Optimal Scalar Quantization for Matrix Multiplication: Closed-Form Density and Phase Transition

Calvin Ang[*] Stanford University

Sungyoon Kim[*] Stanford University

Mert Pilanci Stanford University

Abstract

We study entrywise scalar quantization of two matrices prior to multiplication. Given $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{k\times n}$, we quantize entries of A and B independently using scalar quantizers with K_X and K_Y levels per entry, and form $\widehat{C} = \widehat{A}\,\widehat{B}$. The objective is to minimize the matrix multiplication mean-squared error (MSE) $\mathcal{E} \triangleq \mathbb{E}[\|AB - \widehat{A}\widehat{B}\|_F^2]$ under a pair-i.i.d. inner-product model. In the high-resolution regime $K_X, K_Y \to \infty$, we derive a sharp K^{-2} asymptotic expansion for $\mathcal{E}$, identify the exact optimal leading constants, and characterize asymptotically optimal quantization center densities in terms of conditional second moments. We then specialize to correlated Gaussian multiplicative pairs, obtaining a closed-form optimal point density

$$\lambda^\star(u) \;\propto\; \exp\!\left(-\frac{u^2}{6}\right)\left((1-\rho^2)+\rho^2u^2\right)^{1/3}, \qquad u = \frac{x}{\sigma_X},$$

with the same form for y/σ_Y, and prove a correlation-driven phase transition: the density is unimodal at the origin for $|\rho| \le 1/\sqrt{3}$ and becomes bimodal for $|\rho| > 1/\sqrt{3}$ with peaks at $u_{\mathrm{peak}} = \pm\sqrt{3 - 1/\rho^2}$. We show our method's applicability in synthetic experiments such as matrix multiplication quantization and least squares optimization, as well as quantization of large language model key and query activations.

1 Introduction

Quantized matrix multiplication is central in modern machine learning inference and scientific computing. In many deployments, one does *not* care about entrywise reconstruction error of operands A and B; rather, the relevant distortion is the error induced *after multiplication*. This perspective is closely connected to classical high-rate theory [1] and to the broader quantization literature [4], which show that optimal quantizer design depends fundamentally on the downstream distortion criterion. At the same time, low-precision quantization has become a practical necessity in modern GPU inference: recent hardware and software stacks increasingly rely on ultra-low-precision formats such as NVFP4 [9], INT8 [2], and FP8 [8] to improve throughput, memory efficiency, and energy efficiency for large-scale matrix multiplication workloads arising in large language model inference . These trends make it especially important to understand quantization schemes that are optimal for the *product* itself, rather than for the separate reconstruction of the input matrices. However, matrix multiplication introduces a distinct bilinear distortion structure: the error contributed by one operand is filtered by the other, and the importance of each entry depends jointly on both factors. Related recent work has studied quantized matrix multiplication in different settings, including nested lattice quantization [10]. In contrast, our focus is on the optimal design of scalar quantizers specifically

[*]Equal contribution.

Preprint. Under review.

for matrix multiplication, deriving high-rate laws tailored to product distortion and showing how multiplicative structure and statistical dependence fundamentally change the optimal point density.

This paper asks:

> *Which scalar quantizers for entries of A and B minimize the expected Frobenius MSE of the product computed from quantized operands?*

Contributions.

- We derive a high-rate characterization of the optimal achievable matrix multiplication MSE, proving a sharp K^{-2} scaling law with an *exact* leading constant.
- We reduce the matrix objective to two *weighted scalar* MSE quantization problems driven by conditional second moments, and we identify the optimal companding point densities.
- For correlated Gaussian multiplicative pairs, we obtain closed-form optimal densities and prove a correlation-induced unimodal-to-bimodal "phase transition".
- Provide experimental validation of our results on synthetically generated matrices for matrix multiplication and quantized least squares as well as on key and query activations on the GPT-2 and Qwen3 family of models.

2 High-Rate Analysis of Matrix Multiplication MSE

This section derives the high-rate reduction from matrix multiplication MSE to two decoupled weighted scalar criteria, and states the resulting sharp K^{-2} constant. All asymptotic steps are justified in Appendix A and Appendix B.

2.1 Problem Formulation

Let $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$, and $C = AB$. Let Q_X and Q_Y be scalar quantizers with $|\mathrm{range}(Q_X)| \le K_X$ and $|\mathrm{range}(Q_Y)| \le K_Y$. Define

$$\widehat{A}_{i\ell} = Q_X(A_{i\ell}), \qquad \widehat{B}_{\ell j} = Q_Y(B_{\ell j}), \qquad \widehat{C} = \widehat{A}\,\widehat{B}.$$

We measure performance by the matrix multiplication MSE

$$\mathcal{E}(Q_X, Q_Y) \triangleq \mathbb{E}\big[\,\big\|C - \widehat{C}\big\|_{\mathrm{F}}^2\,\big].$$

The quantizer design is governed by the distribution of multiplicative entry-pairs feeding each inner product. We assume each pairs (X, Y) in the matrix multiplication has identical joint distributions.

Assumption 1 (Pair-i.i.d. inner products)**.** *For each output entry $C_{ij} = \sum_{\ell=1}^{k} A_{i\ell}B_{\ell j}$, the pairs $\{(A_{i\ell}, B_{\ell j})\}_{\ell=1}^{k}$ are i.i.d. copies of a generic pair (X, Y) with finite moments up to order $4+\epsilon$ for some $\epsilon > 0$. Because of this, the entries $\{C_{ij}\}_{i\in[m],j\in[n]}$ are identically distributed (although dependent).*

Under Assumption 1, when we denote $(X, Y) \sim \mathcal{D}$ we have

$$\mathcal{E}(Q_X, Q_Y) = nm\,\mathbb{E}_{(X_i,Y_i)\sim\mathcal{D}}\Big[\big(\sum_{i=1}^{k} X_iY_i - \sum_{i=1}^{k} \hat{X}_i\hat{Y}_i\big)^2\Big].$$

We use the standard fixed-rate identification

$$R_X \triangleq \log_2 K_X, \qquad R_Y \triangleq \log_2 K_Y,$$

and analyze the high-resolution regime $K_X, K_Y \to \infty$ (or equivalently, $R_X, R_Y \to \infty$).

2.2 Sharp High-rate Constant and Optimal Point Densities

Let $D \triangleq XY - \widehat{X}\widehat{Y}$ for a single pair (X, Y). Independence across samples (X_i, Y_i) yields

$$\mathbb{E}_{(X_i,Y_i)\sim\mathcal{D}}\big[(\sum_{i=1}^{k} X_iY_i - \sum_{i=1}^{k} \hat{X}_i\hat{Y}_i)^2\big] = k\,\mathbb{E}[D^2] + k(k-1)\mathbb{E}[D]^2. \tag{1}$$

Appendix A proves that, for the companding quantizers used in the high-rate design, $\mathbb{E}[D] = O(K_X^{-2} + K_Y^{-2})$. Consequently the second term in (1) is $o(K_X^{-2} + K_Y^{-2})$ and does not affect the K^{-2} leading constant.

Define quantization errors $e_X \triangleq X - \widehat{X}$ and $e_Y \triangleq Y - \widehat{Y}$. Then

$$D = XY - \widehat{X}\widehat{Y} = Ye_X + Xe_Y - e_Xe_Y. \tag{2}$$

Squaring produces

$$D^2 = Y^2e_X^2 + X^2e_Y^2 + e_X^2e_Y^2 + 2XYe_Xe_Y - 2Ye_X^2e_Y - 2Xe_Xe_Y^2. \tag{3}$$

Appendix A shows that the mixed terms on the second line have expectation $O(K_X^{-2}K_Y^{-2})$ for the companding quantizers used in the high-rate design. Since $K_X^{-2}K_Y^{-2} = o(K_X^{-2} + K_Y^{-2})$ as $K_X, K_Y \to \infty$, we obtain the sharp leading reduction

$$\mathbb{E}[D^2] = \mathbb{E}[Y^2e_X^2] + \mathbb{E}[X^2e_Y^2] + o(K_X^{-2} + K_Y^{-2}). \tag{4}$$

Because e_X is a deterministic function of X and e_Y is a deterministic function of Y, conditioning yields

$$\mathbb{E}[Y^2e_X^2] = \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\, e_X^2\big], \qquad \mathbb{E}[X^2e_Y^2] = \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\, e_Y^2\big].$$

Define the conditional second moments as

$$w_X(x) \triangleq \mathbb{E}[Y^2 \mid X = x], \qquad w_Y(y) \triangleq \mathbb{E}[X^2 \mid Y = y]. \tag{5}$$

Then (1) and (4), along with the fact that $\mathbb{E}[D] = O(K_X^{-2} + K_Y^{-2})$ imply the two-scalar high-rate reduction

$$\begin{aligned}\mathcal{E}(Q_X, Q_Y) = mnk\Big(&\mathbb{E}[w_X(X)(X - Q_X(X))^2] + \mathbb{E}[w_Y(Y)(Y - Q_Y(Y))^2]\Big) \\ &+ o(K_X^{-2} + K_Y^{-2}).\end{aligned} \tag{6}$$

Now let's denote f_X and f_Y the marginal densities of X and Y, and further assume a regularity condition. These conditions are used to justify the asymptotic analysis and to derive weighted optimal rates.

Assumption 2 (High-rate regularity)**.** *The pair (X, Y) admits a joint density $f_{X,Y}$ that is twice continuously differentiable. The functions f_X, f_Y, w_X, w_Y are continuous and strictly positive on $\mathbb{R}$, and*

$$\int_{\mathbb{R}} (f_X(x)w_X(x))^{1/3}\,\mathrm{d}x < \infty, \qquad \int_{\mathbb{R}} (f_Y(y)w_Y(y))^{1/3}\,\mathrm{d}y < \infty.$$

Moreover, the conditional mean functions $\mu_{Y|X}(x) \triangleq \mathbb{E}[Y \mid X = x]$ and $\mu_{X|Y}(y) \triangleq \mathbb{E}[X \mid Y = y]$ are continuously differentiable and the integrability conditions used in Appendix A and Appendix B hold.[2]

Now denote the integrals in Assumption 2 as

$$I_X \triangleq \int_{-\infty}^{\infty} (f_X(x)w_X(x))^{1/3}\,\mathrm{d}x, \qquad I_Y \triangleq \int_{-\infty}^{\infty} (f_Y(y)w_Y(y))^{1/3}\,\mathrm{d}y. \tag{7}$$

An application of Cauchy-Schwartz gives the optimal high-rate quantizer for matrix multiplication.

[2]These conditions are mild for common smooth models (e.g., jointly Gaussian pairs) and are stated explicitly where they are used.

Theorem 1 (High-rate optimal matrix multiplication MSE)**.** *Under Assumptions 1–2,*

$$\inf_{Q_X,Q_Y} \mathcal{E}(Q_X,Q_Y) = mnk\left(\frac{I_X^3}{12\,K_X^2}+\frac{I_Y^3}{12\,K_Y^2}\right)+o\!\left(\frac{1}{K_X^2}+\frac{1}{K_Y^2}\right). \tag{8}$$

Moreover, asymptotically optimal K_X- and K_Y-level quantizers are companding quantizers with point densities

$$\lambda_X^\star(x)=\frac{(f_X(x)w_X(x))^{1/3}}{I_X}, \qquad \lambda_Y^\star(y)=\frac{(f_Y(y)w_Y(y))^{1/3}}{I_Y}. \tag{9}$$

Proof. Appendix A proves the decoupling (6) with a remainder $o(K_X^{-2}+K_Y^{-2})$ and shows that the bias term in (1) is negligible at the K^{-2} scale. Appendix B proves a weighted scalar high-rate theorem (including both achievability and converse) and identifies the unique optimizing point densities. Combining these two appendices yields (8) and (9). □

Corollary 1 (Rate form and optimal bit split)**.** *Let $R_X=\log_2 K_X$ and $R_Y=\log_2 K_Y$. Then the leading term in* (8) *can be written as*

$$\inf_{Q_X,Q_Y} \mathcal{E}(Q_X,Q_Y) = mnk\left(\alpha_X\,2^{-2R_X}+\alpha_Y\,2^{-2R_Y}\right)+o(2^{-2R_X}+2^{-2R_Y}),$$

where $\alpha_X \triangleq I_X^3/12$ and $\alpha_Y \triangleq I_Y^3/12$. If the pair rate *$R \triangleq R_X+R_Y$ is fixed and large, then the minimizer of the leading term satisfies*

$$R_X^\star=\frac{R}{2}+\frac{1}{4}\log_2\!\left(\frac{\alpha_X}{\alpha_Y}\right), \qquad R_Y^\star=\frac{R}{2}-\frac{1}{4}\log_2\!\left(\frac{\alpha_X}{\alpha_Y}\right),$$

equivalently $K_X/K_Y=(\alpha_X/\alpha_Y)^{1/4}$.

2.3 Special Case: Correlated Gaussian

Assume the multiplicative pair (X,Y) is bivariate Gaussian:

$$\begin{bmatrix}X\\Y\end{bmatrix} \sim \mathcal{N}\!\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right), \qquad \rho\in(-1,1). \tag{10}$$

Applying Theorem 1 to the Gaussian case enables a specialization to Gaussian joint densities. See Appendix C for a proof.

Corollary 2 (Closed-form asymptotically optimal point density)**.** *Under* (12)*, the optimal companding point densities in* (9) *take the closed form*

$$\lambda_X^\star(x)=\frac{\exp\!\left(-\frac{x^2}{6\sigma_X^2}\right)\left((1-\rho^2)+\rho^2\frac{x^2}{\sigma_X^2}\right)^{1/3}}{\displaystyle\int_{-\infty}^{\infty}\exp\!\left(-\frac{t^2}{6\sigma_X^2}\right)\left((1-\rho^2)+\rho^2\frac{t^2}{\sigma_X^2}\right)^{1/3}\mathrm{d}t},$$

$$\lambda_Y^\star(y)=\frac{\exp\!\left(-\frac{y^2}{6\sigma_Y^2}\right)\left((1-\rho^2)+\rho^2\frac{y^2}{\sigma_Y^2}\right)^{1/3}}{\displaystyle\int_{-\infty}^{\infty}\exp\!\left(-\frac{t^2}{6\sigma_Y^2}\right)\left((1-\rho^2)+\rho^2\frac{t^2}{\sigma_Y^2}\right)^{1/3}\mathrm{d}t}.$$

If we write the normalized variable as $u=x/\sigma_X$, we have

$$\lambda^\star(u)\ \propto\ \exp\!\left(-\frac{u^2}{6}\right)\left((1-\rho^2)+\rho^2u^2\right)^{1/3}.$$

The density has a qualitative difference compared to Gaussian companders: specifically, the optimal λ^* goes through a phase transition as ρ increases: at first it has a single mode, but as ρ increases two modes appear. See Fig. 1 for a schematic, and Appendix D for a proof.

Theorem 2 (Unimodal-to-bimodal transition at $|\rho|=1/\sqrt{3}$)**.** *Let $\lambda^\star(u)$ be the optimal density shape above.*

1. *If $\rho^2 < 1/3$, then $\lambda^\star(u)$ is unimodal with a unique global maximum at $u = 0$.*
2. *If $\rho^2 > 1/3$, then $\lambda^\star(u)$ is bimodal: $u = 0$ is a strict local minimum and the two symmetric global maxima occur at*
$$u_{\text{peak}} = \pm\sqrt{3 - \frac{1}{\rho^2}}.$$
3. *If $\rho^2 = 1/3$, the curvature at $u = 0$ vanishes (critical splitting point).*

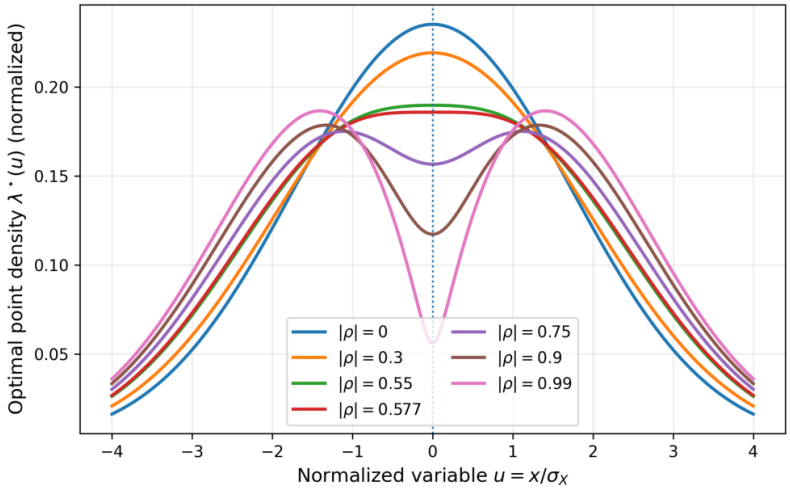

Figure 1: Optimal density phase transition. When $\rho = 0$, there is only a single mode. As ρ increases, an additional mode emerges at the critical value $|\rho| = 1/\sqrt{3}$.

At last, direct substitution gives the high-rate loss for the proposed matmul quantization.

Corollary 3 (Gaussian high-rate constant for matrix multiplication MSE)**.** *Let $K_X = K_Y = K$. Define*
$$J(\rho) \triangleq \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{6}\right)\left((1-\rho^2) + \rho^2 u^2\right)^{1/3} \mathrm{d}u.$$
Then
$$\lim_{K\to\infty} K^2 \inf_{Q_X, Q_Y} \mathbb{E}\left[\left\|AB - \widehat{A}\widehat{B}\right\|_{\mathrm{F}}^2\right] = \frac{mkn\,\sigma_X^2\sigma_Y^2}{6\sqrt{2\pi}}\, J(\rho)^3.$$
A closed form for $J(\rho)$ is given in Appendix E.

3 Experiments

We now empirically evaluate the performance of our derived quantization method with a variety of other state-of-the-art quantization methods on synthetic and real-world data.

3.1 Synthetic Experiments: Matrix Multiplication

First, we compare the performance of our optimal matrix multiplication quantizer with various commonly used quantizers for synthetically generated matrices that conform to our correlated Gaussian model. For this experiment, we test a variety of ρ values and compare the quantization error in relative Frobenius norm for the matrices.

3.1.1 Alternate Quantizer Descriptions

1. **MatMul-Opt (Ours).** The theory-optimal companding quantizer derived from Corollary 2 for the correlated Gaussain case. Bin boundaries $\{t_i\}$ are placed by inverting the CDF of the optimal point density
$$\lambda^*(u) \propto \exp\left(-\tfrac{u^2}{6}\right)\left[(1-\rho^2) + \rho^2 u^2\right]^{1/3}, \qquad u = x/\sigma, \tag{11}$$

so that equal probability mass of * falls in each bin. Separate quantizers are built for A and B using their empirical standard deviations $\hat{\sigma}_A$ and $\hat{\sigma}_B$ and the known correlation ρ.

2. **Gaussian Compander.** A companding quantizer whose point density is the $\rho = 0$ special case of the optimal density from standard scalar high-rate theory.

3. **Lloyd-Max.** The iteratively computed optimal scalar quantizer for a Gaussian source $X \sim \mathcal{N}(0, \sigma^2)$ [6]. Starting from quantile-midpoint initialization, the algorithm alternates between (i) setting boundaries to midpoints of adjacent reconstruction levels and (ii) setting levels to the conditional mean of $|phi$ over each bin, until convergence in ℓ^∞ norm up to 200 iterations.

4. **Uniform.** A symmetric uniform quantizer with K evenly-spaced reconstruction levles on $[-c, c]$. The clip value c is selected by a grid search over $c \in [1.5\hat{\sigma},\ 5.0\hat{\sigma}]$ (36 points) to minimize empirical MSE on entries of the matrix being quantized. Separate clip values are calibrated for A and B.

5. **μ-Law.** The ITU-T G.711 μ-law compander [5] with $\mu = 255$, adapted to continuous-valued inputs. The companding function $f(x) = \operatorname{sgn}(x)\, \frac{\ln(1+\mu|x/c|)}{\ln(1+\mu)}$ maps inputs to a uniform grid of K levels in the logarithmically compressed domain; the inverse map recovers reconstruction levels in the original domain. The clip is fixed at $c = 4\hat{\sigma}$. μ-law compression places more quantization levels near zero, which improves SNR for speech-like signals but is not matched to the matrix-multiplication task.

6. **A-Law.** The ITU-T G.711 A-law compander [5] with $A = 87.6$, also adapted to continuous inputs. The two-piece companding function is linear for $|x/c| < 1/A$ and logarithmic otherwise:

$$f(x) = \operatorname{sgn}(x) \cdot \begin{cases} \dfrac{A\,|x/c|}{1 + \ln A} & |x/c| < \frac{1}{A}, \\ \dfrac{1 + \ln(A\,|x/c|)}{1 + \ln A} & \frac{1}{A} \le |x/c| \le 1. \end{cases}$$

   K uniform levels are placed in the companded domain with clip $c = 4\hat{\sigma}$. A-law and μ-law share the same logarithmic motivation and serve as classical telecommunications baselines.

7. **NF4 (Normal Float 4-bit).** A K-level fixed codebook whose points are the quantile midpoints of $\mathcal{N}(0, 1)$: $c_i = \Phi^{-1}\big((i + \frac{1}{2})/K\big)$ for $i = 0, \ldots, K-1$, normalized so that $\max_i |c_i| = 1$, with endpoints clamped to ± 1. This is an approximation of the QLoRA NF4 codebook [3], which places levels at Gaussian quantile midpoints to minimize MSE for normally-distributed weights. A per-matrix scalar scale is selected by grid search over $[0.5\hat{\sigma},\ 4.0\hat{\sigma}]$ (60 points) to minimize empirical MSE.

8. **NV FP4 (E2M1).** NVIDIA's 4-bit floating-point format [9] with 1 sign bit, 2 exponent bits, and 1 mantissa bit with exponent bias 1. This yields 15 distinct finite values: $\{0,\ \pm 0.5,\ \pm 1.0,\ \pm 1.5,\ \pm 2.0,\ \pm 3.0,\ \pm 4.0,\ \pm 6.0\}$. Nearest-neighbor quantization is performed after scaling the codebook by a per-matrix scalar selected by grid search over $[0.25\hat{\sigma},\ 4.0\hat{\sigma}]$ (60 points) to minimize empirical MSE.

3.1.2 Results

We obtained the following results averaged over 500 generated $A \in \mathbb{R}^{128\times 256}$ and $B \in \mathbb{R}^{256\times 128}$ according to our correlated Gaussian model. From this, we can see that our optimal matrix multiplication quantizer consistently outperforms other commonly used quantization methods for our specified tasks.

We include 95% error bars but used enough synthetic data that they are barely visible.

3.2 Application in Quantized Least Squares

We consider solving the least squares problem

$$\min_W \|XW - Y\|_F^2,$$

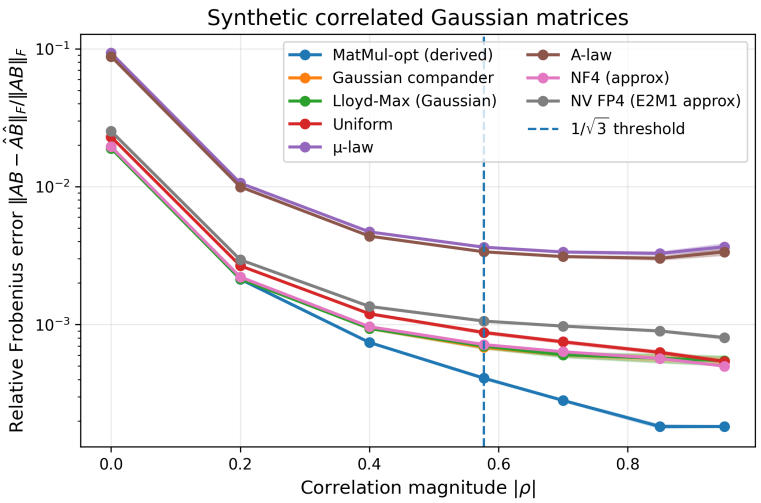

Figure 2: Performance of our optimal quantizer vs. other commonly used quantizers.

where $X \in \mathbb{R}^{n\times d}, W, W^* \in \mathbb{R}^{d\times m}$, $Y = XW^* + \epsilon Z$ and the entries of X, W^* satisfy Assumption 1, and the joint distribution

$$\begin{bmatrix} X \\ W^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right), \qquad \rho = 0.6. \tag{12}$$

Imagine X is so large so that it cannot fit into a single GPU. A natural solution to this would be distributing the matrix over multiple GPUs then solving the problem by appropriately aggregating subproblems, but it may be timely as the loading/deloading time of the data matrix could be a bottleneck in such settings. Hence quantizing X and Y to load them onto a GPU then solving the quantized least squares,

$$\min_W \|Q_X W - Q_Y\|_F^2,$$

may make sense when solving the problem faster could be much more important than solving the problem exactly.

We apply the proposed quantization to the above problem and show that we can have better accuracy in terms of $\|W - W^*\|_F$ under the same bit budget. We compare three different schemes, where

- Scheme 1: quantize X, Y with Gaussian high-rate quantizer, simply considering the marginal distribution.
- Scheme 2: sweep through possible $\rho \in [-1, 1]$ and plot the loss with the best ρ.
- Scheme 3: estimate ρ with

$$\hat{\rho} = \frac{1}{ndm} \sum_{i=1}^{d} \sum_{j=1}^{m} (X\bar{W})_{ij},$$

where $\bar{W}$ is a solution to the subproblem

$$\min_W \|\bar{X}W - \bar{Y}\|_F^2.$$

Here $\bar{X}, \bar{Y}$ are row subsampled matrices of X, Y.

The results show that using nonzero ρ can improve the accuracy of quantized least squares. Sweeping along the possible ρs is better than using a ρ estimate, especially for lower bit budgets. The main reason for this is that for bits=3,4, $\rho = -0.9$ is the optimal ρ for quantized least squares, which is very different from the ground truth $\rho = 0.6$. Such observation implies that there could be a different reason for why the correlation-aware quantization algorithm works better in the least squares setting. For higher rates the optimal ρ becomes similar to ground truth ρ, and is better than naive Gaussian.

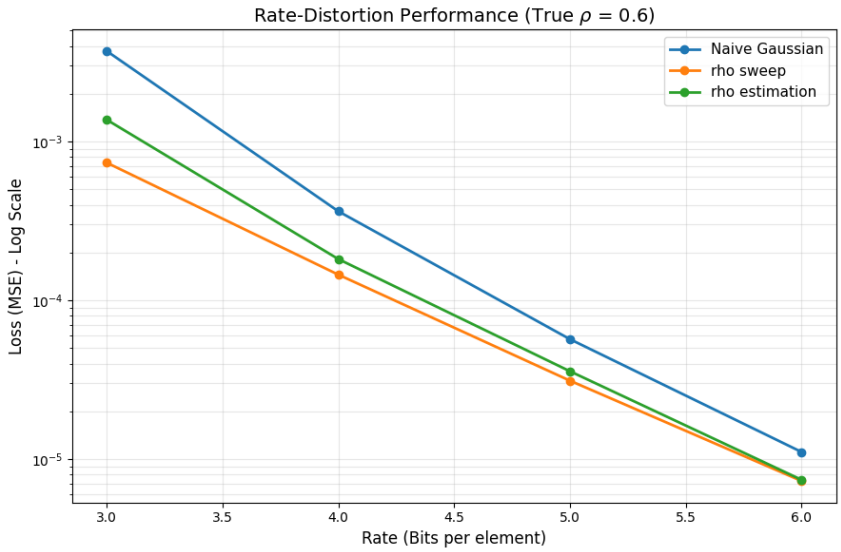

Figure 3: Comparison of three different schemes for solving quantized least squares. The y axis is the difference between ground truth and the quantized solution $\|W - W^*\|_F$.

3.3 Quantization of Transformer-based Models

Activation quantization is critical for efficient inference in Transformer-based models. We compare our method against INT8 quantization [2] and FP8 quantization [8]. In all cases, we apply per-token scaling to Q and K before quantization and use $K = 256$ quantization levels (8 bits).

We empirically observe non-trivial correlation between entries of the query and key activations within attention heads (Figure 4). This motivates applying our correlation-aware matrix multiplication quantizer independently to each attention head. For evaluation, we measure the relative Frobenius error of the pre-softmax attention logits:

$$\frac{\left\|QK^\top - \hat{Q}\hat{K}^\top\right\|_F}{\|QK^\top\|_F},$$

where $\hat{Q}$ and $\hat{K}$ denote quantized activations. This metric directly captures distortion in the attention logits.

3.3.1 Models and Data

We evaluate on the GPT-2 family of models [11] where there key and query activations are directly multiplied in addition so some Qwen3 models [12] which apply rotary embeddings to the key and query activations before multiplying. For the Qwen3 models we tune ρ based on the key and query activations after passing through the rotary embeddings. Sequences are drawn from WikiText-2 [7]. We use 64 non-overlapping sequences of length 128 from the training split for evaluation. We hold out 32 sequences from the validation split for calibration. Per-head statistics are computed by concatenating all evaluation sequences.

3.3.2 Activation Collection

We extract Q and K at each attention layer using PyTorch forward hooks. For GPT-2, we hook the fused `c_attn` module and split the output into Q, K, and V. For Qwen3, we hook the `q_norm`, `k_norm`, and `rotary_emb` submodules per layer; the post-normalization Q and K are buffered, and rotary position embeddings are applied in NumPy to obtain post-RoPE activations.

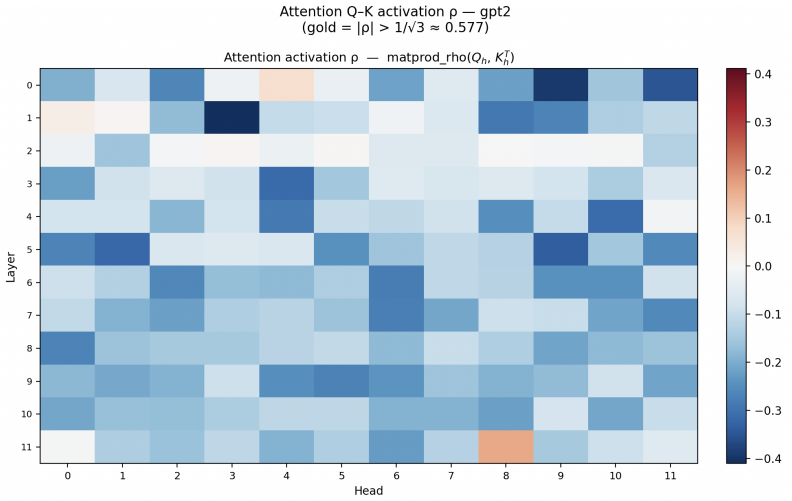

Figure 4: Estimated ρ value for each layer and head in GPT-2 Small

Table 1: Win rate of ρ-tuned vs. INT8 and FP8 across GPT-2 sizes (win = lower relative Frobenius error per head).

Model	ρ-tuned vs FP8 (% wins)	ρ-tuned vs INT8 (% wins)
GPT-2 Small (12 layers, 12 heads)	100%	96.5%
GPT-2 Medium (24 layers, 16 heads)	100%	97.9%
GPT-2 Large (36 layers, 20 heads)	100%	100%
GPT-2 XL (48 layers, 25 heads)	100%	99.7%
Qwen3-0.6B (28 layers, 8 heads)	98.7%	65.6%
Qwen3-1.7B (28 layers, 8 heads)	98.7%	59.4%
Qwen3-8B (36 layers, 8 heads)	86.1%	42.7%

3.3.3 ρ-tuning

For each attention head, we tune $\rho \in [0, 0.95]$ on the calibration set by grid search over 40 evenly spaced values, minimizing the relative Frobenius error above. For all methods, we use the same per-token ℓ_∞ scaling, i.e., we normalize each token vector so all entries lie in $[-1, 1]$. Given a tuned ρ, we construct reproduction points using the closed-form density from Corollary 2.

3.3.4 Results

Table 1 reports the fraction of heads for which ρ-tuned achieves lower relative Frobenius error than the baseline quantizer (a "win").

Across model sizes, ρ-tuned matches or improves upon both INT8 and FP8, with the largest gains observed in larger models for the GPT-2 family of models while the opposite trend was observed on the Qwen3 family. We saw the performance degrade for larger Qwen3 models with the tuned ρ algorithm losing to INT8 for Qwen3-8B. The tuned values of ρ tended to be higher than the estimated values of ρ (see figure 5), which indicates some level of model misspecification in our correlated Gaussian model.

This misspecification seemed to hurt most in the Qwen3 family of models, which we posit is due to the nature of the rotary embeddings and their effect on the activation distributions.

Limitations and Future Work. We plan to build upon our method so that it performs better on key and query quantization in models with rotary embeddings. We will also incorporate outlier-handling techniques commonly used in LLM quantization [2] and report downstream task metrics in addition to logit-level distortion.

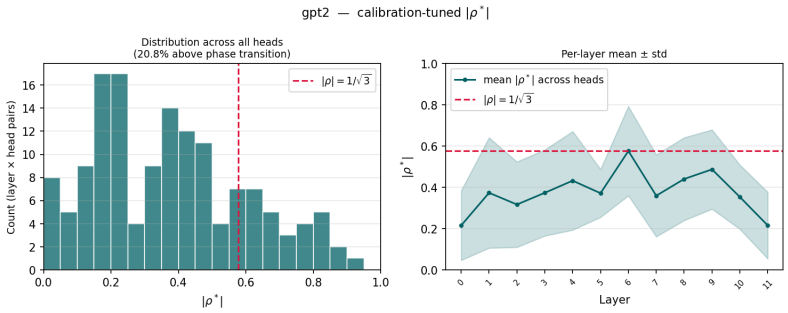

Figure 5: Tuned $|\rho|$ for GPT-small

4 Conclusion

We derived a sharp high-rate characterization of optimal scalar quantization for matrix multiplication MSE under a pair-i.i.d. inner-product model. The leading error constant decomposes into two weighted scalar quantization problems with weights given by conditional second moments, leading to explicit optimal companding densities. For correlated Gaussian multiplicative pairs we obtained a closed-form density and proved a sharp unimodal-to-bimodal phase transition at $|\rho| = 1/\sqrt{3}$. We also benchmarked our derived quantizer on synthetically generated matrices for matrix multiplication and quantized least squares as well as on key and query activation quantization of the GPT-2 family of models.

A Proof of the Decoupling and High-Rate Expansion

This appendix provides a complete proof of the reduction (6) and the negligibility of the bias term in (1) at the K^{-2} scale.

A.1 Exact variance decomposition

Let $D_\ell \triangleq X_\ell Y_\ell - \widehat{X}_\ell \widehat{Y}_\ell$. Then

$$S - \widehat{S} = \sum_{\ell=1}^{k} D_\ell.$$

Because $\{(X_\ell, Y_\ell)\}$ are i.i.d. and quantized entrywise, the D_ℓ are i.i.d. with the same law as D. Hence

$$\mathbb{E}[(S - \widehat{S})^2] = \sum_{\ell=1}^{k} \mathbb{E}[D_\ell^2] + 2 \sum_{\ell<r} \mathbb{E}[D_\ell]\mathbb{E}[D_r] = k\mathbb{E}[D^2] + k(k-1)(\mathbb{E}[D])^2,$$

which is (1).

A.2 Quantizer model and elementary cell identities

Throughout this appendix we analyze companding quantizers $Q_X = Q_{K_X,\lambda_X}$ and $Q_Y = Q_{K_Y,\lambda_Y}$ constructed as in Section 2, with λ_X, λ_Y continuously differentiable and strictly positive.

For $Q_{K,\lambda}$, let boundaries $x_i = G^{-1}(i/K)$ and reproduction points $r_i = G^{-1}((i - \frac{1}{2})/K)$, and denote the ith cell by $I_i = [x_{i-1}, x_i)$ with width $\Delta_i = x_i - x_{i-1}$ and midpoint $m_i \triangleq (x_{i-1} + x_i)/2$.

Lemma 1 (Cell width formula)**.** *For each $i = 1, \ldots, K$, there exists $\xi_i \in I_i$ such that*

$$\Delta_i = \frac{1}{K\,\lambda(\xi_i)}. \tag{13}$$

Proof. By definition $G(x_i) - G(x_{i-1}) = 1/K$. Since G is differentiable with derivative $G'(x) = \lambda(x)$, the mean value theorem gives a point $\xi_i \in (x_{i-1}, x_i)$ such that

$$\frac{1}{K} = G(x_i) - G(x_{i-1}) = G'(\xi_i)(x_i - x_{i-1}) = \lambda(\xi_i)\Delta_i.$$

Rearranging yields (13). □

Lemma 2 (Reproduction points are second-order close to cell midpoints)**.** *Assume λ is continuously differentiable. Then for each fixed compact interval $[-M, M]$ there exists $C_M < \infty$ and K_M such that for all $K \geq K_M$ and all cells I_i intersecting $[-M, M]$,*

$$|r_i - m_i| \leq C_M\,\Delta_i^2. \tag{14}$$

Consequently,

$$\begin{aligned} \int_{x_{i-1}}^{x_i} (x - r_i)\,\mathrm{d}x &= \Delta_i(m_i - r_i) = O(\Delta_i^3), \\ \int_{x_{i-1}}^{x_i} (x - r_i)^2\,\mathrm{d}x &= \frac{\Delta_i^3}{12} + O(\Delta_i^5). \end{aligned} \tag{15}$$

uniformly over such cells.

Proof. Let $g \triangleq G^{-1}$ and $u \triangleq (i - \frac{1}{2})/K$, so that $x_{i-1} = g(u - 1/(2K))$, $x_i = g(u + 1/(2K))$, and $r_i = g(u)$. We know that g is twice continuously differentiable by the inverse function theorem, so taking the second-order Taylor expansion of g around u gives

$$g\left(u \pm \frac{1}{2K}\right) = g(u) \pm \frac{g'(u)}{2K} + \frac{g''(\zeta_\pm)}{8K^2}$$

for some $\zeta_\pm$ between u and $u \pm 1/(2K)$. Averaging the "+" and "−" expansions yields

$$m_i = \frac{x_{i-1} + x_i}{2} = g(u) + \frac{g''(\zeta_+) + g''(\zeta_-)}{16K^2},$$

hence

$$m_i - r_i = \frac{g''(\zeta_+) + g''(\zeta_-)}{16K^2}.$$

We note that by Lemma 1,

$$\Delta_i = \frac{1}{K\lambda(\xi_i)}.$$

Since λ is continuous and strictly positive, it is bounded away from 0 on $[-M, M]$, so $\Delta_i = \Theta(1/K)$ uniformly over cells intersecting $[-M, M]$.

Therefore, on a compact $[-M, M]$, the corresponding u values lie in a compact subset of $(0, 1)$ for all sufficiently large K (since cells intersecting $[-M, M]$ have both endpoints in $g^{-1}([-M - C/K, M + C/K])$ as a consequence of our previous statement), so g'' is bounded there. Thus $|m_i - r_i| = O(1/K^2) = O(\Delta_i^2)$ uniformly over cells intersecting $[-M, M]$, proving (14).

Finally,

$$\begin{aligned}\int_{x_{i-1}}^{x_i} (x - r_i)\,\mathrm{d}x &= \left[\frac{(x - r_i)^2}{2}\right]_{x_{i-1}}^{x_i} = \frac{(x_i - r_i)^2 - (x_{i-1} - r_i)^2}{2} \\ &= \Delta_i(m_i - r_i)\end{aligned}$$

which together with (14) gives the first identity in (15). For the second identity, write $(x - r_i)^2 = (x - m_i)^2 + (m_i - r_i)^2 + 2(x - m_i)(m_i - r_i)$ and integrate over $[x_{i-1}, x_i]$. The cross term integrates to zero by symmetry around m_i, giving

$$\begin{aligned}\int_{x_{i-1}}^{x_i} (x - r_i)^2\,\mathrm{d}x &= \int_{x_{i-1}}^{x_i} (x - m_i)^2\,\mathrm{d}x + \Delta_i(m_i - r_i)^2 \\ &= \frac{\Delta_i^3}{12} + O(\Delta_i^5),\end{aligned}$$

since $\int_{x_{i-1}}^{x_i} (x - m_i)^2\,\mathrm{d}x = \Delta_i^3/12$ exactly and $(m_i - r_i)^2 = O(\Delta_i^4)$. □

A.3 Bias term is negligible at the K^{-2} scale

We prove the key estimate $\mathbb{E}[D] = O(K_X^{-2} + K_Y^{-2})$.

Lemma 3 (A weighted first-moment bound for companders)**.** *Let X have density f and let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable. Let $Q_{K,\lambda}$ be a companding quantizer with continuously differentiable point density $\lambda > 0$. Assume that $q(x) \triangleq g(x)f(x)$ is twice continuously differentiable. Then*

$$\mathbb{E}\big[g(X)\,(X - Q_{K,\lambda}(X))\big] = O\left(\frac{1}{K^2}\right), \tag{16}$$

as $K \to \infty$.

Proof. Let the cells be $I_i = [x_{i-1}, x_i)$ with reproduction r_i, width Δ_i, and midpoint m_i as in Section A.2. Write $e(x) \triangleq x - Q_{K,\lambda}(x)$, so that $e(x) = x - r_i$ for $x \in I_i$. Then

$$\mathbb{E}[g(X)e(X)] = \sum_{i=1}^{K} \int_{I_i} q(x)(x - r_i)\,\mathrm{d}x.$$

Fix $M > 0$ and split the sum into cells intersecting $[-M, M]$ ("interior cells") and the remaining tail cells ("exterior cells").

Step 1: Interior cells. For an interior cell $I_i \subset [-M-\eta, M+\eta]$ (for a fixed small $\eta > 0$ and K large enough), expand q around the midpoint m_i:

$$q(x) = q(m_i) + q'(m_i)(x - m_i) + \frac{1}{2}q''(\zeta_{i,x})(x - m_i)^2,$$

for some $\zeta_{i,x} \in I_i$. Multiply by $(x - r_i) = (x - m_i) + (m_i - r_i)$ and integrate over I_i:

$$\int_{I_i} q(x)(x - r_i)\,\mathrm{d}x = T_{i,0} + T_{i,1} + T_{i,2},$$

where

$$\begin{aligned}
T_{i,0} &\triangleq q(m_i) \int_{I_i} (x - r_i)\,\mathrm{d}x, \\
T_{i,1} &\triangleq q'(m_i) \int_{I_i} (x - m_i)(x - r_i)\,\mathrm{d}x, \\
T_{i,2} &\triangleq \frac{1}{2} \int_{I_i} q''(\zeta_{i,x})(x - m_i)^2(x - r_i)\,\mathrm{d}x.
\end{aligned}$$

We bound each term using Lemma 2.

First, Lemma 2 gives $\int_{I_i}(x - r_i)\,\mathrm{d}x = O(\Delta_i^3)$. Since $q(m_i)$ is bounded on $[-M-\eta, M+\eta]$, we have $T_{i,0} = O(\Delta_i^3)$ uniformly over interior cells.

Second, note that $(x - m_i)(x - r_i) = (x - m_i)^2 + (m_i - r_i)(x - m_i)$ and $\int_{I_i}(x - m_i)\,\mathrm{d}x = 0$ by symmetry. Hence

$$\int_{I_i} (x - m_i)(x - r_i)\,\mathrm{d}x = \int_{I_i} (x - m_i)^2\,\mathrm{d}x = \frac{\Delta_i^3}{12}.$$

Since $q'(m_i)$ is bounded on $[-M-\eta, M+\eta]$, it follows that $T_{i,1} = O(\Delta_i^3)$ uniformly over interior cells.

Third, on I_i we have $|x - m_i| \le \Delta_i/2$ and $|x - r_i| \le |x - m_i| + |m_i - r_i| \le \Delta_i/2 + O(\Delta_i^2) = O(\Delta_i)$. Since q'' is bounded on $[-M-\eta, M+\eta]$, we obtain

$$\begin{aligned}
|T_{i,2}| &\le C \int_{I_i} |x - m_i|^2\ |x - r_i|\ \mathrm{d}x \\
&\le C'\Delta_i \int_{I_i} (x - m_i)^2\,\mathrm{d}x \\
&= C'\Delta_i \frac{\Delta_i^3}{12} \\
&= O(\Delta_i^4).
\end{aligned}$$

Combining $T_{i,0} = O(\Delta_i^3)$, $T_{i,1} = O(\Delta_i^3)$, and $T_{i,2} = O(\Delta_i^4)$ gives

$$\int_{I_i} q(x)(x - r_i)\,\mathrm{d}x = O(\Delta_i^3) \qquad \text{uniformly over interior cells.}$$

Summing over all interior cells and using Lemma 1 (which gives $\Delta_i = O(1/K)$ on compacts), we obtain

$$\begin{aligned}
\sum_{i:\, I_i \cap [-M,M] \neq \emptyset} \int_{I_i} q(x)(x - r_i)\,\mathrm{d}x &= O\left(\sum_{i:\, I_i \cap [-M,M] \neq \emptyset} \Delta_i^3 \right) \\
&= O\left(\frac{1}{K^2} \right),
\end{aligned}$$

because there are $O(K)$ interior cells and each has $\Delta_i^3 = O(1/K^3)$ uniformly.

Step 2: Tail cells. Let $A_M = \{x : |x| > M\}$. Using Cauchy–Schwarz,

$$\left|\int_{A_M} q(x)(x - Q(x))\,\mathrm{d}x\right| \le \left(\int_{A_M} q(x)^2 f(x)^{-1}\,\mathrm{d}x\right)^{1/2} \times \left(\int_{A_M} f(x)(x - Q(x))^2\,\mathrm{d}x\right)^{1/2}.$$

The second factor is the (unweighted) MSE on the tail set and is bounded by $\mathbb{E}[(X - Q(X))^2]^{1/2} = O(1/K)$ for companding quantizers (this is a special case of Appendix B with $w \equiv 1$). The first factor can be made arbitrarily small by choosing M large enough because $q^2/f = g^2 f$ is integrable whenever $\mathbb{E}[g(X)^2] < \infty$, and in our application g is a conditional mean with finite second moment by Assumption 1. Therefore, for any $\varepsilon > 0$ we can pick M so that the tail contribution is at most ε/K in absolute value uniformly in K.

Step 3: Combine. For this fixed M and all sufficiently large K,

$$\mathbb{E}[g(X)e(X)] = O\left(\frac{1}{K^2}\right) + \varepsilon\frac{1}{K}.$$

Letting $\varepsilon \downarrow 0$ and then $K \to \infty$ yields (16). □

Proposition 1 (Bias order for the product error)**.** *Let $D = XY - \widehat{X}\widehat{Y}$ with $\widehat{X} = Q_{K_X,\lambda_X}(X)$ and $\widehat{Y} = Q_{K_Y,\lambda_Y}(Y)$ as above. Then*

$$\mathbb{E}[D] = O(K_X^{-2} + K_Y^{-2}). \tag{17}$$

Consequently,

$$k(k-1)\big(\mathbb{E}[D]\big)^2 = o(K_X^{-2} + K_Y^{-2}) \qquad \text{as} \qquad K_X, K_Y \to \infty.$$

Proof. Write $e_X = X - \widehat{X}$ and $e_Y = Y - \widehat{Y}$. Using (2),

$$\mathbb{E}[D] = \mathbb{E}[Ye_X] + \mathbb{E}[Xe_Y] - \mathbb{E}[e_X e_Y].$$

Since e_X is a function of X only,

$$\mathbb{E}[Ye_X] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\, e_X\big] = \mathbb{E}[\mu_{Y|X}(X)\, e_X].$$

Apply Lemma 3 with $g = \mu_{Y|X}$ and Q_{K_X,λ_X} to obtain $\mathbb{E}[Ye_X] = O(K_X^{-2})$. Similarly, $\mathbb{E}[Xe_Y] = O(K_Y^{-2})$.

For the final term, Cauchy–Schwarz gives

$$|\mathbb{E}[e_X e_Y]| \le \sqrt{\mathbb{E}[e_X^2]\,\mathbb{E}[e_Y^2]}.$$

The unweighted companding MSE satisfies $\mathbb{E}[e_X^2] = O(K_X^{-2})$ and $\mathbb{E}[e_Y^2] = O(K_Y^{-2})$ (a special case of Appendix B), hence $\mathbb{E}[e_X e_Y] = O((K_X K_Y)^{-1})$. Using $2/(K_X K_Y) \le 1/K_X^2 + 1/K_Y^2$ (AM–GM), this is also $O(K_X^{-2} + K_Y^{-2})$. Combining the three bounds yields (17).

Finally, squaring (17) gives $(\mathbb{E}[D])^2 = O((K_X^{-2} + K_Y^{-2})^2) = o(K_X^{-2} + K_Y^{-2})$, completing the proof. □

A.4 Dominant terms in $\mathbb{E}[D^2]$

We now prove that the mixed terms in (3) have expectation $O(K_X^{-2}K_Y^{-2})$.

Lemma 4 (A generic mixed-cell bound)**.** *Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be twice continuously differentiable. Consider companding quantizers $Q_X = Q_{K_X,\lambda_X}$ and $Q_Y = Q_{K_Y,\lambda_Y}$ with continuously differentiable point densities. Let r_i and s_j denote the reproduction points and let Δ_i and δ_j denote the corresponding cell widths for Q_X and Q_Y respectively. Assume the function*

$$M_\phi(x, y) \triangleq \max_{|\alpha|+|\beta|=2} \left|\partial_x^\alpha \partial_y^\beta \phi(x, y)\right|$$

is integrable with respect to $f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$ when weighted by $(\lambda_X(x)\lambda_Y(y))^{-2}$ on compact sets (as made explicit in the proof below). Then, as $K_X, K_Y \to \infty$,

$$\sum_{i=1}^{K_X}\sum_{j=1}^{K_Y}\int_{I_i}\int_{J_j}\phi(x,y)\,(x-r_i)(y-s_j)\,\mathrm{d}y\,\mathrm{d}x = O\Big(\frac{1}{K_X^2K_Y^2}\Big). \tag{18}$$

Proof. Let $I_i = [x_{i-1}, x_i)$ and $J_j = [y_{j-1}, y_j)$ be the cells in x and y. Fix $M > 0$ and split the double sum into (i) rectangles intersecting $[-M, M]^2$ and (ii) rectangles in the complement. We bound both contributions.

Step 1: Rectangles intersecting a compact set. For a rectangle $I_i \times J_j$ intersecting $[-M, M]^2$, pick the center point (r_i, s_j) and perform a second-order Taylor expansion of ϕ around (r_i, s_j):

$$\begin{aligned}\phi(x,y) = \phi(r_i,s_j) &+ \phi_x(r_i,s_j)(x-r_i)\\ &+ \phi_y(r_i,s_j)(y-s_j) + R_{i,j}(x,y),\end{aligned}$$

where the remainder satisfies

$$|R_{i,j}(x,y)| \le \frac{1}{2}M_\phi(\zeta_{i,j}(x,y))\big((x-r_i)^2 + (y-s_j)^2\big)$$

for some $\zeta_{i,j}(x,y) \in I_i \times J_j$.

Multiply the Taylor expansion by $(x-r_i)(y-s_j)$ and integrate over $I_i \times J_j$. The constant term contributes

$$\phi(r_i,s_j)\left(\int_{I_i}(x-r_i)\,\mathrm{d}x\right)\left(\int_{J_j}(y-s_j)\,\mathrm{d}y\right).$$

By Lemma 2, each one-dimensional integral is $O(\Delta_i^3)$ and $O(\delta_j^3)$ on a compact region, hence the constant term contribution is $O(\Delta_i^3\delta_j^3)$.

The ϕ_x term contributes

$$\phi_x(r_i,s_j)\left(\int_{I_i}(x-r_i)^2\,\mathrm{d}x\right)\left(\int_{J_j}(y-s_j)\,\mathrm{d}y\right) = O(\Delta_i^3\delta_j^3),$$

since $\int_{I_i}(x-r_i)^2\,\mathrm{d}x = O(\Delta_i^3)$ by (15). The ϕ_y term is analogous.

For the remainder term, note that on $I_i \times J_j$ we have $|x-r_i| = O(\Delta_i)$ and $|y-s_j| = O(\delta_j)$, so

$$|R_{i,j}(x,y)(x-r_i)(y-s_j)| \le C\,M_\phi(\zeta_{i,j}(x,y))\big(\Delta_i^3\delta_j + \Delta_i\delta_j^3\big).$$

Integrating over $I_i \times J_j$ contributes at most

$$\begin{aligned}C\left(\Delta_i^3\delta_j + \Delta_i\delta_j^3\right)\int_{I_i}\int_{J_j}M_\phi(\zeta_{i,j}(x,y))\,\mathrm{d}y\,\mathrm{d}x \le\\ C'\Delta_i\delta_j\left(\Delta_i^2+\delta_j^2\right)\sup_{I_i\times J_j}M_\phi.\end{aligned}$$

On compact sets, $\sup_{I_i\times J_j}M_\phi$ is bounded and $\Delta_i, \delta_j = O(1/K_X), O(1/K_Y)$, so this remainder contribution is also $O(\Delta_i^3\delta_j^3)$.

Combining all pieces, we have shown that for rectangles intersecting $[-M, M]^2$,

$$\int_{I_i}\int_{J_j}\phi(x,y)(x-r_i)(y-s_j)\,\mathrm{d}y\,\mathrm{d}x = O(\Delta_i^3\delta_j^3),$$

uniformly. Summing over the $O(K_XK_Y)$ rectangles intersecting $[-M, M]^2$ and using $\Delta_i = O(1/K_X)$, $\delta_j = O(1/K_Y)$ on compacts yields a total compact contribution of order $O(K_XK_Y \cdot K_X^{-3}K_Y^{-3}) = O(K_X^{-2}K_Y^{-2})$.

Step 2: Tail rectangles. On the complement of $[-M, M]^2$ we bound the entire integral by absolute values. Since $|x - r_i| \le |x| + |r_i|$ and similarly for $y - s_j$, and since (X, Y) has finite $(4 + \epsilon)$ moments, by choosing M large we can make the tail probability $\Pr((X, Y) \notin [-M, M]^2)$ arbitrarily small. The detailed bound follows the same truncation logic as in Lemma 3: first control the tail by moments and then let $M \to \infty$. This yields a tail contribution that is $o(K_X^{-2} K_Y^{-2})$.

Step 3: Combine and conclude. Let the rectangles which intersect $[-M, M]^2$ be $C := \{(i, j) : I_i \times J_j \cap [-M, M]^2 \neq \emptyset\}$ and $\psi_{i,j}(x, y) := \phi(x, y)\,(x - r_i)(y - s_j)$. Then

$$\sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x = \sum_{(i,j)\in C} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x + \sum_{(i,j)\notin C} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

and from Step 1 we know that $\sum_{(i,j)\in C} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x = O(K_X^{-2} K_Y^{-2})$.

From Step 2 we know that

$$\begin{aligned} &\int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x = o(K_X^{-2} K_Y^{-2}), \ \ \forall (i, j) \notin C \\ \implies &\sum_{(i,j)\notin C} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x = O(K_X K_Y) o(K_X^{-2} K_Y^{-2}) \\ &= O(K_X^{-2} K_Y^{-2}). \end{aligned}$$

Thus we see that

$$\sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} \int_{I_i} \int_{J_j} \psi_{i,j}(x, y)\,\mathrm{d}y\,\mathrm{d}x = O(K_X^{-2} K_Y^{-2})$$

which is exactly (18). □

Proposition 2 (Mixed terms in $\mathbb{E}[D^2]$ are $O(K_X^{-2} K_Y^{-2})$)**.** *Let D be as in Proposition 1. Then the mixed terms in* (3) *satisfy*

$$\mathbb{E}[e_X^2 e_Y^2] = O(K_X^{-2} K_Y^{-2}), \tag{19}$$

$$\mathbb{E}[XY e_X e_Y] = O(K_X^{-2} K_Y^{-2}), \tag{20}$$

$$\mathbb{E}[Y e_X^2 e_Y] = O(K_X^{-2} K_Y^{-2}), \tag{21}$$

$$\mathbb{E}[X e_X e_Y^2] = O(K_X^{-2} K_Y^{-2}). \tag{22}$$

Consequently,

$$\begin{aligned} \mathbb{E}[D^2] &= \mathbb{E}[Y^2 e_X^2] + \mathbb{E}[X^2 e_Y^2] + O(K_X^{-2} K_Y^{-2}) \\ &= \mathbb{E}[Y^2 e_X^2] + \mathbb{E}[X^2 e_Y^2] + o(K_X^{-2} + K_Y^{-2}). \end{aligned}$$

Proof. Each expectation can be written as a double sum over rectangles $I_i \times J_j$.

(i) $\mathbb{E}[XYe_Xe_Y]$. Write

$$\mathbb{E}[XYe_Xe_Y] = \sum_{i=1}^{K_X}\sum_{j=1}^{K_Y}\int_{I_i}\int_{J_j} xy(x-r_i)(y-s_j)f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

Apply Lemma 4 with $\phi(x,y) = xyf_{X,Y}(x,y)$ to obtain (20).

(ii) $\mathbb{E}[Ye_X^2e_Y]$ and $\mathbb{E}[Xe_Xe_Y^2]$. For $\mathbb{E}[Ye_X^2e_Y]$ we write

$$\mathbb{E}[Ye_X^2e_Y] = \sum_{i,j}\int_{I_i}\int_{J_j} y(x-r_i)^2(y-s_j)f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

Fix i and view $(x-r_i)^2$ as a bounded factor of order $O(\Delta_i^2)$ on I_i; then apply the same second-order Taylor argument in y around s_j (as in Lemma 4) to gain an extra factor δ_j^3. Summing over i,j yields $O(K_X^{-2}K_Y^{-2})$. The term $\mathbb{E}[Xe_Xe_Y^2]$ is symmetric.

(iii) $\mathbb{E}[e_X^2e_Y^2]$. Similarly,

$$\mathbb{E}[e_X^2e_Y^2] = \sum_{i,j}\int_{I_i}\int_{J_j} (x-r_i)^2(y-s_j)^2f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

On each rectangle, $(x-r_i)^2 = O(\Delta_i^2)$ and $(y-s_j)^2 = O(\delta_j^2)$, and integrating over the rectangle produces a factor $\Delta_i\delta_j$. Thus each rectangle contributes $O(\Delta_i^3\delta_j^3)$ and summing yields (19).

(iv) Conclusion. Insert (19)–(22) into (3). Since $K_X^{-2}K_Y^{-2} = o(K_X^{-2}+K_Y^{-2})$, the stated expansion follows. □

A.5 Conditional weights and conclusion

Using Proposition 2 in (3) yields

$$\mathbb{E}[D^2] = \mathbb{E}[Y^2e_X^2] + \mathbb{E}[X^2e_Y^2] + o(K_X^{-2}+K_Y^{-2}).$$

Conditioning on X and Y gives

$$\mathbb{E}[Y^2e_X^2] = \mathbb{E}[w_X(X)e_X^2], \qquad \mathbb{E}[X^2e_Y^2] = \mathbb{E}[w_Y(Y)e_Y^2],$$

and combining with Proposition 1 in (1) yields (6).

B Proof of the Weighted Scalar High-Rate Theorem

We prove the weighted high-rate companding theorem used in Theorem 1, including a converse that rules out a better K^{-2} constant for *any* sequence of K-level scalar quantizers.

B.1 Scalar setting

Let X have density f on $\mathbb{R}$ and let $w:\mathbb{R}\to(0,\infty)$ be continuous. For a quantizer Q with at most K reproduction points, define the weighted MSE

$$D(Q) \triangleq \mathbb{E}\big[w(X)(X-Q(X))^2\big] = \int_{\mathbb{R}} f(x)w(x)(x-Q(x))^2\,\mathrm{d}x.$$

Let $D_K^\star \triangleq \inf_{|\mathrm{range}(Q)|\le K} D(Q)$ and define $h(x) \triangleq f(x)w(x)$.

We assume throughout that h is continuous and strictly positive, and that

$$I \triangleq \int_{\mathbb{R}} h(x)^{1/3}\,\mathrm{d}x < \infty. \tag{23}$$

B.2 Nearest-neighbor form

Lemma 5 (Nearest-neighbor regions)**.** *Fix reproduction points $r_1 < \cdots < r_K$. Among all quantizers using this codebook, the minimizer of $D(Q)$ assigns each x to the nearest reproduction point in squared error. Hence optimal quantizers can be taken to have interval cells.*

Proof. We note that by $w > 0$

$$w(x)(x - Q(x))^2 \geq w(x)(x - r_{i_x})^2$$

where $i_x \triangleq \arg\min_i\{(x - r_i)^2\}$ so that we achieve the minimum $D(Q)$ when Q is chosen to assign x to its nearest neighbor among the reproduction points. Furthermore, in one dimension, these intervals $Q^{-1}(r_i) = [x_{i-1}, x_i)$ are characterized by $x_i \triangleq \frac{r_i + r_{i+1}}{2}$.

□

B.3 Achievability: Bennett integral for a fixed point density

Fix a continuous point density $\lambda > 0$ with $\int_{\mathbb{R}} \lambda = 1$ and construct $Q_{K,\lambda}$. Denote its cells by $I_i = [x_{i-1}, x_i)$ with width $\Delta_i = x_i - x_{i-1}$, midpoint m_i, and reproduction point r_i.

Lemma 6 (Bennett integral limit for companders)**.** *Assume λ is continuously differentiable and $\int_{\mathbb{R}} h(x)/\lambda(x)^2 \, \mathrm{d}x < \infty$. Then*

$$\lim_{K\to\infty} K^2 D(Q_{K,\lambda}) = \frac{1}{12} \int_{\mathbb{R}} \frac{h(x)}{\lambda(x)^2} \, \mathrm{d}x. \tag{24}$$

Proof. Write

$$D(Q_{K,\lambda}) = \sum_{i=1}^{K} \int_{I_i} h(x)(x - r_i)^2 \, \mathrm{d}x.$$

Fix $M > 0$ and decompose the sum into cells that intersect $[-M, M]$ and cells that do not. We handle these two contributions separately.

Step 1: Compact contribution. For a cell I_i intersecting $[-M, M]$, Lemma 2 gives

$$\int_{I_i} (x - r_i)^2 \, \mathrm{d}x = \frac{\Delta_i^3}{12} + O(\Delta_i^5),$$

uniformly. Since h is uniformly continuous on a slightly enlarged compact interval and $\Delta_i \to 0$ uniformly on that compact region (by Lemma 1), there exists $\xi_i \in I_i$ such that

$$\int_{I_i} h(x)(x - r_i)^2 \, \mathrm{d}x = h(\xi_i) \left(\frac{\Delta_i^3}{12} + O(\Delta_i^5) \right).$$

Multiply by K^2 and use Lemma 1 to write $\Delta_i = 1/(K\lambda(\eta_i))$ for some $\eta_i \in I_i$:

$$K^2 \int_{I_i} h(x)(x - r_i)^2 \, \mathrm{d}x = \frac{1}{12} \frac{h(\xi_i)}{\lambda(\eta_i)^2} \Delta_i + O\left(\frac{h(\xi_i)}{K^2} \Delta_i \right),$$

because $K^2 \Delta_i^5 = (K^2 \Delta_i^3)\Delta_i^2 = O(\Delta_i^2)\Delta_i = O(K^{-2})\Delta_i$ on compacts where $\Delta_i = O(1/K)$.

Summing over all cells intersecting $[-M, M]$ gives

$$K^2 \sum_{i:\, I_i \cap [-M,M] \neq \emptyset} \int_{I_i} h(x)(x - r_i)^2 \, \mathrm{d}x = \frac{1}{12} \sum_{i:\, I_i \cap [-M,M] \neq \emptyset} \frac{h(\xi_i)}{\lambda(\eta_i)^2} \Delta_i + o_M(1), \tag{25}$$

where $o_M(1) \to 0$ as $K \to \infty$ for each fixed M.

Because $\xi_i, \eta_i \in I_i$ and $\max_{i:\, I_i \cap [-M,M] \neq \emptyset} \Delta_i \to 0$, continuity implies $\frac{h(\xi_i)}{\lambda(\eta_i)^2} = \frac{h(\zeta_i)}{\lambda(\zeta_i)^2} + o(1)$ for some $\zeta_i \in I_i$. Therefore the right-hand side of (25) is a Riemann sum for $\int_{-M}^{M} h(x)/\lambda(x)^2 \, \mathrm{d}x$, and we conclude that

$$\lim_{K\to\infty} K^2 \sum_{i:\, I_i \cap [-M,M] \neq \emptyset} \int_{I_i} h(x)(x - r_i)^2 \, \mathrm{d}x = \frac{1}{12} \int_{-M}^{M} \frac{h(x)}{\lambda(x)^2} \, \mathrm{d}x.$$

Step 2: Tail contribution. Since h/λ^2 is integrable by assumption, we can choose M large enough such that

$$\int_{|x|>M} \frac{h(x)}{\lambda(x)^2}\,\mathrm{d}x \le \varepsilon.$$

A similar bound (using nonnegativity of the integrand and the same cell-width identity as above) shows that for all K,

$$0 \le K^2 \sum_{i:\, I_i \cap [-M,M]=\emptyset} \int_{I_i} h(x)(x-r_i)^2\,\mathrm{d}x \le \frac{1}{12}\int_{|x|>M} \frac{h(x)}{\lambda(x)^2}\,\mathrm{d}x + o(1) \le \frac{\varepsilon}{12} + o(1).$$

Letting $\varepsilon \downarrow 0$ and then $K \to \infty$ gives that the tail contribution vanishes in the K^2-scaled limit.

Step 3: Combine and conclude. Combining the compact and tail contributions and then letting $M \to \infty$ yields (24). □

B.4 Optimize over λ (Hölder inequality)

Given Lemma 6, the leading constant for a fixed point density λ is

$$J(\lambda) \triangleq \int_{\mathbb{R}} \frac{h(x)}{\lambda(x)^2}\,\mathrm{d}x, \qquad \text{subject to} \qquad \lambda > 0,\ \int_{\mathbb{R}} \lambda = 1.$$

Let $a(x) \triangleq h(x)^{1/3}$. Write

$$a(x) = \Big(\frac{a(x)}{\lambda(x)^{2/3}}\Big)\lambda(x)^{2/3}.$$

Hölder with exponents 3 and $3/2$ gives

$$\int_{\mathbb{R}} a \le \left(\int_{\mathbb{R}} \frac{a^3}{\lambda^2}\right)^{1/3} \left(\int_{\mathbb{R}} \lambda\right)^{2/3} = J(\lambda)^{1/3}.$$

Therefore $J(\lambda) \ge (\int a)^3 = I^3$, with equality if and only if $\lambda(x) \propto a(x)$, i.e.,

$$\lambda^\star(x) = \frac{h(x)^{1/3}}{\int_{\mathbb{R}} h(t)^{1/3}\,\mathrm{d}t} = \frac{h(x)^{1/3}}{I}. \tag{26}$$

Combining with Lemma 6 gives the *achievability* statement

$$\limsup_{K\to\infty} K^2 D_K^\star \le \frac{I^3}{12}. \tag{27}$$

B.5 Converse: no smaller K^{-2} constant is possible

We now prove the matching lower bound $\liminf_{K\to\infty} K^2 D_K^\star \ge I^3/12$.

Lemma 7 (A lower bound on the distortion over a compact interval)**.** *Fix $M > 0$ and define the truncated distortion*

$$D_M(Q) \triangleq \int_{-M}^{M} h(x)(x-Q(x))^2\,\mathrm{d}x.$$

Let $h_{\min,M} \triangleq \inf_{x\in[-M,M]} h(x)$, which is strictly positive by continuity. Then for any K-level quantizer Q,

$$D_M(Q) \;\ge\; \frac{1}{12}\sum_{\ell=1}^{L} h_{\ell,\min}\,|J_\ell|^3\,, \tag{28}$$

where $\{J_\ell\}_{\ell=1}^{L}$ is the partition of $[-M,M]$ induced by the quantizer cells (each J_ℓ is an interval on which Q is constant), $L \le K$, and $h_{\ell,\min} \triangleq \inf_{x\in J_\ell} h(x)$.

Proof. On each induced interval J_ℓ the quantizer output is a constant reproduction point, say $Q(x) = r_\ell$. Since $h(x) \geq h_{\ell,\min}$ for $x \in J_\ell$,

$$\int_{J_\ell} h(x)(x - r_\ell)^2 \, \mathrm{d}x \geq h_{\ell,\min} \int_{J_\ell} (x - r_\ell)^2 \, \mathrm{d}x.$$

For an interval of length $|J_\ell|$, the function $r \mapsto \int_{J_\ell} (x - r)^2 \, \mathrm{d}x$ is minimized at the midpoint of J_ℓ, with minimum value $|J_\ell|^3 / 12$. Therefore the integral is at least $|J_\ell|^3 / 12$ for any r_ℓ, giving

$$\int_{J_\ell} h(x)(x - r_\ell)^2 \, \mathrm{d}x \geq \frac{h_{\ell,\min}}{12} |J_\ell|^3 .$$

Summing over $\ell = 1, \ldots, L$ yields (28). □

Lemma 8 (Approximating $\int_{-M}^{M} h^{1/3}$ by a partition of small mesh)**.** *Fix $M > 0$ and let $\omega_M(\delta)$ denote the modulus of continuity of $h^{1/3}$ on $[-M, M]$:*

$$\omega_M(\delta) \triangleq \sup_{\substack{x,y \in [-M,M] \\ |x-y| \leq \delta}} \left| h(x)^{1/3} - h(y)^{1/3} \right| .$$

Then for any partition $\{J_\ell\}_{\ell=1}^{L}$ of $[-M, M]$ into intervals with maximum length at most δ,

$$\sum_{\ell=1}^{L} h_{\ell,\min}^{1/3} \, |J_\ell| \;\geq\; \int_{-M}^{M} h(x)^{1/3} \, \mathrm{d}x \;-\; 2M \, \omega_M(\delta). \tag{29}$$

Proof. Fix an interval J_ℓ. For any $x \in J_\ell$, we have $h_{\ell,\min}^{1/3} \geq h(x)^{1/3} - \omega_M(\delta)$ because $\sup_{t \in J_\ell} \left| h(t)^{1/3} - h(x)^{1/3} \right| \leq \omega_M(\delta)$ and $|J_\ell| \leq \delta$. Integrating over $x \in J_\ell$ yields

$$h_{\ell,\min}^{1/3} \, |J_\ell| \geq \int_{J_\ell} h(x)^{1/3} \, \mathrm{d}x - \omega_M(\delta) \, |J_\ell| .$$

Summing over ℓ and using $\sum_\ell |J_\ell| = 2M$ gives (29). □

Lemma 9 (Compact-set converse constmmant)**.** *Fix $M > 0$ and define*

$$I_M \triangleq \int_{-M}^{M} h(x)^{1/3} \, \mathrm{d}x.$$

Then

$$\liminf_{K \to \infty} K^2 D_K^\star \;\geq\; \frac{I_M^3}{12}. \tag{30}$$

Proof. Since $D(Q) \geq D_M(Q)$ for every quantizer Q, we have $D_K^\star \geq \inf_{|\mathrm{range}(Q)| \leq K} D_M(Q)$, and it is enough to lower bound the latter.

Fix $\delta > 0$. Consider any K-level quantizer Q, and let $\{J_\ell\}_{\ell=1}^{L}$ be the induced partition of $[-M, M]$ (as in Lemma 7), with $L \leq K$.

Case 1: the partition mesh exceeds δ. If $\max_\ell |J_\ell| > \delta$, then Lemma 7 and $h_{\ell,\min} \geq h_{\min,M}$ imply

$$D_M(Q) \geq \frac{h_{\min,M}}{12} \delta^3.$$

Case 2: the partition mesh is at most δ. If $\max_\ell |J_\ell| \leq \delta$, then Lemma 7 gives

$$D_M(Q) \geq \frac{1}{12} \sum_{\ell=1}^{L} \left(h_{\ell,\min}^{1/3} \, |J_\ell| \right)^3 .$$

Apply Hölder to the nonnegative numbers $a_\ell \triangleq h_{\ell,\min}^{1/3}\,|J_\ell|$:

$$\left(\sum_{\ell=1}^{L} a_\ell\right)^3 \le \left(\sum_{\ell=1}^{L} a_\ell^3\right) L^2 \le \left(\sum_{\ell=1}^{L} a_\ell^3\right) K^2,$$

so $\sum_{\ell=1}^{L} a_\ell^3 \ge (\sum_{\ell=1}^{L} a_\ell)^3/K^2$. Therefore

$$D_M(Q) \ge \frac{1}{12K^2}\left(\sum_{\ell=1}^{L} h_{\ell,\min}^{1/3}\,|J_\ell|\right)^3.$$

By Lemma 8, the bracketed term is at least $I_M - 2M\omega_M(\delta)$. Hence

$$D_M(Q) \ge \frac{1}{12K^2}\big(I_M - 2M\omega_M(\delta)\big)^3.$$

Combine the two cases. We have shown that for every K-level quantizer Q,

$$D_M(Q) \ge \min\left\{\frac{h_{\min,M}}{12}\delta^3,\ \frac{1}{12K^2}\big(I_M - 2M\omega_M(\delta)\big)^3\right\}.$$

Taking the infimum over all Q with at most K levels preserves the inequality. Multiply by K^2 and let $K \to \infty$. For fixed δ, the second term in the minimum dominates for large K, yielding

$$\liminf_{K\to\infty} K^2 D_K^\star \ge \frac{1}{12}\big(I_M - 2M\omega_M(\delta)\big)^3.$$

Finally let $\delta \downarrow 0$. Since $h^{1/3}$ is uniformly continuous on $[-M, M]$, $\omega_M(\delta) \to 0$, and we obtain (30). □

Theorem 3 (Weighted scalar high-rate constant and optimal density)**.** *Under* (23)*,*

$$\lim_{K\to\infty} K^2 D_K^\star = \frac{1}{12}\left(\int_{\mathbb{R}} (f(x)w(x))^{1/3}\,\mathrm{d}x\right)^3. \tag{31}$$

Moreover, the unique minimizing point density for the Bennett functional is (26)*, and the sequence of companders $Q_{K,\lambda^\star}$ achieves the limit in* (31)*.*

Proof. The achievability bound (27) follows from Lemma 6 and the Hölder minimization. For the converse, Lemma 9 implies that for every M,

$$\liminf_{K\to\infty} K^2 D_K^\star \ge \frac{I_M^3}{12}, \qquad I_M = \int_{-M}^{M} h(x)^{1/3}\,\mathrm{d}x.$$

Letting $M \to \infty$ and using monotone convergence (since $h^{1/3} \ge 0$ and integrable) gives $\lim_{M\to\infty} I_M = I$, hence $\liminf_{K\to\infty} K^2 D_K^\star \ge I^3/12$. Together with (27) this yields (31). The optimality and uniqueness of $\lambda^\star$ follow from the equality condition in Hölder. □

C Proof of Theorem 2

By Theorem 1, $\lambda_X^\star(x) \propto (f_X(x)w_X(x))^{1/3}$ and $\lambda_Y^\star(y) \propto (f_Y(y)w_Y(y))^{1/3}$.

For joint Gaussian (12), the conditional law satisfies

$$Y \mid X = x \sim \mathcal{N}\left(\rho\frac{\sigma_Y}{\sigma_X}x,\ \sigma_Y^2(1-\rho^2)\right).$$

Hence

$$\begin{aligned} w_X(x) = \mathbb{E}[Y^2 \mid X = x] &= \mathrm{Var}(Y \mid X = x) + (\mathbb{E}[Y \mid X = x])^2 \\ &= \sigma_Y^2(1-\rho^2) + \rho^2\frac{\sigma_Y^2}{\sigma_X^2}x^2. \end{aligned} \tag{32}$$

Similarly,

$$w_Y(y) = \sigma_X^2(1-\rho^2) + \rho^2 \frac{\sigma_X^2}{\sigma_Y^2} y^2. \tag{33}$$

The marginal $X \sim \mathcal{N}(0, \sigma_X^2)$ has density $f_X(x) \propto \exp(-x^2/(2\sigma_X^2))$, so

$$(f_X(x) w_X(x))^{1/3} \;\propto\; \exp\!\left(-\frac{x^2}{6\sigma_X^2}\right) \left((1-\rho^2) + \rho^2 \frac{x^2}{\sigma_X^2}\right)^{1/3},$$

and normalization yields the stated $\lambda_X^\star$. The same reasoning with (33) yields $\lambda_Y^\star$.

D Proof of Theorem 2

Let $u = x/\sigma_X$. Up to an additive constant, the log-density is

$$\ell(u) = -\frac{u^2}{6} + \frac{1}{3} \log\big((1-\rho^2) + \rho^2 u^2\big).$$

Differentiate:

$$\ell'(u) = u \left[-\frac{1}{3} + \frac{2\rho^2}{3((1-\rho^2) + \rho^2 u^2)}\right].$$

Thus $u = 0$ is always stationary. For $u \neq 0$, stationarity requires

$$-\frac{1}{3} + \frac{2\rho^2}{3((1-\rho^2) + \rho^2 u^2)} = 0 \iff (1-\rho^2) + \rho^2 u^2 = 2\rho^2 \iff u^2 = 3 - \frac{1}{\rho^2}.$$

Real nonzero stationary points exist iff $\rho^2 > 1/3$, and then they are $\pm\sqrt{3 - 1/\rho^2}$.

The curvature at the origin is

$$\ell''(0) = \frac{3\rho^2 - 1}{3(1-\rho^2)}.$$

If $\rho^2 < 1/3$, then $\ell''(0) < 0$ and, with no other stationary points, the density is unimodal with maximum at 0. If $\rho^2 > 1/3$, then $\ell''(0) > 0$ so 0 is a strict local minimum and the only other stationary points are the two symmetric maxima computed above, yielding bimodality. At $\rho^2 = 1/3$, $\ell''(0) = 0$ (critical point).

E Closed Form for the Normalizer Integral

Define

$$J(\rho) = \int_{-\infty}^{\infty} e^{-u^2/6} \big((1-\rho^2) + \rho^2 u^2\big)^{1/3} \, \mathrm{d}u.$$

By even symmetry and the substitution $t = u^2$ (so $\mathrm{d}u = \frac{1}{2} t^{-1/2} \, \mathrm{d}t$),

$$J(\rho) = \int_0^{\infty} e^{-t/6} \big((1-\rho^2) + \rho^2 t\big)^{1/3} t^{-1/2} \, \mathrm{d}t. \tag{34}$$

Factoring $(1-\rho^2)$ and rescaling t yields, for $\rho \neq 0$,

$$J(\rho) = \frac{\sqrt{\pi}\,(1-\rho^2)^{5/6}}{|\rho|} \, U\!\left(\frac{1}{2}, \frac{11}{6}, \frac{1-\rho^2}{6\rho^2}\right), \tag{35}$$

where $U(a, b, z)$ is Tricomi's confluent hypergeometric function. Also $J(0) = \sqrt{6\pi}$.

References

[1] W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

[2] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[4] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[5] Nuggehally S Jayant and Peter Noll. *Digital coding of waveforms: principles and applications to speech and video*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1984.

[6] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[7] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

[8] Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. Fp8 formats for deep learning, 2022.

[9] NVIDIA Technical Blog. Introducing nvfp4 for efficient and accurate low-precision inference, June 2025. NVIDIA Technical Blog.

[10] Or Ordentlich and Yury Polyanskiy. Optimal quantization for matrix multiplication. *IEEE Transactions on Information Theory*, 2025.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[12] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.