\text{QEP}}\|_F^2 \leq \mathbb{E}\|\delta_{L+1}^{\text{BASE}}\|_F^2$.

Entrywise subtractive dither on an unbounded grid, with independent entries and a step fixed in advance, satisfies the hypotheses with $\Psi_l(T) = \sigma_l^2 \text{tr}(T) I$; with entry-dependent steps, Ψ_l is the corresponding fixed weighted trace map. The weight matrices may be rectangular, subject to the rank restriction discussed after the proof.

Proof. Fix a layer of the QEP trajectory and write $Z = \hat{X}_l$, $X = X_l$, $P = P_l$, $W = W_l$, $\alpha = \alpha_l$. If $\alpha = 0$ the correction is skipped, $T = Z$, and the two identities below hold trivially. Otherwise, since $ZP = Z$, the next pre-activation is $\hat{W}Z = WT + E_l Z$ with $T = (1 - \alpha)Z + \alpha XP$, and $T - X = (Z - X)(I - \alpha P)$. Two pointwise identities follow from $P = P^\top = P^2$:

$$(Z - X)(Z - X)^\top - (T - X)(T - X)^\top = \alpha(2 - \alpha)(Z - X)P(Z - X)^\top \succeq 0, \quad (3)$$

$$(1 - \alpha)ZZ^\top + \alpha XX^\top - TT^\top = \alpha X(I - P)X^\top + \alpha(1 - \alpha)(Z - XP)(Z - XP)^\top \succeq 0. \quad (4)$$

The second-moment relation extends from bounded T to every $\mathcal{F}_l$ -measurable positive semidefinite T , as an identity in $[0, \infty]$, by truncating T on $\{\|T\| \leq n\}$ and monotone convergence. Taking the trace in (4) gives $\|T\|_F^2 \leq (1 - \alpha)\|Z\|_F^2 + \alpha\|X\|_F^2$, and with $\kappa_l = \|\Psi_l^*(I)\|_2$ we have $\mathbb{E}[\|E_l Z\|_F^2 \mid \mathcal{F}_l] = \text{tr} \Psi_l(ZZ^\top) \leq \kappa_l \|Z\|_F^2$. Hence $\mathbb{E}\|\hat{X}_{l+1}\|_F^2 \leq 2\|W\|_2^2[(1 - \alpha)\mathbb{E}\|\hat{X}_l\|_F^2 + \alpha\|X_l\|_F^2] + 2\kappa_l \mathbb{E}\|\hat{X}_l\|_F^2$, so all second moments of the activations and of the errors are finite by induction, for both schemes. The corrected weight itself need not be square integrable, since it contains $\hat{H}_l^{-1}$; the output involves only the projection P .

The two terms WT and $E_l Z$ are now square integrable, so the conditional-mean assumption removes their cross terms, and

$$C_{l+1}^{\text{QEP}} \preceq W C_l^{\text{QEP}} W^\top + \Psi_l(G_l^{\text{QEP}}), \quad G_{l+1}^{\text{QEP}} \preceq W[(1 - \alpha)G_l^{\text{QEP}} + \alpha X_l X_l^\top] W^\top + \Psi_l(G_l^{\text{QEP}}),$$

using (3) and (4). For BASE, $\hat{X}_{l+1} = (W + E_l)\hat{X}_l$ gives the exact recursions $C_{l+1}^{\text{BASE}} = W C_l^{\text{BASE}} W^\top + \Psi_l(G_l^{\text{BASE}})$ and $G_{l+1}^{\text{BASE}} = W G_l^{\text{BASE}} W^\top + \Psi_l(G_l^{\text{BASE}})$. BASE is unbiased, $\mathbb{E}\hat{X}_l^{\text{BASE}} = X_l$, hence $G_l^{\text{BASE}} = X_l X_l^\top + C_l^{\text{BASE}} \succeq X_l X_l^\top$.

Both orderings hold with equality at $l = 1$. If they hold at l , then, Ψ_l being positive, $C_{l+1}^{\text{QEP}} \preceq W C_l^{\text{BASE}} W^\top + \Psi_l(G_l^{\text{BASE}}) = C_{l+1}^{\text{BASE}}$, and, using $X_l X_l^\top \preceq G_l^{\text{BASE}}$, $G_{l+1}^{\text{QEP}} \preceq W G_l^{\text{BASE}} W^\top + \Psi_l(G_l^{\text{BASE}}) = G_{l+1}^{\text{BASE}}$. Taking traces gives the last claim. $\square$

The induction carries the pair (C_l, G_l) ; the uncentred moment is needed because the new perturbation enters multiplied by the input. No inclusion between the row spaces of X_l and $\hat{X}_l$ is assumed: the term $X(I - P)X^\top$ in (4) is exactly what the missing inclusion costs, and it has the

right sign. At a layer of positive strength, invertibility of $\widehat{H}_l$ requires $\widehat{X}_l$ to have full row rank; layers of zero strength are not restricted. A linear map cannot increase rank, so a layer whose output feeds a later corrected layer may be narrower than its predecessors but not wider than the rank of its input; the last layer is not restricted. If $\widehat{H}_l^{-1}$ is replaced by the Moore–Penrose pseudo-inverse, P_l is still the orthogonal projection onto the row space of $\widehat{X}_l$, the proof is unchanged and no restriction on widths remains; that is a different algorithm from the one in [1].

The two inequalities in the proof discard non-negative terms. Keeping them gives the gain exactly.

Proposition 9 (Exact gain). *Assume the hypotheses of Theorem 8. Along the QEP trajectory put, for layers with $\alpha_l > 0$,*

$$\begin{aligned} S_l &= \alpha_l(2 - \alpha_l) W_l \mathbb{E}[\delta_l P_l \delta_l^\top] W_l^\top, \\ B_l &= W_l \mathbb{E}[\alpha_l X_l(I - P_l)X_l^\top + \alpha_l(1 - \alpha_l)(\widehat{X}_l - X_l P_l)(\widehat{X}_l - X_l P_l)^\top] W_l^\top, \end{aligned}$$

and $S_l = B_l = 0$ when $\alpha_l = 0$. Then $\Delta_l = C_l^{\text{BASE}} - C_l^{\text{QEP}}$ and $\Lambda_l = G_l^{\text{BASE}} - G_l^{\text{QEP}}$ satisfy $\Delta_1 = \Lambda_1 = 0$ and

$$\Delta_{l+1} = W_l \Delta_l W_l^\top + \Psi_l(\Lambda_l) + S_l, \quad (5)$$

$$\Lambda_{l+1} = (1 - \alpha_l) W_l \Lambda_l W_l^\top + \Psi_l(\Lambda_l) + \alpha_l W_l C_l^{\text{BASE}} W_l^\top + B_l, \quad (6)$$

in which every term is positive semidefinite. In particular

$$\mathbb{E}\|\delta_{L+1}^{\text{BASE}}\|_F^2 - \mathbb{E}\|\delta_{L+1}^{\text{QEP}}\|_F^2 \geq \sum_{l: \alpha_l > 0} \alpha_l(2 - \alpha_l) \mathbb{E}\|W_L \cdots W_l \delta_l^{\text{QEP}} P_l\|_F^2, \quad (7)$$

where the sum runs over the layers of positive strength only, the projection P_l not being defined elsewhere; the comparison of Theorem 8 is strict as soon as one term on the right is non-zero.

Proof. The relations (3) and (4) are identities. Keeping their right-hand sides in the proof of Theorem 8 turns the two inequalities for QEP into $C_{l+1}^{\text{QEP}} = W_l C_l^{\text{QEP}} W_l^\top + \Psi_l(G_l^{\text{QEP}}) - S_l$ and $G_{l+1}^{\text{QEP}} = W_l[(1 - \alpha_l)G_l^{\text{QEP}} + \alpha_l X_l X_l^\top] W_l^\top + \Psi_l(G_l^{\text{QEP}}) - B_l$; the expectations are finite because P_l is a projection, and P_l is not separated from δ_l . Subtracting from the exact recursions for BASE gives (5) at once, and (6) after using $G_l^{\text{BASE}} = X_l X_l^\top + C_l^{\text{BASE}}$, which yields $G_l^{\text{BASE}} - (1 - \alpha_l)G_l^{\text{QEP}} - \alpha_l X_l X_l^\top = (1 - \alpha_l)\Lambda_l + \alpha_l C_l^{\text{BASE}}$. Positivity of Δ_l and Λ_l is Theorem 8. Dropping $\Psi_l(\Lambda_l) \succeq 0$ in (5) and iterating gives $\Delta_{L+1} \succeq \sum_l W_L \cdots W_{l+1} S_l W_{l+1}^\top \cdots W_L^\top$, whose trace is the right-hand side of (7), the terms with $\alpha_l = 0$ being zero. $\square$

The two recursions separate two effects. S_l is the error removed at layer l . The source in (6) is a reduction of the energy that excites later rounding errors; it reaches the output only through the maps Ψ of later layers, and is not seen by a per-layer comparison. Since the recursions are linear with positive coefficients, the gain is a sum of non-negative contributions of the sources S_l and $\alpha_l W_l C_l^{\text{BASE}} W_l^\top + B_l$, each weighted by a positive semidefinite matrix that depends only on the teacher, the maps Ψ_l and the strengths; the bound (7) keeps the first kind only. A gain at an intermediate layer need not reach the output: later weights and noise maps may ignore the directions in which it occurs.

Corollary 10 (Schedules with a common row space). *Assume the hypotheses of Theorem 8 and let a and b be two schedules fixed in advance with $0 \leq a_l \leq b_l \leq 1$. Suppose that along both trajectories $X_l P_l = X_l$ almost surely at every layer of positive strength; this is the case when X_1 has full row*

rank and all teacher and quantized weight matrices are square and invertible. Then $C_l^b \preceq C_l^a$ and $G_l^b \preceq G_l^a$ for every l . For each such schedule,

$$C_{l+1} = (1 - \alpha_l)^2 W_l C_l W_l^\top + \Psi_l(C_l) + \Psi_l(X_l X_l^\top).$$

In any class of fixed schedules that satisfy these hypotheses with the same maps Ψ_l and that contains the full-strength schedule, full strength minimizes the mean-square output error.

Proof. With $X_l P_l = X_l$ the target is $T = (1 - \alpha_l) \hat{X}_l + \alpha_l X_l$, so $\delta_{l+1} = (1 - \alpha_l) W_l \delta_l - E_l \hat{X}_l$. By induction every schedule is unbiased, $\mathbb{E} \delta_l = 0$, hence $G_l = X_l X_l^\top + C_l$ and the displayed recursion holds. If $C_l^b \preceq C_l^a$, positivity of the maps and $(1 - b_l)^2 \leq (1 - a_l)^2$ give $C_{l+1}^b \preceq C_{l+1}^a$. $\square$

Example 11 (A closed form for the gain). Fix a calibration input X_1 with d rows and full row rank, and a schedule. Let all layers be square of size d and let the teacher weights be independent random matrices, almost surely invertible, with $\mathbb{E}[W_l^\top W_l] = \gamma_l I$ for deterministic constants $0 < \gamma_l < \infty$; for instance $W_l = I + V_l$ with the entries of V_l independent and uniform on $[-a_l, a_l]$, $a_l < 1/(2d)$, where $\gamma_l = 1 + da_l^2/3$. Suppose that, conditionally on all teacher weights and on the past, E_l has zero mean and $\mathbb{E}[E_l M E_l^\top] = \nu_l \text{tr}(M) I$ with deterministic constants $0 \leq \nu_l < \infty$ common to the two schemes, and that the quantized weights are almost surely invertible. Put $q_l = \mathbb{E} \|\delta_{l+1}\|_F^2 / \mathbb{E} \|X_{l+1}\|_F^2$, with expectations over the teacher as well, and $\nu_l = d \nu_l / \gamma_l$. Taking traces and the expectation over the teacher in Corollary 10 gives $q_l = [(1 - \alpha_l)^2 + \nu_l] q_{l-1} + \nu_l$ with $q_0 = 0$, hence

$$q_L^{\text{BASE}} = \prod_{l=1}^L (1 + \nu_l) - 1, \quad q_L^{\text{QEP}} = \sum_{j=1}^L \prod_{i=j}^L \nu_i \quad \text{at full strength.}$$

For $\nu_l = \nu > 0$ the ratio $q_L^{\text{BASE}}/q_L^{\text{QEP}} = \sum_{k < L} (1 + \nu)^k / \sum_{k < L} \nu^k$ increases with the depth and tends to L as $\nu \downarrow 0$ (for $\nu = 0$ both errors vanish): at fine resolution full compensation leaves the rounding noise of the last layer only, a gain of $10 \log_{10} L$ dB. This is an average over random teachers and a ratio of expectations, in a class where all row spaces coincide; it is not a statement about a fixed network or about bottlenecks.

Three limits should be stated. First, Theorem 8 compares any fixed schedule with the uncorrected scheme. Outside the setting of Corollary 10, monotonicity between two non-zero schedules can fail; the condition of that corollary is empty for the zero schedule and so does not exclude this. Take $X = I_4$, $W_1 = [I_2 \mid 0]$, $W_2 = (1, 0)$, $W_3 = 1$, with errors $E_1 = [0 \mid \text{diag}(u_1, u_2)]$, $E_2 = (0, \eta)$, $E_3 = 0$, where $u_1, u_2 = \pm 1$ and $\eta = \pm \frac{1}{2}$ are independent fair signs, and $\alpha_3 = 1$. The mean-square output error is $\frac{3}{2}$ without correction, $\frac{17}{26}$ at $\alpha_2 = \frac{1}{2}$ and $\frac{3}{4}$ at $\alpha_2 = 1$. Second, the theorem does not cover feedback coefficients recomputed along each trajectory, which change Ψ_l ; strengths chosen after seeing the rounding, which break the fixed schedule and the conditional mean; or overload, which introduces a conditional bias. Third, it does not extend to the bounds U^M . For three scalar linear layers with unit weights and input, strengths $(0, 1, 0)$, and independent errors $\pm t, \pm 4, \pm \frac{4}{3}$ with fair signs and $0 < t < \frac{6}{67}$, one finds $\mathbb{E}(U^{\text{QEP}})^2 - \mathbb{E}(U^{\text{BASE}})^2 = t(6 - 67t)/9 > 0$. These errors satisfy the hypotheses of the theorem but are not a uniform dither; for uniform dither the comparison of the bounds is not settled here.

5 ReLU networks: two obstructions and a conditional guarantee

The comparison of Theorem 8 does not survive a nonlinearity. We give the two ways in which it fails and a condition under which it is recovered up to an explicit excess.

Example 12 (The mean-square comparison fails for ReLU). Let $L = 2$, all quantities scalar, both activations ReLU, $X = 1$, $W_1 = 1$, $W_2 = -\frac{1}{4}$, and let both layers use subtractive dither on $\mathbb{Z}$, with errors u, v independent and uniform on $[-\frac{1}{2}, \frac{1}{2})$. The teacher output is 0. With $z = 1 + u \in [\frac{1}{2}, \frac{3}{2})$, the uncorrected output is $z(v - \frac{1}{4})_+$; at $\alpha_2 = 1$ the QEP target is $-1/(4z)$ and the QEP output is $(vz - \frac{1}{4})_+$. Then

$$\mathbb{E}[(\hat{X}_3^{\text{BASE}})^2] = \frac{13}{2304} \approx 0.00564, \quad \mathbb{E}[(\hat{X}_3^{\text{QEP}})^2] = \frac{1}{72} - \frac{\log 3}{192} \approx 0.00817.$$

No target is out of range and all Gram matrices are positive. Given the first layer, the correction restores the conditional mean of the pre-activation to its teacher value $-\frac{1}{4}$ (without it the conditional mean is $-z/4$). The conditional variance, $z^2/12$, is the same with and without the correction; the unconditional variance of the pre-activation drops from $\frac{55}{576}$ to $\frac{13}{144}$. The teacher's ReLU is inactive there, so what reaches the output is the upper tail of the noise. The corrected output is positive less often, with probability $\frac{1}{2} - \frac{\log 3}{4} < \frac{1}{4}$, but its positive tail carries more energy. A zero conditional mean before the activation does not survive the activation.

Example 12 has noise crossing the activation boundary. That is not the only obstruction. A unit with a fixed 0/1 activation pattern across the calibration samples reads its pre-activation error through a diagonal mask D , while the correction mixes samples through P . The next proposition says when the correction is compatible with the mask for every possible inherited error.

Proposition 13 (Masked contraction). *Let P be an orthogonal projection, D a diagonal matrix with entries in $\{0, 1\}$, and $0 < \alpha \leq 1$. Then $\|a(I - \alpha P)D\|_2 \leq \|aD\|_2$ for every row vector a if and only if $PD = DP$.*

Proof. If $PD = DP$ then $a(I - \alpha P)D = aD(I - \alpha P)$ and $I - \alpha P$ is a contraction. Conversely, for every a with $aD = 0$ the inequality forces $aPD = 0$, that is, $(I - D)PD = 0$; transposing, $DP(I - D) = 0$, hence $PD = DPD = DP$. $\square$

The proof uses only that D is a symmetric idempotent. The condition characterizes contraction for all inherited errors; in a given network the inherited error is constrained by the teacher and the input, so commutation is not necessary for the mean-square comparison in that network. The same computation gives a weighted form: for any symmetric positive semidefinite M , the inequality $r(I - \alpha P)M(I - \alpha P)r^\top \leq rMr^\top$ holds for every row vector r if and only if $PM = MP$; in an orthonormal basis adapted to P , with $M = \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{pmatrix}$, the difference $M - (I - \alpha P)M(I - \alpha P)$ has blocks $\alpha(2 - \alpha)M_{11}$, αM_{12} and 0, and is positive semidefinite exactly when $M_{12} = 0$. The activation mask is one instance of a metric on the sample space.

The defect of commutation can be computed from the calibration data: $\|PD - DP\|_F^2 = 2 \sum_{i: d_i=1} \sum_{j: d_j=0} P_{ij}^2$, the squared coupling, through the fit, between samples on which the unit is active and samples on which it is inactive. Here P acts on the index space of the calibration samples; the coupling comes from fitting one weight matrix to all samples, not from any mixing of inputs at inference. Theorem 15 below uses the spectral norm of the commutator, which this Frobenius norm bounds from above.

Example 14 (Failure without any change of activation). Let $L = 2$ with both activations ReLU,

$$X = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \quad W_1 = I_2, \quad W_2 = \begin{pmatrix} 1 & -2 \\ -1 & -1 \end{pmatrix},$$

and let every weight of both layers be quantized with subtractive dither of the same step, with errors independent and uniform on $[-t, t]$; $\alpha_2 = 1$. The teacher pre-activations of the second layer

are bounded away from zero by 1. For $0 < t \leq \frac{1}{10}$, in both schemes and for every realization of the errors, every unit has the teacher's activation state on every sample, and all Gram matrices and quantized weights are invertible. Nevertheless

$$\mathbb{E}\|\delta_3^{\text{QEP}}\|_F^2 - \mathbb{E}\|\delta_3^{\text{BASE}}\|_F^2 \geq t^2 \left(\frac{2}{27} - 42t(1+t) \right) > 0 \quad \text{for } 0 < t \leq \frac{1}{600},$$

and the difference equals $\frac{2}{27}t^2 + O(t^3)$; the proof is in Appendix A. Only the first output unit on the first sample is active. On that entry, with h_{ij} the normalized first-layer errors, the error inherited from the first layer is $t(h_{11} - h_{12})$ to first order without correction and $t(-2h_{12} + 4h_{21})/3$ with it; the second-layer error contributes the same term to both. The correction has moved error from samples where the unit is closed onto the sample where it is open. Here P does not commute with the activation masks.

When the projections respect the activation patterns the linear argument goes through, and when they nearly do the comparison holds up to an explicit cost.

Partition the calibration samples into cells, with J_c the diagonal 0/1 projection onto cell c , such that within each cell all samples have the same teacher activation pattern at every layer. Let $D_{l,c}$ be the diagonal 0/1 pattern of cell c at the output of layer l , and put $F_{l,c} = D_{l,c}W_l$ and $\Psi_{l,c}(T) = D_{l,c}\Psi_l(T)D_{l,c}$. Along the QEP trajectory let $k_{l,c} = \|P_l J_c - J_c P_l\|_2$ if $l \geq 2$ and $\alpha_l > 0$, and $k_{l,c} = 0$ otherwise. Define deterministic positive semidefinite matrices by $K_{1,c} = V_{1,c} = 0$ and

$$K_{l+1,c} = F_{l,c}K_{l,c}F_{l,c}^\top + \Psi_{l,c}(V_{l,c}) + \alpha_l F_{l,c} \mathbb{E}[k_{l,c} \delta_l^{\text{QEP}} (\delta_l^{\text{QEP}})^\top] F_{l,c}^\top, \quad (8)$$

$$V_{l+1,c} = (1 - \alpha_l)F_{l,c}V_{l,c}F_{l,c}^\top + \Psi_{l,c}(V_{l,c}) + \alpha_l \mathbb{E}[k_{l,c}] F_{l,c}X_l X_l^\top F_{l,c}^\top. \quad (9)$$

Theorem 15 (ReLU networks with stable activations). *Let all activations be ReLU and let the strengths be fixed in advance. Assume:*

- (a) *almost surely both quantized trajectories have the teacher's activation pattern at every layer and on every sample;*
- (b) *the hypotheses of Theorem 8 on E_l and $\hat{H}_l$.*

Then, for every l and c ,

$$\begin{aligned} \mathbb{E}[\delta_l^{\text{QEP}} J_c (\delta_l^{\text{QEP}})^\top] &\preceq \mathbb{E}[\delta_l^{\text{BASE}} J_c (\delta_l^{\text{BASE}})^\top] + K_{l,c}, \\ \mathbb{E}[\hat{X}_l^{\text{QEP}} J_c (\hat{X}_l^{\text{QEP}})^\top] &\preceq \mathbb{E}[\hat{X}_l^{\text{BASE}} J_c (\hat{X}_l^{\text{BASE}})^\top] + V_{l,c}, \end{aligned}$$

and therefore

$$\mathbb{E}\|\delta_{L+1}^{\text{QEP}}\|_F^2 \leq \mathbb{E}\|\delta_{L+1}^{\text{BASE}}\|_F^2 + \sum_c \text{tr } K_{L+1,c}.$$

In particular, if $P_l J_c = J_c P_l$ almost surely for every c and every $l \geq 2$ with $\alpha_l > 0$, then $\mathbb{E}\|\delta_{L+1}^{\text{QEP}}\|_F^2 \leq \mathbb{E}\|\delta_{L+1}^{\text{BASE}}\|_F^2$.

Proof. Let P and J be orthogonal projections, $0 \leq \alpha \leq 1$ and $k = \|PJ - JP\|_2$. Then

$$(I - \alpha P)J(I - \alpha P) \preceq J + \alpha k I, \quad PJP \preceq J + k I. \quad (10)$$

Indeed, write $x = p + q$ with $p = Px$. Since $PJ(I - P) = P(PJ - JP)(I - P)$, we have $|p^\top Jq| \leq k\|p\|\|q\| \leq \frac{1}{2}k\|x\|^2$, and the two quadratic forms to be bounded equal $-\alpha(2 - \alpha)p^\top Jp - 2\alpha p^\top Jq$ and $-q^\top Jq - 2p^\top Jq$.

Fix a layer and a cell, with the notation of the proof of Theorem 8 and $J = J_c$, $k = k_{l,c}$. From $T - X = (Z - X)(I - \alpha P)$ and (10),

$$(T - X)J(T - X)^\top \preceq (Z - X)J(Z - X)^\top + \alpha k (Z - X)(Z - X)^\top.$$

Since $(1 - \alpha)ZJZ^\top + \alpha XPJPX^\top - TJT^\top = \alpha(1 - \alpha)(Z - XP)J(Z - XP)^\top \succeq 0$, (10) also gives

$$TJT^\top \preceq (1 - \alpha)ZJZ^\top + \alpha XJX^\top + \alpha k XX^\top.$$

At the first layer and at layers with $\alpha = 0$ we have $T = Z$ and both hold with $k = 0$. No Gram matrix of ZJ_c is inverted.

By (a), on the samples of cell c the next activations are $F_{l,c}T + D_{l,c}E_lZ$ for QEP, $F_{l,c}Z + D_{l,c}E_lZ$ for BASE and $F_{l,c}X$ for the teacher. Write $C_{l,c}^M$ and $G_{l,c}^M$ for the two cell moments in the statement. Integrability is obtained as in Theorem 8, since ReLU does not increase norms. The conditional-mean assumption removes the cross terms, so

$$\begin{aligned} C_{l+1,c}^{\text{QEP}} &\preceq F_{l,c}C_{l,c}^{\text{QEP}}F_{l,c}^\top + \Psi_{l,c}(G_{l,c}^{\text{QEP}}) + \alpha_l F_{l,c}\mathbb{E}[k_{l,c}\delta_l^{\text{QEP}}(\delta_l^{\text{QEP}})^\top]F_{l,c}^\top, \\ G_{l+1,c}^{\text{QEP}} &\preceq F_{l,c}[(1 - \alpha_l)G_{l,c}^{\text{QEP}} + \alpha_l X_l J_c X_l^\top]F_{l,c}^\top + \Psi_{l,c}(G_{l,c}^{\text{QEP}}) + \alpha_l \mathbb{E}[k_{l,c}]F_{l,c}X_l X_l^\top F_{l,c}^\top. \end{aligned}$$

The commutator norm $k_{l,c}$ is random and correlated with δ_l^{QEP} , so the expectation of their product is not factored. For BASE the recursions are exact, $C_{l+1,c}^{\text{BASE}} = F_{l,c}C_{l,c}^{\text{BASE}}F_{l,c}^\top + \Psi_{l,c}(G_{l,c}^{\text{BASE}})$ and $G_{l+1,c}^{\text{BASE}} = F_{l,c}G_{l,c}^{\text{BASE}}F_{l,c}^\top + \Psi_{l,c}(G_{l,c}^{\text{BASE}})$, and by (a) BASE is unbiased on the cell, so $G_{l,c}^{\text{BASE}} \succeq X_l J_c X_l^\top$. Both claims hold with equality at $l = 1$. If they hold at l , inserting them in the two inequalities and using the positivity of $\Psi_{l,c}$ gives them at $l + 1$, with (8) and (9). The squared Frobenius error is the sum over cells of $\text{tr } C_{L+1,c}$. $\square$

Condition (a) must hold almost surely, not on an event selected afterwards, since conditioning on the absence of activation changes does not in general preserve the zero conditional mean; it can be certified from the margins of the teacher pre-activations and a bound on the errors. Commutation says that the correction does not project samples of one activation pattern onto another. Stability and commutation are independent: with a single sample commutation is automatic, and Example 12 still fails because (a) does; Example 14 satisfies (a) and has non-commuting projections.

Exact commutation is restrictive. Together with the invertibility of $\hat{H}_l$ it is equivalent to $\sum_c \text{rank}(\hat{X}_l J_c) = d_l$, where d_l is the number of rows of $\hat{X}_l$: the activations of the different cells must span complementary subspaces, so the number of cells with non-zero activations at that layer is at most d_l . It holds when all samples share one pattern, and it fails as soon as two cells have non-zero activations and one of them spans the whole space. It is not vacuous. With

$$X = \begin{pmatrix} 1 & 2 & -1 & -1 \\ -1 & -1 & 1 & 2 \end{pmatrix}, \quad W_1 = I_2, \quad W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad W_3 = \begin{pmatrix} 1 & -3 \\ 1 & 1 \end{pmatrix},$$

and cells $\{1, 2\}$ and $\{3, 4\}$, the two cells have different activation patterns at every layer, and all hypotheses hold with commuting projections for independent errors $\pm \frac{1}{20}$ on every weight. Exact enumeration of the 4096 realizations gives mean-square output error $3287607/16000000 \approx 0.2055$ without correction and ≈ 0.0532 with $\alpha = (0, 1, 1)$.

The defect term makes the commuting case stable under perturbation. If Z_0 has full row rank with smallest singular value at least s , P_0 is its row-space projection and $\|Z - Z_0\|_2 \leq \eta < s$, then the row-space projection P of Z satisfies $\|P - P_0\|_2 \leq 2\eta/(s - \eta)$, and hence $\|PJ - JP\|_2 \leq \|P_0J - JP_0\|_2 + 2\|P - P_0\|_2$ for every projection J . A lower bound on the singular values is needed: the row-space projection of $\begin{pmatrix} 1 & 0 & 0 \\ 0 & \eta & 0 \end{pmatrix}$ has commutator norm $\frac{1}{2}$ with $\text{diag}(1, 1, 0)$ for every $\eta > 0$.

Corollary 16 (Small errors). *Fix a ReLU network, the calibration input, the cells and the strengths. Suppose that all teacher pre-activations are non-zero, that X_l has full row rank whenever $\alpha_l > 0$, and that the teacher projections P_l^0 satisfy $\|P_l^0 J_c - J_c P_l^0\|_2 \leq \varepsilon_0$ for $l \geq 2$ with $\alpha_l > 0$. Let $E_l = tN_l$ with $\|N_l\|_2$ bounded, $\mathbb{E}[N_l | \mathcal{F}_l] = 0$ and $\mathbb{E}[N_l T N_l^\top | \mathcal{F}_l] = \Psi_l^0(T)$ for a deterministic positive linear map common to both schemes. Then there are $t_0 > 0$ and a constant κ , depending on the network, the input and the bound on N_l , such that for $0 < t < t_0$*

$$\mathbb{E}\|\delta_{L+1}^{\text{QEP}}\|_F^2 \leq \mathbb{E}\|\delta_{L+1}^{\text{BASE}}\|_F^2 + \kappa(\varepsilon_0 t^2 + t^3).$$

Proof. For both schemes $\|\delta_{l+1}\|_F \leq (\|W_l\|_2 + t\|N_l\|_2)\|\delta_l\|_F + t\|N_l\|_2\|X_l\|_F$ on every realization, since the entrywise ReLU is 1-Lipschitz in the Frobenius norm and $I - \alpha P$ is a contraction; hence $\|\delta_l\|_F = O(t)$ uniformly, and $\|\delta_l\|_2 \leq \|\delta_l\|_F$. For small t the margins of the teacher give (a), the quantized activations entering layers of positive strength have full row rank with singular values bounded below, and by the perturbation bound above $k_{l,c} \leq \varepsilon_0 + O(t)$. In (9) the source is $O(\varepsilon_0 + t)$, so $V_{l,c} = O(\varepsilon_0 + t)$. In (8) the source is $O((\varepsilon_0 + t)t^2)$ and $\Psi_{l,c}(V_{l,c}) = t^2 D_{l,c} \Psi_l^0(V_{l,c}) D_{l,c} = O((\varepsilon_0 + t)t^2)$. The depth is fixed, so $\text{tr } K_{L+1,c} = O(\varepsilon_0 t^2 + t^3)$. $\square$

When the teacher projections commute exactly, the excess is $O(t^3)$, an upper bound, while the errors themselves are in general of order t^2 , although the projections along the quantized trajectory do not commute in general. An instance: X_1 with rows $(2, 2, 1, 1)$, $(1, 1, 2, 2)$, $(0, 0, \eta, -\eta)$, $W_1 = [I_2 \mid 0]$, $W_2 = \begin{pmatrix} 1 & -1 \\ -1 & 3/2 \end{pmatrix}$, cells $\{1, 2\}$ and $\{3, 4\}$, $\alpha = (0, 1)$, $0 < \eta \leq \frac{1}{10}$ and independent errors uniform on $[-t, t]$ with $t \leq 10^{-3}$: the two cells have different patterns, (a) holds on the whole support, and $0 < k_{2,c} \leq 8\eta t / (1 - 2\eta t)$ outside a null set.

The bound of Theorem 15 can be nearly attained and can be loose. Three saved instances with independent sign-valued additive errors have certificates combining exact rational output risks with rational intervals for the commutator norms. The stored binary parameters are embedded exactly as rationals. Every sign realization is checked for agreement with the teacher’s masks along both trajectories and for the required Gram inverses at corrected layers. The certificates use the stated recursion, with zero first-layer commutator contribution. They support a ratio of actual excess to $\sum_c \text{tr } K_{L+1,c}$ of approximately 0.999994, a three-layer case in which dropping $\Psi_{l,c}(V_{l,c})$ makes the bound fail, and another ratio of approximately 0.74.

The accompanying code repository’s `numerics/saved_witnesses/` directory contains the configurations. The near-attainment case is `checkG_climb_3_1.json`; the omitted-term case is `checkG_climb_variant_a_1.json`; the third is `checkG_search_seed5.json`. Their rational risks and interval endpoints are indexed by the same filenames in `SAVED_WITNESS_REPLAY.json`; `replay_fixed_witnesses.py` replays only these saved inputs. These instances support near attainment and the need for the later-noise term, not optimality of the constant or a guarantee for actual quantizers. In Example 14 the bound exceeds the actual excess by more than an order of magnitude. Neither result removes condition (a), and we do not claim that trained networks satisfy it.

6 Fixed rounding errors and the published guarantee

Sections 4 and 5 establish stochastic network comparisons under explicit assumptions. In contrast, the guarantee published with QEP is asserted for arbitrary fixed perturbation matrices shared by the two schemes. Under this deterministic convention, the comparison fails already for two scalar linear layers. Here the corrected scheme denotes the idealized correction in that guarantee; actual codebook minimizers are treated separately in Section 7.

To our knowledge the only published statement of this kind is the guarantee given with Quantization Error Propagation (QEP) [1], which adds a closed-form correction to the weights before quantizing them, scaled by a propagation strength $\alpha_l \in [0, 1]$. Theorem 5.2 of [1] states that the output error of the propagated scheme is bounded by that of the uncorrected scheme; its formal version, Theorem B.5, states that a Lipschitz upper bound U^{QEP} on the output error satisfies $U^{\text{QEP}} \leq U^{\text{BASE}}$ for any fixed perturbation matrices and every choice of strengths. Proposition 5.4 adds that the output error decreases monotonically as each α_l approaches one. In the published discussion of the paper the authors describe Theorem 5.2 as “a rigorous guarantee that solving the layer-wise local optimization problem ... reduces the final quantization error”¹ and use it to distinguish QEP from a concurrent method with the same local objective.

Claims under examination.

- [1, Theorem B.5]: for every choice of $\{\alpha_l\} \subset [0, 1]$, $U^{\text{QEP}} \leq U^{\text{BASE}}$, via the per-layer inequality $\|R_l^{\text{QEP}}\|_F \leq \|R_l^{\text{BASE}}\|_F$ for all l .
- [1, Theorem 5.2, informal]: $\|f_\theta(X) - f_{\hat{\theta}^{\text{QEP}}}(X)\|_F \leq \|f_\theta(X) - f_{\hat{\theta}^{\text{BASE}}}(X)\|_F$ for the quantizers defined by the two layer objectives.
- [1, Proposition 5.4]: the output error of QEP decreases monotonically as each α_l approaches 1.

The appendix of [1] works to first order in $\{E_l\}$ and writes the same symbol E_l for the perturbation in both schemes. The counterexample of Theorem 17 respects both conventions.

Theorem 17 (Counterexample to Theorem B.5 of [1]). *Let $L = 2$, all matrices 1×1 , $\sigma_1 = \sigma_2 = \text{id}$, $\gamma_1 = \gamma_2 = 1$, $W_1 = W_2 = 1$, $X = 1$, $\alpha_1 = \alpha_2 = 1$, and $E = (\varepsilon, -\varepsilon)$ with $0 < \varepsilon < \frac{1}{2}$. Then all hypotheses of Theorem B.5 hold, including invertibility of the Gram matrices and of all quantized weights, and*

$$\begin{aligned} R_1^{\text{BASE}} &= R_1^{\text{QEP}} = -\varepsilon, & \hat{X}_2^{\text{BASE}} &= \hat{X}_2^{\text{QEP}} = 1 + \varepsilon, \\ R_2^{\text{BASE}} &= \varepsilon^2, & R_2^{\text{QEP}} &= \varepsilon + \varepsilon^2, \\ U^{\text{BASE}} &= \varepsilon + \varepsilon^2, & U^{\text{QEP}} &= 2\varepsilon + \varepsilon^2, \\ |X_3 - \hat{X}_3^{\text{BASE}}| &= \varepsilon^2, & |X_3 - \hat{X}_3^{\text{QEP}}| &= \varepsilon + \varepsilon^2. \end{aligned}$$

Hence $U^{\text{QEP}} - U^{\text{BASE}} = \varepsilon > 0$ and the QEP output error is strictly larger, for every $0 < \varepsilon < \frac{1}{2}$.

Proof. In the first layer $\delta_1 = 0$, so the correction vanishes and both schemes use $\hat{W}_1 = 1 + \varepsilon$, giving $\hat{X}_2 = 1 + \varepsilon$ and $R_1 = 1 - (1 + \varepsilon) = -\varepsilon$. In the second layer $\delta_2 = -\varepsilon$ and $\hat{H}_2 = (1 + \varepsilon)^2 > 0$.

BASE. $\hat{W}_2 = 1 - \varepsilon$, output $(1 - \varepsilon)(1 + \varepsilon) = 1 - \varepsilon^2$, so $R_2 = \varepsilon^2$.

QEP. $W_2^* = 1 + \delta_2 \hat{X}_2 / \hat{X}_2^2 = 1/(1 + \varepsilon)$, $\hat{W}_2 = 1/(1 + \varepsilon) - \varepsilon$, output $1 - \varepsilon(1 + \varepsilon) = 1 - \varepsilon - \varepsilon^2$, so $R_2 = \varepsilon + \varepsilon^2$.

With $\gamma_s = \|W_s\|_2 = 1$, $U^M = |R_1^M| + |R_2^M|$. Since σ_2 is the identity, the output error equals $|R_2^M|$. For $0 < \varepsilon < \frac{1}{2}$ the quantized weights $1 + \varepsilon$, $1 - \varepsilon$ and $1/(1 + \varepsilon) - \varepsilon$ are positive. (The algebraic identities hold for every $\varepsilon > 0$; the QEP second-layer weight vanishes at $\varepsilon = (\sqrt{5} - 1)/2$ and the BASE one at $\varepsilon = 1$.) $\square$

¹ Authors’ rebuttal of 30 July 2025 in the public discussion of [1], OpenReview forum [a313K9khhbL](#), under “Weakness and Question 2”; archived copy consulted on 4 October 2026.

The gap is first order: $\max_k |E_k| = \varepsilon$ and $(U^{\text{QEP}} - U^{\text{BASE}})/\varepsilon = 1$ throughout, while the two bounds and the two output errors differ from their first-order expressions by exactly ε^2 . The first-order convention of [1] therefore cannot absorb it. Both schemes see the same second-layer input, so the failure does not come from comparing different trajectories.

Remark 18 (The sign matters). With $E = (\varepsilon, \varepsilon)$ the same computation gives $U^{\text{BASE}} = 3\varepsilon + \varepsilon^2$, $U^{\text{QEP}} = 2\varepsilon + \varepsilon^2$ and output errors $2\varepsilon + \varepsilon^2$ against $\varepsilon + \varepsilon^2$: QEP is better. The ordering depends on the relation between the inherited error and the new perturbation, which is what the identity of Section 3.1 makes precise.

Theorem 19 (Failure of monotonicity in the shared-perturbation model). *In the setting of Theorem 17, keep $\alpha_1 = 1$ and let $\alpha_2 = \alpha \in [0, 1]$. Then the QEP output is $1 - \alpha\varepsilon - \varepsilon^2$ and the output error is $\alpha\varepsilon + \varepsilon^2$. For $0 \leq \alpha < \beta \leq 1$ the error at β exceeds the error at α by $(\beta - \alpha)\varepsilon > 0$.*

Proof. $W_2^*(\alpha) = 1 - \alpha\varepsilon/(1 + \varepsilon)$, so $\widehat{W}_2 = 1 - \alpha\varepsilon/(1 + \varepsilon) - \varepsilon$ and the output is $(1 + \varepsilon)\widehat{W}_2 = 1 - \alpha\varepsilon - \varepsilon^2$. $\square$

This contradicts Proposition 5.4 of [1] read in the shared-perturbation model of its appendix. It does not contradict Corollary B.6 of [1]: that corollary bounds the residual by $\|W_l \delta_l(I - \alpha P_l)\|_F + \|E_l\|_2 \|\widehat{X}_l\|_F$, which here equals $(2 - \alpha)\varepsilon + \varepsilon^2$, is decreasing in α , and remains a valid upper bound on $\alpha\varepsilon + \varepsilon^2$ for $\alpha \leq 1$. The bound is monotone; the error it bounds is not.

Where the proof breaks. Let $P_l = \widehat{X}_l^\top \widehat{H}_l^{-1} \widehat{X}_l$, the orthogonal projection onto the row space of $\widehat{X}_l$. The proof of Theorem B.5 correctly derives

$$R_l^{\text{QEP}} = W_l \delta_l(I - \alpha_l P_l) - E_l \widehat{X}_l, \quad \widetilde{R}_l^{\text{BASE}} := W_l \delta_l - E_l \widehat{X}_l,$$

where $\widetilde{R}_l^{\text{BASE}}$ is the BASE residual on the same $\widehat{X}_l$, and correctly notes $\|W_l \delta_l(I - \alpha_l P_l)\|_F \leq \|W_l \delta_l\|_F$. It then asserts, “by the triangle inequality”,

$$\|W_l \delta_l(I - \alpha_l P_l) - E_l \widehat{X}_l\|_F \leq \|\widetilde{R}_l^{\text{BASE}}\|_F.$$

This does not follow. Shrinking one term of a difference does not shrink the difference. In Theorem 17, $W_2 \delta_2 = -\varepsilon$ and $E_2 \widehat{X}_2 = -\varepsilon - \varepsilon^2$ nearly cancel in the BASE residual; QEP removes the first term and leaves the second intact.

The proof contains a second unjustified step: having fixed $\widehat{X}_l = \widehat{X}_l^{\text{QEP}}$, it concludes by “taking $\widehat{X}_l = \widehat{X}_l^{\text{BASE}}$ ”, although the two trajectories differ in general. Theorem 17 isolates the first gap, since there the trajectories coincide up to the layer in question. The second gap causes failure on its own, as Example 26 shows.

Three further facts about the shared-perturbation statements are established in Appendix B. The failure is not an artifact of $P = I$ or of scalar networks: it occurs with a proper projection, lifts to matrices, and occurs in a family that does not reduce to independent scalar channels (Example 25). The substitution of one trajectory for the other fails on its own, in a three-layer network where the first step causes no harm (Example 26). The composite inequality $\|\delta_{L+1}^{\text{QEP}}\|_F \leq U^{\text{BASE}}$, also asserted in Theorem B.5 of [1], fails for finite perturbations (Example 27); it holds once each factor $\|W_s\|_2$ in U^{BASE} is replaced by $\|W_s\|_2 + \|E_s\|_2$, and therefore to first order.

7 Actual nearest-codeword quantizers

The shared-perturbation model fixes E_l independently of the target. Real quantizers do not behave this way.

Remark 20 (The correction is piecewise constant). Let Q be a deterministic nearest-codeword map with a fixed codebook and scale, and $\widehat{W}(\alpha) = Q(W + \alpha K)$ with $K = W\delta\widehat{X}^\top\widehat{H}^{-1}$. With the layer input and all other layers held fixed, on any interval of α on which no entry of $W + \alpha K$ crosses a decision boundary, $\widehat{W}(\alpha)$ is constant, so the quantized layer and its loss are constant. Within such an interval the change in the continuous target is cancelled exactly by the change in rounding error: for α, α_0 in the same interval, $B_\alpha = B_{\alpha_0} - (\alpha - \alpha_0)K\widehat{X}$. The loss therefore changes only at boundary crossings, which are determined by the path of the target and by the codebook; the quadratic of Proposition 1, which holds B fixed, does not describe them.

Remark 21 (No local harm at full strength with exact minimization). For any candidate Q and common input $\widehat{X}$,

$$\|WX - Q\widehat{X}\|_F^2 = \|W\delta(I - P)\|_F^2 + \|(W^*(1) - Q)\widehat{X}\|_F^2.$$

Hence an exact minimizer over the codebook of the QEP objective with $\alpha = 1$ attains a layer residual no larger than any other codeword assignment on that input, including the one chosen by the uncorrected objective. This is the standard least-squares decomposition and is implicit in [1, Eq. (5)]. It shows that the shared-perturbation instance of Theorem 17 cannot arise from exact minimization of both objectives on a common input with a common codebook. It says nothing about the network output.

The next result shows that output-level domination fails for exact minimizers.

Theorem 22 (An actual finite-codebook family). *Let the codebook be $\{0, 1, 2, 3\}$ and consider a three-layer network with scalar weights $a, b, c > 0$, ReLU activations and two calibration samples $X = (-1, 1)$. Let each layer be quantized by exact minimization over the codebook, the uncorrected scheme using the objective of [1, Eq. (1)] with its own quantized inputs and QEP using [1, Eq. (3)] with $\alpha = 1$ and no damping. Suppose*

$$\frac{1}{2} < a < \frac{3}{2}, \quad \frac{3}{2} < b < \frac{5}{2}, \quad \frac{5}{2} < ab < 3, \quad \frac{1}{2} < c < \frac{3}{2}, \quad \frac{3}{2} < abc < \frac{7}{2}.$$

Then every layer minimizer in both schemes is unique, the uncorrected scheme selects $(1, 2, 1)$ and QEP selects $(1, 3, 1)$, the continuous targets lie strictly inside $(0, 3)$, and all input Gram matrices are positive. Writing $t = abc$ for the teacher output on the active sample, the output errors are $|t - 2|$ and $|t - 3|$, so

$$\text{QEP is strictly worse} \iff t < \frac{5}{2}, \quad \text{QEP is strictly better} \iff t > \frac{5}{2}.$$

Moreover, the better of the two attains the minimum output error over all 64 codeword triples.

Proof sketch. Since $a > 0$, the first sample is mapped to 0 by every layer and the second stays positive; the network acts on the second sample as the scalar chain $x \mapsto abc x$. The uncorrected scheme rounds a, b, c to 1, 2, 1. QEP rounds a to 1; its second-layer target is $ab \in (\frac{5}{2}, 3)$, which rounds to 3; its third-layer target is $abc/3 \in (\frac{1}{2}, \frac{7}{6})$, which rounds to 1. The outputs are 2 and 3. Global optimality follows because every codeword triple outputs a non-negative integer and 2, respectively 3, is the nearest integer to t on the corresponding side of $\frac{5}{2}$. $\square$

We state plainly what this family is. It is a scalar chain: one sample is inactive throughout and the ReLU never acts on a sign change. The threshold $t = \frac{5}{2}$ is the midpoint of the two attainable outputs. The content of the theorem is that exact layer-wise minimization, with interior targets, unique minimizers and no clipping, produces a QEP output that is strictly worse on an open set of parameters. In the second layer QEP’s choice is locally better (the pre-activation 3 is closer to the teacher’s ab than 2 is), and the third layer’s grid is too coarse to exploit it. For instance, $(a, b, c) = (\frac{7}{5}, 2, \frac{18}{25})$ gives teacher output 2.016, uncorrected error 0.016 and QEP error 0.984; changing c to $\frac{11}{10}$ gives teacher output 3.08, and QEP is better.

The same instance shows non-monotonicity in the propagation strength for an actual quantizer. At $(\frac{7}{5}, 2, \frac{18}{25})$, keeping $\alpha_1 = \alpha_3 = 1$, with second-layer strength $\alpha_2 = \frac{1}{2}$ the selected codewords are $(1, 2, 1)$ and the squared output error is $4/15625$; with $\alpha_2 = \frac{3}{4}$ they are $(1, 3, 1)$ and it is $15129/15625$. The decision boundary is at $\alpha_2 = \frac{5}{8}$. This instance is formalized, including the tie at the boundary, where the smaller codeword is selected.

8 Numerical illustrations

The statements of Sections 4 and 5 are proved. The computations of this section show how large the quantities are in small synthetic networks. They are of three kinds. For Figure 1 and Table 2 the actual algorithm is run: the corrected target is computed with the Gram inverse, the weights are rounded once per run with entrywise subtractive dither on an unbounded grid, and the rounded weights are applied to all calibration samples. For the left panel of Figure 2 the rounding errors are sampled from the uniform law that this dither produces, and for the right panel sign-valued errors are enumerated; no quantizer is called in these two. Models, strengths, sample sizes and reported quantities were fixed before running. Scripts and raw outputs accompany the paper.

How large the gain is. Figure 1 concerns the class of Example 11 with $d = 4$ and eight calibration samples, for depths up to 64. The simulated normalized errors agree with the closed form for a fixed teacher and for random teachers: of the 168 simulated values, 156 lie within two standard errors of the prediction, and the largest deviation is 2.7 standard errors. Values at different depths come from the same runs and are not independent, so these counts are descriptive. Without compensation the error grows with the depth as $(1 + \nu)^L - 1$; with full compensation it stays close to ν , the rounding noise of the last layer; with $\alpha = \frac{1}{2}$ it settles at an intermediate level. At fine resolution the gain follows $10 \log_{10} L$ over the depths shown. At the coarser resolutions and the largest depths shown it exceeds this value, since the uncorrected error then grows geometrically: at depth 64 the simulated gains are 18.1, 18.2, 19.4 and 26.1 dB for $\nu = 10^{-4}, 10^{-3}, 10^{-2}$ and 5×10^{-2} , against $10 \log_{10} 64 \approx 18.1$.

Where the gain comes from. Table 2 concerns a linear network with two bottlenecks, where the row spaces of the original and quantized activations differ and Corollary 10 does not apply. The gain predicted by Proposition 9, from sources estimated on one batch of runs, agrees within two standard errors with the gain measured on an independent batch. In this network between 2% and 3% of the gain comes from the reduced excitation of later rounding noise, the part that a per-layer comparison does not see; the rest is error removed directly. The gain increases with the strength here, which Theorem 8 does not guarantee.

ReLU networks. The left panel of Figure 2 is Example 14. Over all runs no activation changed state, in either scheme and at either layer, and the excess of QEP over BASE divided by t^2 lies

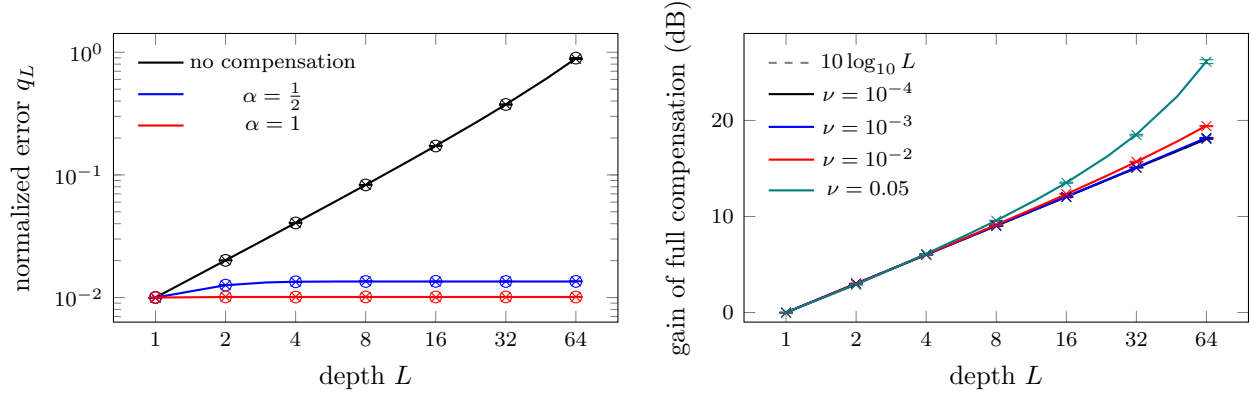


Figure 1: Gain as a function of depth in the class of Example 11, $d = 4$, with the actual algorithm and dithered rounding of the weights. Lines are the closed form; marks are simulations with two standard errors. Left: $\nu = 10^{-2}$; circles, fixed teacher $W_l = I$ (2×10^4 runs); crosses, random teachers (100 teachers, 200 runs each, errors clustered by teacher). Right: ratio of the two errors, random teachers, with standard errors paired by teacher. At depth one the two schemes coincide.

α	removed directly	through later noise	predicted gain	measured gain	error of QEP
0.25	0.0882	0.0017	0.0898	0.0897 ± 0.0022	0.2093
0.5	0.1356	0.0030	0.1386	0.1392 ± 0.0024	0.1598
0.75	0.1579	0.0040	0.1619	0.1621 ± 0.0024	0.1369
1	0.1640	0.0046	0.1686	0.1686 ± 0.0024	0.1305

Table 2: The two parts of the gain in Proposition 9, in a linear network $8 \rightarrow 6 \rightarrow 4 \rightarrow 2$ with dither of step 0.3 and strengths $\alpha_2 = \alpha_3 = \alpha$. The prediction is computed from the recursions (5)–(6) with sources estimated on 10^4 runs; the measured gain comes from 10^4 independent runs and is given with two standard errors. The mean-square error without compensation is 0.2991.

between 0.071 and 0.076, with a standard error of 0.002, for t from 5×10^{-4} to 10^{-1} , against the leading coefficient $\frac{2}{27} \approx 0.074$. The harm is a second-order effect of constant relative size, present at error sizes well beyond the interval for which the lower bound is proved. The right panel is the bottleneck family described after Corollary 16, with sign-valued errors and exact enumeration. The projections do not commute with the cells, the largest commutator norm decreases as t , and the bound of Theorem 15 on the excess decreases as t^3 ; relative to the uncorrected error it is about 10^{-5} at $t = 10^{-4}$ and 3×10^{-3} at $t = 3 \times 10^{-2}$. In this family the correction helps: the error of QEP is about 28% of that of BASE at every t , so the bound is far from attained. No activation changed state.

These are small networks chosen to exhibit the statements; they do not test whether trained networks satisfy the hypotheses. The Monte Carlo values were computed twice, with two independent implementations and random generators, one of them free of any numerical library. On identical inputs the two agree to rounding error. Run independently they agree overall: of the 168 values of Figure 1, 151 differ by at most two combined standard errors and the largest difference is 3.06; these values are correlated and the comparison is descriptive.

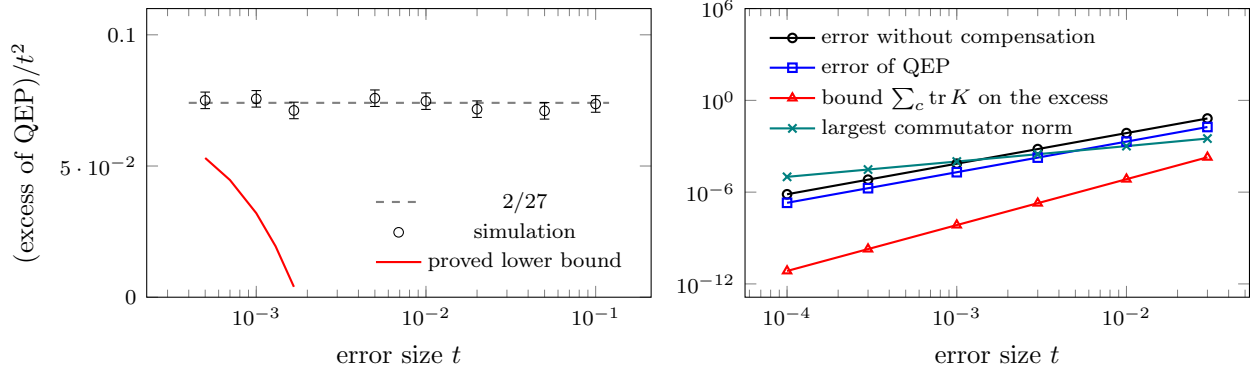


Figure 2: ReLU networks. Left: Example 14, rounding errors sampled from the uniform law of subtractive dither, 10^6 independent runs per value of t , two standard errors; no activation changed state in any run. Right: the bottleneck family after Corollary 16 with $\eta = \frac{1}{10}$ and sign-valued errors, exact enumeration; here the correction helps, the bound on the excess decreases as t^3 and the commutator as t .

9 Relation to experiments

The comparisons of Sections 3–7 hold the calibration input fixed; the randomness is that of rounding. A quantized model is used on other inputs, and whether an improvement obtained on calibration data carries over to them is a different question, on which those results say nothing. This section reports what we observed on both questions: published observations, a layer-wise diagnostic on two small models, and experiments on two 8B models. None of the experiments satisfies the hypotheses of our theorems, and they are not tests of them.

9.1 Published observations

Table 1 of [1] reports WikiText-2 perplexity for 60 pairs (four base methods, five models, three bit widths) with and without QEP. In 53 pairs QEP is better. In 7 pairs it is worse, all at 2 bits: RTN on Llama-2-7B and Llama-2-13B, GPTQ on Llama-2-13B and Llama-3-8B, and AWQ on Llama-2-7B, Llama-2-70B and Mistral-7B; in all seven the perplexity is above 10^3 both with and without QEP. These experiments use $\alpha_l = \frac{1}{2}$ by default, except in the MLP layers of Llama-2-70B, where $\alpha_l = 0$. [17] uses the same propagated objective under the name activation drift correction, reports that it becomes “increasingly unstable in deeper layers” of larger models, and chooses a per-layer coefficient between drift-corrected and original calibration statistics by search; in most deeper layers of the Qwen3-8B configuration they report, the selected coefficient is close to the uncorrected end. In their reply of 30 July 2025 to reviewer pLsN in the same public discussion as above, under “Propagation Strength” and “Minor Question 1”, the authors of [1] report perplexities for GPTQ at 3 bits: on Llama-2-7B, 10.881 without QEP, 7.898 at $\alpha = \frac{1}{2}$ and 8.169 at $\alpha = 1$; on Llama-2-70B, 4.860, 4.102 and 4.771, where the middle entry uses $\alpha = 0$ in the MLP down-projection matrices and $\frac{1}{2}$ elsewhere. The authors attribute the loss at full strength to overfitting. Perplexity is not the quantity addressed by Proposition 5.4 of [1]. We have not run these configurations and do not attribute any of these observations to the mechanism of Section 3.1.

9.2 A layer-wise diagnostic on two small models

If, for actual rounding errors, the ratio c/d of Section 3.1 is typically far below the threshold of Corollary 2, the failure mode of Section 6 has little practical relevance. We measured this ratio on SmolLM2-135M [22] and Pythia-160M [23], each with two calibration texts, WikiText-2 and a public-domain novel. All linear maps inside the transformer blocks, 210 and 48 respectively, are first rounded to the nearest point of an unbounded uniform grid with step $\max_j |W_{ij}|/7$ in row i . On this common quantized trajectory, and for each linear map separately, we compare the layer residual with and without the correction at $\alpha = \frac{1}{2}$, using 4096 calibration tokens; the corrected layer is not propagated further. The correction is computed from a truncated singular value decomposition of the quantized input, keeping singular values above 10^{-10} times the largest, without damping; all linear maps have full effective rank except one of the 46 maps of Pythia-160M with $d > 0$, for each text. Here c is computed with the rounding error of the uncorrected layer. The labels in Table 3 use the following numerical conventions, with all energies divided by the number of tokens. A linear map enters the ratios when $d > 10^{-12} \max(\|W\delta\|_F^2, 10^{-30})$; three maps of SmolLM2-135M and two of Pythia-160M do not. With the rounding error held fixed, a map is counted as worse when $2c - (2 - \alpha)d > 10^{-8} \max(d, 2|c|, \|B\|_F^2, 10^{-30})$, where B is the rounding error of the uncorrected layer applied to the quantized input. After re-rounding, on calibration or held-out tokens, the comparison uses the squared Frobenius residual per token, $\|WX - Q\hat{X}\|_F^2/N$ for the quantized weight Q and N tokens: a map is counted as worse when this quantity with the correction exceeds its value without the correction by more than 10^{-8} times the larger of the two values (or 10^{-30}). On held-out tokens the correction fitted on the calibration tokens is reused without refitting. The protocol, the thresholds and the sample were fixed before any measurement.

	SmolLM2-135M		Pythia-160M	
	WikiText-2	novel	WikiText-2	novel
Linear maps with $d > 0$	207	207	46	46
Range of c/d	$[-0.063, 0.070]$	$[-0.065, 0.065]$	$[-0.012, 0.025]$	$[-0.038, 0.023]$
Worse, same rounding error	0	0	0	0
Worse, re-rounded, calibration	0	0	0	0
Worse, re-rounded, held-out	19	0	11	12
Weights outside row range	0.19%	0.21%	9.1%	9.3%

Table 3: Layer-wise diagnostic at $\alpha = \frac{1}{2}$ with deterministic rounding. “Worse” counts linear maps whose squared residual is larger with the correction than without it, up to the numerical tolerances stated in the text. By Corollary 2, with the rounding error held fixed the correction is worse exactly when $c/d > \frac{3}{4}$. The held-out row applies the correction fitted on the calibration tokens to 2048 other tokens. The first five rows come from one measurement on the unbounded grid. The last row comes from a second measurement, fixed in advance separately, on the same weights, corrections and quantized trajectory: it is the fraction, pooled over all weights of all 210 and 48 linear maps, of corrected weights $W + \alpha C$, where $C = W\delta\hat{X}^\top\hat{H}^{-1}$ (with the truncated singular value inverse specified above), whose magnitude exceeds the largest magnitude of the original row.

The ratio c/d never exceeds 0.07 in absolute value and its median is below 10^{-3} in absolute value: the cross term is small relative to the energy of the projected inherited error, and the condition of Corollary 2 is far from violated. This ratio is not a correlation coefficient, and a small value does not mean that the two errors are uncorrelated. Re-rounding the corrected weights does not change the counts on the calibration tokens. With subtractive dither and 16 seeds, over the 8096 pairs of a linear map and a seed, $|c/d|$ stays below 0.20; no pair is worse with the rounding

error held fixed; two pairs, the same linear map of SmolLM2-135M on the novel, are worse after re-rounding on the calibration tokens; and on held-out tokens the number of linear maps that are worse ranges over the seeds from 17 to 20, from 0 to 2, and equals 11 and 12 in the four columns. On these models we did not observe a cancellation large enough to make the correction worse with the rounding error held fixed.

What we did observe is of the other kind. On held-out tokens the correction fitted on the calibration tokens increases the residual of a sizeable fraction of the linear maps, up to 12 of the 48 maps of Pythia-160M. In addition, the correction moves weights out of the range of the original row, for 9% of the weights of Pythia-160M at $\alpha = \frac{1}{2}$ and 17–18% at $\alpha = 1$, so a bounded codebook with a range set by the original weights would overload, which is the situation of Section 3.3. This diagnostic intervenes on one layer at a time on a common trajectory and does not run the two schemes to the output.

9.3 Choosing the strength on two 8B models

Setting. We quantized Qwen3-8B and Meta-Llama-3-8B with our own implementation of a QEP-style correction on top of round-to-nearest; it is not a reproduction of QEP, GPTQ, AWQ or QuIP#. The seven linear maps of each transformer block are quantized to 4 bits with an asymmetric uniform codebook (integer codes 0–15, a scale and zero point per row and group of 128 input channels); embeddings, normalizations and the output head stay in BF16, and the dequantized weights are executed in BF16, so no speed is claimed. The maps are processed in dependency order in four groups (query–key–value, attention output, gate–up, down), maps that share an input being quantized together; each policy follows its own quantized trajectory, and the inputs of every group are taken from the actual modules. For a group with quantized input Z and teacher input X , the correction is fitted by ridge least squares with $\lambda = \text{tr}(ZZ^\top)/d$, the mean eigenvalue of the Gram matrix, and every corrected candidate is re-quantized with recomputed scales and rounded deterministically. The ridge matters for reading the results: the matrix $P_\lambda = Z^\top(ZZ^\top + \lambda I)^{-1}Z$ that replaces P is not a projection, and a direction in which the Gram matrix has eigenvalue s is compensated in proportion $\alpha s/(s+\lambda)$. “Strength 1” below means external strength 1 with the ridge retained; on the fitting inputs and before re-quantization, in a direction with $s = \lambda$ it compensates half of the inherited error.

For each model we drew, three times, twelve disjoint 512-token windows from WikiText-2 training text: four (2048 tokens) to fit the corrections and eight (4096 tokens) to choose strengths. Evaluation uses the first 16384 tokens of the WikiText-2 test text and of *Pride and Prejudice* (Project Gutenberg 1342, with the wrapper removed and the front matter kept), in non-overlapping 512-token windows with the context reset for each window; neither enters any decision. Each model uses its own tokenizer. The policies are: no correction (RTN); strength $\frac{1}{2}$ in every group; strength $\frac{1}{2}$ with the down projections uncorrected; and four rules that choose, for each group, a strength in $\{0, \frac{1}{2}, 1\}$ on the selection data, by the output error of the group (local), the error after the next block kept at RTN (block), the loss of the full network with the remaining blocks kept at RTN (network), or a first-order estimate of that loss from its gradient with respect to the block output, applied to the re-quantized candidates (gradient). The replication protocol, including all seven policies, was fixed after an exploratory round on validation text and before any of the reported evaluations.

What we observed. Every compensating policy has a lower mean perplexity than RTN in each model and text (Table 4), and fixed strength $\frac{1}{2}$ is lower than RTN in each of the twelve combinations of model, text and draw; the network rule on the novel with Llama-3-8B is above RTN in one draw.

Model	Text	RTN	$\alpha = \frac{1}{2}$	local	block	network	gradient
Qwen3-8B	WikiText-2	12.643	12.320	12.119	12.199 (1)	11.970 (2)	11.870 (3)
	novel	12.875	12.618	12.665	12.552 (3)	12.613 (2)	12.582 (1)
Llama-3-8B	WikiText-2	8.264	8.031	8.018	8.017 (2)	8.075 (0)	8.028 (0)
	novel	5.786	5.625	5.586	5.627 (1)	5.766 (0)	5.689 (1)

Table 4: Perplexity at 4 bits, mean over three calibration draws. In parentheses, the number of draws, out of three, in which the rule has a lower perplexity than the local rule. RTN does not depend on the draw.

No rule beats the local rule consistently. The gradient rule wins all three draws on Qwen3-8B with WikiText-2 and none on Llama-3-8B with WikiText-2. On Llama-3-8B the network rule loses to the local rule in all six comparisons, although, strength 0 being always a candidate, none of its 816 decisions over both models and three draws increases the full-network loss on the selection data relative to leaving that group uncorrected. These two facts compare different things, the chosen candidate with no correction in the current state and one complete policy with another; the next paragraph compares the two policies directly. The gradient rule took 4.1–4.8 times the construction time of the local rule in our prototype.

More selection data. To see whether the selection data were simply too few, we kept the 2048 fitting tokens of two of the draws, enlarged the selection data in nested sets of 4096, 16384 and 32768 tokens, added a separate 8192-token diagnostic set from the same source that enters no decision, and evaluated on new windows (tokens 65536 to 81920) of both texts. These runs record the loss of both final policies on every data set, so the same pair can be compared with the same metric on different data. On Llama-3-8B with 4096 selection tokens, averaged over the two draws, the network policy has a lower loss per token than the local policy on the selection data, by 0.0210, and a higher one on the diagnostic data (+0.0084), on WikiText-2 (+0.0069) and on the novel (+0.075): the ordering of the two policies reverses when the data change. With more selection data the advantage on the selection data shrinks (−0.0116 at 16384 tokens, −0.0087 at 32768) and so does the deficit elsewhere, to +0.0025 on WikiText-2 and +0.030 on the novel, behind in every draw on the novel. On Qwen3-8B the network rule is ahead at every size, most at 16384 tokens (−0.033 and −0.018) and less at 32768. Selection by the output error of the current block does not help consistently either. The candidates themselves were always fitted on the same 2048 tokens, so these results do not separate the quality of the candidates from the accuracy of choosing among them.

Where the overestimate lies. Measuring each policy against RTN separates the two rules. For a policy, take its gain over RTN on the selection data minus its gain over RTN on the diagnostic data; the difference between the two policies’ values is the change of their ordering reported above. With 4096 selection tokens, averaged over the two draws, this overestimate is +0.022 on Llama-3-8B and +0.034 on Qwen3-8B for the network rule, and −0.007 and −0.006 for the local rule. A further run with the fitting data of the first draw, two new selection draws and 32 source windows used by no policy gave the same ordering in all four cases, with 0.029 to 0.049 for the network rule and 0.008 to 0.022 for the local rule. The network rule chooses by the loss of the full network on the selection data, which is also the quantity reported on those data; the local rule chooses by a linear output error. A small standard error does not exclude the overestimate: on Llama-3-8B in the first draw the network rule is ahead of the local rule on the selection data by 0.025, about six standard

errors over the eight windows, and behind on the diagnostic data. The overestimate is comparable in the two models; the gains are not. On the selection data the network rule is ahead of the local rule by 0.052 on Qwen3-8B and by 0.021 on Llama-3-8B, and only the first survives. Fixed strength $\frac{1}{2}$ makes no choice on the selection data and pays no such cost. Reusing the same eight windows for every decision is not the explanation. In the same run, giving each decision eight windows not used before in its path, with everything else fixed, did not reduce the overestimate of the network rule: summed over the decisions and measured on a fixed set of eight held-out source windows, it went from 0.031–0.043 with reused windows to 0.070–0.083 with fresh ones, and the mean final loss on held-out source windows did not improve. Fresh windows remove reuse across decisions, but not the choice, at each decision, of the candidate that scores best on the very windows on which its gain is then measured. This run used a protocol, predictions and a stopping rule fixed before inference; fresh windows also change the amount of text seen, so it does not isolate reuse at an equal data budget.

Changing the suffix. Choosing group by group, each rule above scores a candidate with the later blocks still at RTN, and then changes those blocks. To see whether this explains the failures, we replayed 48 recorded decisions (two draws per model; three blocks per draw, four groups each; the groups of a fourth, final block serve as a control) with the earlier state and the three candidates fixed, scoring them once with the RTN suffix and once with the corrected suffix reconstructed from the recorded decisions. In 11 of the 48 the preferred candidate on the selection data reverses by more than 10^{-4} in loss; none of these 11 reversals recurs, for the same pair and in the same direction, on the diagnostic data. Dependence on the suffix exists, but we found no evidence that it accounts for the losses on new text.

The strength from the fixed-error quadratic. Remark 3 warns that the minimizer of a quadratic with the rounding error held fixed is not by itself a rule for an actual quantizer. We measured how far it goes at 2 bits (codes 0–3, same group structure), in 48 slices (two draws per model, three blocks spread over the depth, four groups), each candidate re-quantized with recomputed scales and applied to a common RTN trajectory. With S the output of the ridge correction, the strength minimizing $\|R_0 - \alpha S\|_F^2$ on the selection data, clipped to $[0, 1]$, ranks the re-quantized strengths with a mean rank correlation of 0.977 to 0.994 across models and draws, and on the diagnostic data it lowers the normalized residual of the group by 6.4 to 9.2% relative to strength $\frac{1}{2}$, on average within each model and draw (in 47 of the 48 slices). Seven of the 48 values were above 1 before clipping, the largest 1.16; since the ridge has already shrunk the correction, these numbers say nothing about the best unshrunk strength. The fixed pair $\{\frac{1}{2}, 1\}$, chosen between on the selection data with strength 0 retained as a fallback (as in the grid of ten non-zero strengths), already retains 98.2 to 99.6% of the diagnostic gain of a ten-point grid. These are linear residuals, not perplexities. The scale is recomputed for each candidate, which places this setting outside Remark 20; a high rank correlation shows that the quadratic predicts the ordering, not that the loss is a smooth function of the strength.

9.4 What the experiments do and do not show

On a fixed calibration input, Sections 4–7 give conditions under which compensation cannot hurt and mechanisms by which it can. In the experiments, compensation helped on average relative to no compensation, and on the small models the cancellation of Section 6 was not observed with the rounding error held fixed; we did not measure the cross term on the 8B models, so we neither confirm nor exclude it there. The held-out comparisons reveal a further limitation:

gains measured on fitting or selection data need not persist on other inputs. A correction that lowers a layer residual on calibration tokens raises it on others, and a policy that is better than another on the selection data can be worse on new text. This is separate from the two increases on calibration tokens observed after re-rounding with dither in the small-model diagnostic, and from the fixed-input failures of Sections 6 and 7. We have not identified the cause of the held-out losses. The overestimate is larger for the rule that chooses by the reported quantity, and reusing the selection data across decisions does not explain it; the finite fitting data, the size and domain of the selection data, the ridge and the interaction between layers are not separated by these experiments. Calibrating quantized weights against the output of a whole block rather than a single layer is not new [28, 29]; our network rule goes further and scores candidates by the loss of the full network. What these experiments add is a measurement, in one setting, of how far such choices carry over to new text. The scope is narrow: two models, two or three draws, two texts that may appear in pretraining data, windows rather than independent documents, our own RTN-backed implementation and dequantized execution.

10 Formal verification

The statements below are checked in Lean 4 (version 4.34.1) with Mathlib at commit `d13f23b`. The accompanying artifact contains 96 modules with 969 explicitly named theorem declarations, auxiliary lemmas included; every one of them depends only on the axioms `propext`, `Classical.choice` and `Quot.sound`. The number of declarations is not a count of results. The artifact also contains positive controls and ten deliberately false statements that are required to fail. The table lists the eleven modules for the statements of Sections 3.1, 6 and 7, with 153 declarations; the remaining modules are described after it.

Module	Content	Decl.
<code>QepSharedE</code>	Theorem 17, the sign remark, arbitrarily small ε	15
<code>QepSharedEExtended</code>	Exact parameter domain of Theorem 17: the quantized weights are non-zero for every $\varepsilon > 0$ except $\varepsilon = 1$ and $\varepsilon^2 + \varepsilon = 1$	6
<code>QepAlpha</code>	Theorem 19; validity and monotonicity of the bound of [1, Cor. B.6]	8
<code>QepProjectionExample</code>	First part of Example 25	7
<code>QepMatrixFamily</code>	Second part of Example 25 for $D = \varepsilon I$ and general inputs, and the instance $X = [I_2 \mid \mathbf{1}]$	25
<code>QepCorrected</code>	Proposition 1, Corollary 2(i), finite weighted form of Theorem 4	10
<code>QepCorollary</code>	Corollary 2(ii)–(iv)	12
<code>QepGram</code>	$P = \hat{X}^\top \hat{H}^{-1} \hat{X}$ is a symmetric idempotent with $\hat{X}P = \hat{X}$ when $\hat{X}$ has full row rank; Corollary 2(i) for the actual QEP residuals; the decomposition of Remark 21	13
<code>QepDitherFinite</code>	Proposition 7 for independent finitely supported errors with zero mean and equal second moments	6
<code>QepInterior</code>	Theorem 22 and its two numerical instances	43
<code>QepInteriorAlpha</code>	The α_2 instance of Section 7, with its switch at $\frac{5}{8}$	8

In `QepSharedE` the residuals and the bound U are computed from the recursion of Section 2, not assumed. `QepInterior` defines both encoders by an explicit four-way minimization and derives the selected codewords.

The remaining modules were reviewed independently of their authors. They cover the following, each within the stated scope.

- Theorem 8 for finitely supported error laws, with the actual Gram matrices, the expectation step, the joint induction on the two moments, and layers of zero strength.
- Example 12 in full: the recursion with continuous dither, both integrals, the strict inequality and the probability of a positive output.
- For the actual continuous dither on a fixed finite grid: the distribution of the saturating quantizer, the two bounded-codebook examples of Section 3.3, and per-layer energy and cross-term statements on a common input with an arbitrary history, under explicit integrability and rank assumptions.
- Rounding with feedback, for the actual sequential process with continuous dither: with parameters fixed in advance, the relation between final errors and innovations, the covariance of the final error, the output energy on an unbounded grid and the overload term for a finite codebook; with measurable parameters depending on earlier randomness, the joint law of the standardized innovations only.
- Examples 26 and 27 and the repaired composite bound; the non-commutative family of Example 25, with the spectral norm of D .
- For Theorem 15, algebra only: the cell-wise identities; the two inequalities (10) and the two matrix bounds derived from them, under a bilinear bound on the commutator that is not identified with the spectral norm; and an order-propagation step with defect terms in which the per-layer inequalities are hypotheses. These are not linked to the actual Gram projections, the activations or the expectations.
- Proposition 13; the schedule example and the example for the bounds in Section 4.

Not formalized: Theorem 4 as an abstract conditional-expectation statement (the finite weighted version and the fixed-grid versions above are formalized); Theorem 8 for general error laws, where the integrability argument near singular Gram matrices is a paper proof; Proposition 9, Corollary 10 and Example 11, of which the first two were checked by exact enumeration on small networks and all three by the simulations of Section 8; the distributional property of subtractive dither on an unbounded grid for a general lattice, which we cite, and Proposition 7 for continuous dither; Theorem 15 as a statement about the actual stochastic network, Corollary 16, the perturbation bound for projections and the weighted form of Proposition 13; the expectation in Example 14, which is proved on paper in Appendix A; the energy statements for rounding with feedback when the parameters depend on earlier randomness; Remark 20; the consequence drawn in Remark 21 for an exact discrete minimizer; and the matrix family of Example 25 for general diagonal D . Verification consists of standard elaboration and kernel checking with printed axiom dependencies; no independent proof checker was run.

11 Limitations and open questions

- The theorems compare the two schemes on a fixed calibration input. They say nothing about new inputs, where the largest losses we observed occur (Section 9); the overestimate we measured is larger for the rule that selects by full-network loss, but reuse of the selection data does not explain it and we have not established the cause. The cross term c/d was measured only on two small models with single-layer interventions; how often $2c > (2 - \alpha)d$ occurs in larger models, at lower bit widths or along each scheme’s own trajectory is not known to us.

- The composite inequality $\|\delta_{L+1}^{\text{QEP}}\|_F \leq U^{\text{BASE}}$, also asserted in [1, Theorem B.5], fails as an exact statement (Example 27) and holds to first order, in the repaired form given in Appendix B.
- At the network level, Theorem 8 covers linear networks and Example 12 shows that the mean-square comparison fails for ReLU networks. Theorem 15 bounds the cost of non-commuting projections for ReLU networks but requires stable activations, which we have not established for any trained network; a statement that tolerates a small probability of activation changes is open, and other activations are not treated. How $\mathbb{E}(U^{\text{QEP}})^2$ and $\mathbb{E}(U^{\text{BASE}})^2$ compare for linear networks under uniform dither is not settled in this paper.
- We have not analyzed damping, although the experiments of Section 9 use it. With a ridge term, $T - X = (Z - X)(I - \alpha P_\lambda)$ still contracts the inherited error on the fitting inputs, before re-quantization, but the inequality for the activation moment in the proof of Theorem 8 does not carry over as stated, and a network comparison is not established. With a ridge term the matrix playing the role of P is not idempotent and Proposition 1 must be replaced by the general expansion $\|A - B - \alpha S\|_F^2 - \|A - B\|_F^2 = \alpha^2 \|S\|_F^2 - 2\alpha \langle A - B, S \rangle_F$ for the correction direction S .
- The counterexamples to the shared-perturbation theorem are small. Example 25 gives a matrix family and a non-commutative family; we have not looked for counterexamples with large or structured weights.

12 Related work

RTN, GPTQ [2], AWQ [3] and QuIP [4] are the layer-wise methods to which QEP is applied in [1]; QuIP# [5] extends QuIP with lattice codebooks. GPTAQ [6] adopts an asymmetric calibration objective of the same form and is the method from which the authors of [1] distinguish their work by the guarantee examined here. Qronos [7], NestQuant [16] and WaterSIC [17] use the same propagated objective, and [15] studies the residual introduced by compensation itself. Error analyses that account for the quantization of earlier layers exist: SPFQ [14] gives full-network error bounds for a stochastic path-following quantizer, and [8] gives error bounds for OPTQ and Qronos and compares the two bounds. These are high-probability or worst-case bounds for specified algorithms; they do not assert that, realization by realization or for the whole network, correction is never worse than no correction, which is the statement examined here. For a single layer, the information-theoretic limits of quantizing for a matrix product are characterized in [18], with nested-lattice constructions, and under an i.i.d. Gaussian weight model at high rate [17] reports a rate gap of about 0.255 bit to the information-theoretic limit. The independence of dithered quantization error from the input goes back to [19]; see [20, 21]. Subtractive dither has been applied to the compression of pretrained network weights [9], and stochastic rounding to low-precision training [10]; neither work compares corrected and uncorrected layer-wise quantization. Treating active and inactive ReLU samples differently in layer-wise reconstruction, with a cascade analysis, goes back to Net-Trim [11]. Quantizing a residual after predicting from an already quantized past is the structure of predictive coding and noise shaping, where the effect of noisy prediction and the limits on feedback gain are classical [12, 13]. Closer still are the classical analyses of finite-word-length effects. In roundoff-noise analysis of digital filters the output noise is computed through Gramians and minimized over realizations [24], and error feedback is designed to reduce it; the noise gain is then a quadratic function of the feedback, with a gain term and a cross term and an optimal strength [25]. The per-layer identity of Section 3.1 has this form and is elementary. The second-moment recursion of the uncorrected scheme is that of a linear system with multiplicative noise, and the propagation of orderings by positive maps is standard in that theory [26]. The

theory of dither is surveyed in [27]. What is specific here is that the correction is a projection determined by the realized activations of the corrected scheme, so that it is neither a fixed feedback filter nor a fixed state transition, and that two trajectories are compared; the step to be proved is that this correction reduces both the error moment and the moment that excites later noise. That error feedback can amplify correlated errors is therefore not new; what we establish is a joint comparison of error and activation moments along two realized trajectories, its exact gain, and the limits imposed by nonlinear geometry and finite codebooks.

13 Conclusion

Under the hypotheses of Theorem 8, compensating inherited quantization error with strengths fixed in advance does not increase the mean-square output error of a linear network. The proof compares the schemes along their own trajectories through both error and activation moments; the exact gain separates direct error removal from reduced excitation of later rounding noise. Stronger compensation is no worse under additional row-space conditions, with a closed-form depth gain in a specified model. Outside that setting we exhibit several ways in which the comparison fails. For ReLU networks, stable activations alone are insufficient; commutators between calibration projections and activation-pattern partitions bound the excess. For fixed perturbations, the per-layer comparison depends on the error cross term. With actual finite codebooks, even unique exact layer-wise minimizers without overload can yield a larger output error. Local improvement is therefore not a network guarantee. All of this holds the calibration input fixed; on two 8B models, the harder part was carrying an improvement on calibration data over to new text.

Use of AI tools. Claude, ChatGPT and Codex assisted in developing counterexamples, parts of the analysis, numerical and Lean code, and manuscript revisions. The mathematical arguments are given in the paper and its appendices; Section 10 states the scope and limitations of the formal checks.

A Proof for Example 14

Setting. Columns are samples. Let

$$X_1 = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \quad W_1 = I_2, \quad W_2 = \begin{pmatrix} 1 & -2 \\ -1 & -1 \end{pmatrix},$$

so that $X_2 = \text{ReLU}(W_1 X_1) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, $W_2 X_2 = \begin{pmatrix} 1 & -2 & -1 \\ -1 & -1 & -2 \end{pmatrix}$ and $X_3 = \text{ReLU}(W_2 X_2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Fix $t > 0$ and let $E_1 = tH$, $E_2 = tG$, where the eight entries of $H = (h_{ij})$, $G = (g_{ij})$ are independent and uniform on $[-1, 1]$. Put $Q_1 = W_1 + E_1$, $Z_2 = \text{ReLU}(Q_1 X_1)$ and

$$Q_2^B = W_2 + E_2, \quad Q_2^Q = W_2 + W_2(X_2 - Z_2)Z_2^\top(Z_2 Z_2^\top)^{-1} + E_2.$$

For either scheme $Y = \text{ReLU}(Q_2 Z_2)$ and $L = \|Y - X_3\|_F^2$. Let $P_t = Z_2^\top(Z_2 Z_2^\top)^{-1}Z_2$, $w = (1, -2)$, $e_1 = (1, 0, 0)^\top$,

$$K = \begin{pmatrix} h_{11} - h_{12} & 0 & h_{11} + h_{12} \\ 0 & -h_{21} + h_{22} & h_{21} + h_{22} \end{pmatrix},$$

$$b = h_{11} - h_{12}, \quad c_t = wK(I - P_t)e_1, \quad c_0 = \frac{1}{3}(-2h_{12} + 4h_{21}).$$

Theorem 23. (a) For $0 < t \leq \frac{1}{10}$ and every realization of (H, G) : all layer-1 pre-activations and all layer-2 pre-activations of both schemes have the strict signs of the teacher's; $Z_2 = X_2 + tK$; $\sigma_{\min}(Z_2) \geq 1 - \sqrt{8}t \geq 1 - 3t > 0$; and Q_1, Q_2^B, Q_2^Q are invertible.

(b) For $0 < t \leq \frac{1}{10}$, $\mathbb{E}L_{\text{QEP}} - \mathbb{E}L_{\text{BASE}} = t^2 \mathbb{E}(c_t^2 - b^2)$, with $\mathbb{E}b^2 = \frac{2}{3}$ and $\mathbb{E}c_0^2 = \frac{20}{27}$.

(c) For $0 < t \leq \frac{1}{10}$ and every H : $\|P_t - P_0\|_2 \leq 4t(1+t)$, $|c_t| \leq 6$, $|c_0| \leq 2$, $|c_t - c_0| \leq 24t(1+t)$.

(d) For $0 < t \leq \frac{1}{10}$,

$$\mathbb{E}L_{\text{QEP}} - \mathbb{E}L_{\text{BASE}} \geq t^2 \left(\frac{2}{27} - 42t(1+t) \right),$$

and the right-hand side is strictly positive for $0 < t \leq \frac{1}{600}$.

Proof. Step 1 (layer-1 gates). $Q_1 X_1 = X_1 + tH X_1$ and row i of $H X_1$ is $(h_{i1} - h_{i2}, -h_{i1} + h_{i2}, h_{i1} + h_{i2})$, with entries of modulus at most 2. The entries of X_1 are ± 1 , so no sign changes when $2t < 1$, and $Z_2 = X_2 + tK$. Write $k_{11} = h_{11} - h_{12}$, $k_{13} = h_{11} + h_{12}$, $k_{22} = h_{22} - h_{21}$, $k_{23} = h_{21} + h_{22}$ (so $|k_{ij}| \leq 2$) and $a = 1 + tk_{11}$, $c = 1 + tk_{13}$, $b' = 1 + tk_{22}$, $d = 1 + tk_{23}$; then $Z_2 = \begin{pmatrix} a & 0 & c \\ 0 & b' & d \end{pmatrix}$ with $a, b', c, d \in [1 - 2t, 1 + 2t]$.

Step 2 (Gram matrix). $\|K\|_F^2 = 2 \sum_{i,j} h_{ij}^2 \leq 8$. Since $X_2 X_2^\top = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ has eigenvalues 1, 3, $\sigma_{\min}(X_2) = 1$, and Weyl's inequality gives $\sigma_{\min}(Z_2) \geq 1 - t\|K\|_2 \geq 1 - \sqrt{8}t$. Hence $Z_2 Z_2^\top$ is invertible and $\|Z_2^\top (Z_2 Z_2^\top)^{-1}\|_2 \leq (1 - \sqrt{8}t)^{-1}$.

Step 3 (layer-2 pre-activations). $Q_2^B Z_2 = W_2 X_2 + tW_2 K + tG Z_2$, and

$$Q_2^Q Z_2 = W_2 Z_2 + W_2 (X_2 - Z_2) P_t + tG Z_2 = W_2 X_2 + tW_2 K (I - P_t) + tG Z_2.$$

Step 4 (rows of $W_2 K$). The rows are $r_1 = wK = (k_{11}, -2k_{22}, k_{13} - 2k_{23})$ and $r_2 = (-k_{11}, -k_{22}, -k_{13} - k_{23})$. The functions $\|r_i\|_2^2$ are convex in H , so they are maximized at vertices of the cube, where $(k_{11}, k_{13}) \in \{(\pm 2, 0), (0, \pm 2)\}$ and independently $(k_{22}, k_{23}) \in \{(\pm 2, 0), (0, \pm 2)\}$. Enumerating the sign choices gives $\|r_1\|_2^2 \leq 36$ and $\|r_2\|_2^2 \leq 16$ at every vertex, so

$$\|wK\|_2 \leq 6, \quad \|r_2\|_2 \leq 4.$$

Step 5 (layer-2 gates). The (i, j) entry of $W_2 K (I - P_t)$ is $r_i (I - P_t) e_j$, of modulus at most $\|r_i\|_2 \leq 6$; the same bound holds for $W_2 K$. The (i, j) entry of $G Z_2$ has modulus at most $z_{1j} + z_{2j} \leq 2 + 4t$. Thus both pre-activations differ from $W_2 X_2$ entrywise by at most $t(8 + 4t) \leq 0.84 < 1$ for $t \leq \frac{1}{10}$, whereas every entry of $W_2 X_2$ has modulus at least 1.

Step 6 (invertibility). $\sigma_{\min}(Q_1) \geq 1 - t\|H\|_F \geq 1 - 2t > 0$. The eigenvalues of $W_2 W_2^\top = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}$ are $\frac{1}{2}(7 \pm \sqrt{13})$, so $\sigma_{\min}(W_2) \geq 1.3$ and $\|W_2\|_2 \leq 2.305$; hence $\sigma_{\min}(Q_2^B) \geq 1.3 - 2t > 0$. Finally $Q_2^Q = W_2 - tW_2 K Z_2^\top (Z_2 Z_2^\top)^{-1} + tG$, and by Step 2 the perturbation has norm at most $t(2.305\sqrt{8}/(1 - \sqrt{8}t) + 2) \leq 1.11 < 1.3$ for $t \leq \frac{1}{10}$. This proves (a).

Step 7 (the loss is one scalar). By Step 5, Y has the support of X_3 , so $L = (Y_{11} - 1)^2$ with Y_{11} the $(1, 1)$ pre-activation. As $z_{21} = 0$ and $z_{11} = a$, Step 3 gives

$$Y_{11}^B - 1 = t(b + g_{11}a), \quad Y_{11}^Q - 1 = t(c_t + g_{11}a).$$

Step 8 (cancellation of the last-layer error). b, c_t, a depend on H only and g_{11} is independent of H with mean zero, so for $\xi \in \{b, c_t\}$, $\mathbb{E}(\xi + g_{11}a)^2 = \mathbb{E}\xi^2 + \frac{1}{3}\mathbb{E}a^2$. Subtracting gives the identity in (b).

Step 9 ($t = 0$). The orthogonal complement of the row space of X_2 is spanned by $m_0 = (1, 1, -1)^\top$, so $I - P_0 = \frac{1}{3}m_0m_0^\top$ and $c_0 = \frac{1}{3}(k_{11} - 2k_{22} - k_{13} + 2k_{23}) = \frac{1}{3}(-2h_{12} + 4h_{21})$. With $\mathbb{E}h_{ij}^2 = \frac{1}{3}$: $\mathbb{E}b^2 = \frac{2}{3}$, $\mathbb{E}c_0^2 = \frac{4+16}{27} = \frac{20}{27}$, and $|c_0| \leq 2$.

Step 10 (projection bound). $m = (cb', ad, -ab')^\top$ is nonzero and orthogonal to both rows of Z_2 , so $I - P_t = mm^\top / \|m\|_2^2$. For unit vectors u, v spanning lines at angle $\theta \in [0, \frac{\pi}{2}]$, $\|uu^\top - vv^\top\|_2 = \sin \theta$; therefore

$$\|P_t - P_0\|_2 = \sin \angle(m, m_0) = \frac{\text{dist}(m_0, \text{span}\{m\})}{\|m_0\|_2} \leq \frac{\|m - m_0\|_2}{\sqrt{3}}.$$

Each coordinate of $m - m_0$ is $\pm((1 + t\alpha)(1 + t\beta) - 1)$ with $|\alpha|, |\beta| \leq 2$, of modulus at most $4t + 4t^2$. Hence $\|P_t - P_0\|_2 \leq 4t(1 + t)$.

Step 11. $|c_t| \leq \|wK\|_2 \|(I - P_t)e_1\|_2 \leq 6$ and $|c_t - c_0| = |wK(P_0 - P_t)e_1| \leq 6 \cdot 4t(1 + t) = 24t(1 + t)$. This proves (c).

Step 12 (lower bound). With $\delta = c_t - c_0$, $c_t^2 - c_0^2 = 2c_0\delta + \delta^2 \geq -2|c_0||\delta|$, so

$$\mathbb{E}c_t^2 - \mathbb{E}c_0^2 \geq -48t(1 + t)\mathbb{E}|c_0| \geq -48t(1 + t)\sqrt{20/27} \geq -42t(1 + t),$$

because $\sqrt{20/27} \leq \frac{7}{8}$. By Steps 8–9, $\mathbb{E}L_{\text{QEP}} - \mathbb{E}L_{\text{BASE}} = t^2(\frac{2}{27} + \mathbb{E}c_t^2 - \mathbb{E}c_0^2) \geq t^2(\frac{2}{27} - 42t(1 + t))$. Since $t \mapsto 42t(1 + t)$ is increasing and equals $\frac{42 \cdot 601}{360000} < 0.0702 < \frac{2}{27}$ at $t = \frac{1}{600}$, the bound is positive on $(0, \frac{1}{600}]$. $\square$

Remark 24. (i) Since $4t(1 + t) \leq 12t$ and $6 \leq \sqrt{40}$, the cruder estimates $\|P_t - P_0\|_2 \leq 12t$, $|c_t^2 - c_0^2| \leq 960t$ and positivity for $0 < t \leq 10^{-5}$ also hold. (ii) Apart from Steps 8 and 9 the argument is deterministic and uses only $|h_{ij}|, |g_{ij}| \leq 1$; Step 8 uses that G is independent of H with $\mathbb{E}g_{11} = 0$, and Step 9 uses the second moments of the h_{ij} . If the h_{ij} are i.i.d. with mean zero and second moment s , the leading coefficient is $\frac{20s}{9} - 2s = \frac{2s}{9}$ (so $\frac{2}{9}$ for symmetric signs ± 1). (iii) $c_t(H) = -c_{-t}(-H)$, so for a symmetric law $t \mapsto \mathbb{E}c_t^2$ is even and the true correction is $O(t^2)$; numerically $\mathbb{E}(c_t^2 - b^2) \approx \frac{2}{27} - 0.283t^2$ and stays positive on all of $(0, \frac{1}{10}]$; this is a numerical observation. The interval $(0, \frac{1}{600}]$ is the proved one and is not sharp.

B Further counterexamples in the shared-perturbation model

Example 25 (Not an artifact of $P = I$). Take one output, one input dimension and two calibration samples: $W = 1$, $\hat{X} = (1, 0)$, $X = (2, 1)$, $E = 1$, $\alpha = 1$. Then $A := W\delta = (1, 1)$, $P = \text{diag}(1, 0)$, $B := E\hat{X} = (1, 0)$. The QEP residual is $A(I - P) - B = (-1, 1)$ with squared norm 2; the BASE residual on the same input is $A - B = (0, 1)$ with squared norm 1.

The counterexample of Theorem 17 also lifts to matrices. Let $d \geq 2$, $L = 2$, identity activations, $W_1 = W_2 = I_d$, $X = [I_d \mid \mathbf{1}] \in \mathbb{R}^{d \times (d+1)}$, $\alpha_1 = \alpha_2 = 1$, and $E_1 = D$, $E_2 = -D$ with $D = \text{diag}(\varepsilon_1, \dots, \varepsilon_d)$, $0 < \varepsilon_i < \frac{1}{2}$. Both schemes use $\hat{W}_1 = I + D$ and $\hat{X}_2 = (I + D)X$, which has full row rank, and $P_2 \neq I_{d+1}$. The correction is $-D(I + D)^{-1}$, so $\hat{W}_2^{\text{BASE}} = I - D$ and $\hat{W}_2^{\text{QEP}} = (I + D)^{-1} - D$, all invertible, and

$$R_1 = -DX, \quad R_2^{\text{BASE}} = D^2X, \quad R_2^{\text{QEP}} = (D + D^2)X.$$

Hence $U^{\text{QEP}} > U^{\text{BASE}}$; for $D = \varepsilon I$ the gap is exactly $\varepsilon\|X\|_F$. The mechanism is the scalar one applied coordinate by coordinate.

A family that does not reduce to independent scalar channels: let $N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $D = \text{diag}(a, b)$ with $a > 2b > 0$, $W_1 = I$, $W_2 = D$, $X = I$, $\alpha = (1, 1)$, and $E_1 = tN$, $E_2 = t(ND - DN) = (b - a)tN$ with $t > 0$. Both schemes have $\hat{X}_2 = I + tN$. The outputs are $D + btN$ for BASE and $D + (b - a)tN$

for QEP, so the output errors are bt and $(a - b)t$, and $U^{\text{BASE}} = (a + b)t$, $U^{\text{QEP}} = (2a - b)t$. All weights and Gram matrices are invertible for every t . Here E_2 is a commutator; when $a = b$ it vanishes, and QEP is exact while BASE has error bt .

Example 26 (The trajectory substitution fails independently). Take $L = 3$, scalars, identity activations, $W = (1, 1, 1)$, $X = 1$, $\alpha = (1, 1, 1)$ and $E(t) = t(\frac{1}{2}, -\frac{1}{10}, -\frac{7}{27})$ with $0 < t \leq 1$. All quantized weights are positive. On QEP's own inputs, the QEP residual is no larger in magnitude than the BASE residual evaluated on the same inputs, at every layer: the per-layer comparison on a common input holds throughout. Yet on each scheme's own trajectory the third-layer residuals satisfy

$$|R_3^{\text{QEP}}| - |R_3^{\text{BASE}}| = \frac{t(64 + 69t - 14t^2)}{540} > 0,$$

with a positive first-order coefficient $16/135$. At $t = 1$ the BASE residual is 0 and the QEP residual is $119/540$. A layer-wise comparison on common inputs therefore does not transfer to the two actual trajectories, even to first order.

Theorem B.5 of [1] also asserts the composite inequality $\|\delta_{L+1}^{\text{QEP}}\|_F \leq U^{\text{BASE}}$, which compares the actual QEP output error with the BASE bound. In Theorem 17 it holds with equality. As an exact statement it fails as well, although not to first order in the example we know.

Example 27 (The composite inequality fails for finite perturbations). Take $L = 3$, scalars, identity activations, $W = (1, 1, 1)$, $X = 1$, $\alpha = (1, 1, 1)$, and $E = (-\frac{1}{2}, -\frac{1}{2}, 3)$. The BASE weights are $(\frac{1}{2}, \frac{1}{2}, 4)$, with output 1 and residuals $(\frac{1}{2}, \frac{3}{4}, 0)$, so $U^{\text{BASE}} = \frac{5}{4}$. The QEP weights are $(\frac{1}{2}, \frac{3}{2}, \frac{13}{3})$, with output $\frac{13}{4}$ and output error $\frac{9}{4} > U^{\text{BASE}}$. Along the ray $E(t) = t(-\frac{1}{2}, -\frac{1}{2}, 3)$, $0 < t \leq 1$, the difference between the QEP output error and U^{BASE} equals $t(3t - 1)/2$, which is positive only for $t > \frac{1}{3}$; the first-order coefficients are 3 and $\frac{7}{2}$, so the inequality holds to first order along this ray.

A repaired composite bound. In the shared-perturbation model, put $w_l = \|W_l\|_2$ and $e_l = \|E_l\|_2$, and let

$$\tilde{U}^{\text{BASE}} = \sum_{k=1}^L \gamma_k \|R_k^{\text{BASE}}\|_F \prod_{s=k+1}^L \gamma_s (w_s + e_s).$$

For Lipschitz activations and strengths in $[0, 1]$, whenever the corrected recursion is defined,

$$\|\delta_{L+1}^{\text{QEP}}\|_F \leq \tilde{U}^{\text{BASE}}.$$

Only the Gram inverses at positive-strength layers are required; the weights need not be invertible. For a fixed finite teacher and input, if $e_l(t) \leq t f_l$ with fixed finite f_l , then, uniformly over defined corrected trajectories and strengths,

$$\begin{aligned} 0 &\leq \tilde{U}^{\text{BASE}}(t) - U^{\text{BASE}}(t) \leq Ct^2, \\ \|\delta_{L+1}^{\text{QEP}}(t)\|_F &\leq U^{\text{BASE}}(t) + Ct^2. \end{aligned}$$

for some finite C and $0 < t \leq t_0$. This is an additive remainder, not an exact comparison with the original bound.

Proof of the repaired composite bound. Write $Z_l = \hat{X}_l^{\text{QEP}}$, $Y_l = \hat{X}_l^{\text{BASE}}$, $B_l = X_l - Y_l$, $D_l = Z_l - Y_l$, and $b_l = \|B_l\|_F$, $d_l = \|D_l\|_F$, $h_l = b_l + d_l$. Then $X_l - Z_l = B_l - D_l$, $h_1 = 0$ and $\|X_l - Z_l\|_F \leq h_l$. Let

P_l be the row-space projection of Z_l when $\alpha_l > 0$, and set $P_l = 0$ otherwise. Both P_l and $I - \alpha_l P_l$ have norm at most one. The exact difference of the pre-activations is

$$\widehat{W}_l^{\text{QEP}} Z_l - \widehat{W}_l^{\text{BASE}} Y_l = W_l D_l (I - \alpha_l P_l) + E_l D_l + \alpha_l W_l B_l P_l.$$

The Lipschitz property therefore gives

$$\begin{aligned} d_{l+1} &\leq \gamma_l [(w_l + e_l) d_l + \alpha_l w_l b_l], \\ b_{l+1} &\leq \gamma_l \|R_l^{\text{BASE}}\|_F. \end{aligned}$$

Since $\alpha_l w_l \leq w_l + e_l$, addition yields

$$h_{l+1} \leq \gamma_l \|R_l^{\text{BASE}}\|_F + \gamma_l (w_l + e_l) h_l.$$

Unrolling from $h_1 = 0$ proves the claimed bound. The residuals and the projection have stayed on their respective trajectories throughout.

For the remainder, on the uncorrected trajectory $R_l^{\text{BASE}}(t) = (W_l + E_l(t))(X_l - Y_l(t)) - E_l(t)X_l$, so

$$\|R_l^{\text{BASE}}(t)\|_F \leq (w_l + t f_l) b_l(t) + t f_l \|X_l\|_F.$$

Together with $b_{l+1}(t) \leq \gamma_l \|R_l^{\text{BASE}}(t)\|_F$ and $b_1 = 0$, finite induction gives $b_l(t) = O(t)$ and $\|R_l^{\text{BASE}}(t)\|_F = O(t)$ uniformly. Each downstream product changes by $O(t)$ when its factors w_s are replaced by $w_s + e_s(t)$: expand the finite product, with $0 \leq e_s(t) \leq t f_s$. This does not divide by w_s and also covers zero weights. Multiplying the product differences by the $O(t)$ residuals and summing proves the nonnegative $O(t^2)$ remainder. $\square$

Thus the composite inequality of [1] is valid to first order in this precise sense, unlike the ordering of the two original bounds. For Example 27, $\tilde{U}^{\text{BASE}} = 6$.

References

- [1] Y. Arai and Y. Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.09629.
- [2] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.
- [3] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, 2024.
- [4] J. Chee, Y. Cai, V. Kuleshov, and C. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [5] A. Tseng, J. Chee, Q. Sun, V. Kuleshov, and C. De Sa. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning (ICML)*, PMLR 235:48630–48656, 2024.

- [6] Y. Li, R. Yin, D. Lee, S. Xiao, and P. Panda. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *International Conference on Machine Learning (ICML)*, PMLR 267:36690–36706, 2025. arXiv:2504.02692.
- [7] S. Zhang, H. Zhang, I. Colbert, and R. Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11695.
- [8] H. Zhang, S. Zhang, I. Colbert, and R. Saab. Provable post-training quantization: Theoretical analysis of OPTQ and Qronos. arXiv:2508.04853 (version 2, April 2026).
- [9] Y. Choi, M. El-Khamy, and J. Lee. Universal deep neural network compression. arXiv:1802.02271, 2018.
- [10] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning (ICML)*, 2015.
- [11] A. Aghasi, A. Abdi, N. Nguyen, and J. Romberg. Net-Trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1611.05162.
- [12] R. Zamir, Y. Kochman, and U. Erez. Achieving the Gaussian rate–distortion function by prediction. *IEEE Transactions on Information Theory*, 54(7):3354–3364, 2008.
- [13] O. Ordentlich and U. Erez. Performance analysis and optimal filter design for sigma-delta modulation via duality with DPCM. arXiv:1501.01829.
- [14] J. Zhang and R. Saab. SPFQ: A stochastic algorithm and its error analysis for neural network quantization. arXiv:2309.10975, 2023.
- [15] S. Li, J. Deng, K. Xu, R. Deng, H. Gu, M. Jiang, H. Shen, and K. Huang. Rethinking residual errors in compensation-based LLM quantization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2604.07955.
- [16] S. Savkin, E. Porat, O. Ordentlich, and Y. Polyanskiy. NestQuant: Nested lattice quantization for matrix products and LLMs. In *International Conference on Machine Learning (ICML)*, PMLR 267:53042–53062, 2025.
- [17] E. Lifar, S. Savkin, O. Ordentlich, and Y. Polyanskiy. WaterSIC: Information-theoretically (near) optimal linear layer quantization. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.04956.
- [18] O. Ordentlich and Y. Polyanskiy. Optimal quantization for matrix multiplication. *IEEE Transactions on Information Theory*, 72(3):1943–1972, 2026.
- [19] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, 1964.
- [20] R. M. Gray and T. G. Stockham Jr. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.
- [21] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, 1996.

- [22] L. B. Allal et al. SmolLM2: When smol goes big — data-centric training of a small language model. arXiv:2502.02737, 2025.
- [23] S. Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.
- [24] C. T. Mullis and R. A. Roberts. Synthesis of minimum roundoff noise fixed point digital filters. *IEEE Transactions on Circuits and Systems*, 23(9):551–562, 1976.
- [25] T. Hinamoto, H. Ohnishi, and W.-S. Lu. Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation optimization. *IEEE Transactions on Circuits and Systems I*, 50(1):23–33, 2003.
- [26] B. Bamieh and M. Filo. An input–output approach to structured stochastic uncertainty. arXiv:1806.07473, 2018.
- [27] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy. Quantization and dither: A theoretical survey. *Journal of the Audio Engineering Society*, 40(5):355–375, 1992.
- [28] Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu. BRECCQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2102.05426.
- [29] W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, P. Gao, Y. Qiao, and P. Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.13137.