

High-Rate Quantized Matrix Multiplication: Theory and Practice

Or Ordentlich and Yury Polyanskiy

Abstract—This work investigates the problem of quantized matrix multiplication (MatMul), which has become crucial for the efficient deployment of large language models (LLMs). We consider two settings: 1) Generic MatMul, where both matrices must be quantized (weight+activation quantization); and 2) weight-only quantization, where the second matrix is only known through covariance matrix Σ_X of its columns. For each setting, we first review the fundamental information-theoretic tradeoff between quantization rate and distortion (high-rate theory), and then analyze the performance of several popular quantization schemes, comparing them to these fundamental limits. Specifically, we discuss rate loss (compared to information theoretic optima) of absmax INT and floating-point (FP) quantization, for which we also derive remarkably accurate heuristic approximations. Weight-only quantization is related to the problem of weighted mean squared error (WMSE) source coding, whose classical (reverse) waterfilling solution dictates how one should distribute rate between coordinates of the vector. We show how waterfilling can be used to improve practical LLM quantization algorithms (GPTQ), which at present allocate rate equally. This new scheme (termed “WaterSIC”) only uses scalar INT quantizers, but its high-rate performance is basis free (it depends only on the determinant of Σ_X and, thus, unlike existing schemes, is immune to applying random rotations) and is within a multiplicative factor of $\frac{2\pi e}{12}$ (or 0.25 bit/entry) of the information-theoretic distortion limit (!). GPTQ’s performance is affected by the choice of basis, but for a random rotation and actual Σ_X from Llama-3-8B we find GPTQ to be within 0.1 bit (depending on the layer type) of WaterSIC, suggesting that GPTQ with random rotation is also near optimal (for high-rate quantization).

I. INTRODUCTION

Matrix multiplication (MatMul) is the work-horse of AI. Consequently, great amounts of research efforts are invested in exploring ways in which MatMul can be executed with as few resources as possible, while maintaining satisfactory accuracy for the underlying application.

One of the most prominent techniques for achieving this goal is quantization, where full-precision matrices

are first compressed to R bits per entry, and MatMul is computed based on the compressed descriptions of the matrices. In LLMs, quantization serves two main purposes: 1) Reducing the IO burden, which is often the bottleneck for MatMul. This is achieved since the number of bytes that needs to be moved around the hardware scales linearly with R ; 2) Representing the matrices with small data-types, e.g., INT8, FP8, INT4, FP4, such that on top of IO savings one can also accelerate compute by using faster multipliers.

In light of the above, the last few years have seen an explosion in the number of papers published on quantized matrix multiplication for LLMs, mostly in the AI and machine learning literature. However, the fundamental information theoretic limits on the tradeoff between quantization rate and distortion in quantized MatMul were only recently established [1], and only for the case of generic MatMul, where we have no prior knowledge on the characteristics of the matrices to be multiplied when designing the quantizers.

The purpose of the first part of this paper is to put forth information theoretic benchmarks for generic quantized MatMul, and use them for evaluating the quality of several popular quantization schemes, and their gap to the fundamental limit.

The second part of this paper deals with a quantized MatMul scenario where second-order statistics of one of the matrices participating in the MatMul is available when designing the quantizer. In LLMs this scenario is quite typical since the statistics of activations can be estimated using a calibration set. To keep this paper short and concise, in this setup, we focus on one-sided quantization, where only the matrix on which there is no prior information needs to be quantized, while the second matrix remains in full-resolution. This corresponds to weight-only quantization in LLMs. For this quantization scenario we develop the fundamental limits for high-resolution quantization, and compare the performance of the canonical GPTQ/LDLQ [2], [3] quantization schemes to them. We also leverage insights from the fundamental limits to develop a novel quantization scheme called WaterSIC, which is a variant of GPTQ/LDLQ.

We stress from the outset a critical distinction in the meaning of the word “rate” between information

O. Ordentlich is with the Hebrew University of Jerusalem, Israel (or.ordentlich@mail.huji.ac.il). Y. Polyanskiy is with the MIT, USA (yp@mit.edu). The work of OO was supported by the Israel Science Foundation (ISF), grant No. 2878/25. The work of YP was supported in part by the MIT-IBM Watson AI Lab and by the National Science Foundation under Grant No CCF-2131115.

theorists and practitioners. The former, focusing on fundamental limits, assume that quantization algorithm inputs a vector/matrix of N entries and outputs NR bits, where each bit is dependent on all of the input entries (e.g., as in vector quantization). For the latter, quantization algorithm takes each entry, applies a simple rule to each entry and produces one of 2^R possible discrete outputs (e.g., as in standard INTM or FPM formats). Recently, so called *microscaling* formats started offering a middle ground between the two: a simple scaling operation is applied to a small group of 16 or 32 entries, after which each entry is still processed individually (e.g., as in NVFP4 and MXFP4). The purpose of this survey is to quantify the gap between fundamental limits and the restrictive practical approaches, specifically in the context of LLM quantization.

In order to enable clean and simple analysis, in this paper we restrict attention to high-resolution quantization. This essentially boils down to assuming that the strength of the quantization noise is significantly weaker than that of the signal, which in turn, enables to neglect several terms in the distortion analysis. While the high-resolution assumption may lead to erroneous conclusions for low rate, say $R < 2$ bits per entry, for relatively high-quantization rate it usually provides sufficiently accurate expressions for the purpose of this survey paper, with the advantage of simplified analysis.

A. Organization and summary of results

Below we give a short overview of the content of this survey paper. In Section II we consider the problem of rate- R quantized “generic” MatMul, where no a priori knowledge of the matrices is available when designing the quantizers. We give an overview of known theoretic upper and lower bounds and conclude that in this setup the smallest expected distortion for the ij th entry of the matrix product is $K(i, j) \cdot 2 \cdot 2^{-2R}$, where $K(i, j)$ is the product of squared ℓ_2 norms of the vectors participating in the corresponding inner product, normalized by their dimension.

In Section III we analyze the performance of several popular MatMul quantization schemes including INT multipliers with absmax scaling, floating-point (FP) multipliers with absmax scaling and NVFP4, as well as the recently proposed NestQuant [4] scheme. For INT and FP multipliers we develop approximations on the attained distortion of the form $K(i, j) \cdot 2 \cdot 2^{-2R_{\text{eff}}}$, from which their gap to optimality in bits is immediately seen to be $R - R_{\text{eff}}$, where R is the actual number of bits per entry these schemes use. The FP analysis appears to be new and is empirically shown to be quite accurate. As an example, we show that for FP we get $R_{\text{eff}} \approx \mathcal{M} + 2.23$ bit, where \mathcal{M} is the number of mantissa bits.

In Section IV we discuss the case of *weight-only quantization*: we are interested in the matrix product $X^\top W$ and only W needs to be quantized to \hat{W} , whereas X is given in full resolution. The problem is made interesting through the fact that while quantization of W cannot know X , it has access to covariance matrix Σ_X of (the columns of) X . We show that, in essence, using standard “isotropic” codebook and Σ_X -oblivious quantization, one attains distortion $D \approx (\frac{1}{n} \text{tr} \Sigma_X) 2^{-2R}$. At the same time, the classical waterfilling formula says that optimal codebook (with optimal quantization) can attain $D \approx |\Sigma_X|^{1/n} 2^{-2R}$, with the difference given by the gap in AM-GM inequality for eigenvalues of Σ_X . It turns out that even with isotropic codebook, but Σ_X -aware encoding one can still attain the same performance.

Low-complexity algorithms for weight-only quantization in LLMs have been pioneered by [5] (a Σ_X -oblivious rounding to nearest integer) and GPTQ [2]. The latter was originally derived from [6], but was soon found to be equivalent to classical algorithms known as *successive interference cancellation (SIC)* in communication and Babai’s algorithm in computer science, cf. Section V. We show that GPTQ (that uses constant rate per entry) attains distortion decaying as $D \approx \frac{2\pi e}{12} (\frac{1}{n} \sum_{i=1}^n U_{i,i}^2) 2^{-2R}$, where $U^\top U = \Sigma_X$ is the upper-triangular (Cholesky) decomposition of Σ_X . At the same time, by adjusting rate per-coordinate (“WaterSIC” algorithm) one can attain $D \approx \frac{2\pi e}{12} (\prod_{i=1}^n U_{i,i}^2)^{\frac{1}{n}} 2^{-2R}$. Since $\prod_{i=1}^n U_{i,i} = \sqrt{|\Sigma_X|}$ this performance is only a factor away from information-theoretically optimal, or within a rate gap of at most 0.25 bit.

Figures 5 and 9 exemplify rate gains for all these four (two theoretical and two practical) algorithms on the example of layers of Llama-3-8B and wiki2 calibration data.

B. High-rate assumption and uniform errors.

Some of the results in this survey (including INT and FP analysis) work under the heuristic assumption that quantization can be modeled as additive zero-mean, uniformly distributed error. Here we describe why this is accurate in the high-rate regime.

As an example, consider a high-dimensional vector $x = (x_1, \dots, x_n)$. Its approximation over an ϵ -grid (quantization) is defined as $\tilde{x}_i = \epsilon[x_i/\epsilon]$. The errors $e_i = \tilde{x}_i - x_i$ clearly are deterministic functions of x_i and should not be modeled as random. However, when ϵ is small and entries of x are not “in any special position” with respect to ϵ -grid, we expect the residuals to land fairly uniformly over the interval $[-\epsilon/2, \epsilon/2)$. (This can be easily checked empirically, by sampling x_i i.i.d. P from any unit-variance distribution P with smooth den-

sity and taking $\epsilon \ll 1$; see Fig. 1a.) Furthermore, generally the empirical average $\frac{1}{n} \sum_i x_i e_i \approx \frac{1}{n} \sum_i \tilde{x}_i e_i \approx 0$, cf. [7], [8].¹

Consequently, under the $\epsilon \ll 1$ assumption we can model the effect of quantization as passing data through an additive uniform noise channel:

$$\tilde{x}_i = x_i + e_i, \quad e_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[-\epsilon/2, \epsilon/2]. \quad (1)$$

(This can be made rigorous by replacing uniform quantization with dithered uniform quantization [9].)

Note that generally, our entries will be assumed to be of zero mean and $O(1)$ (empirical) variance and hence \tilde{x}_i 's range over $O(1/\epsilon)$ possible values. Thus, we can think of relationship between ϵ and rate as $\epsilon \asymp 2^{-R}$. Overall, the high-rate assumption allows us to model quantization error as stochastic, independent uniform and additive.

Another way in which high-rate (low ϵ) assumption is helping us is in ignoring so-called ‘‘linear shrinkage’’ factors. Specifically, it turns out that despite \tilde{x} being composed of nearest elements of the ϵ -grid to x , the best estimate of the vector x given \tilde{x} is NOT \tilde{x} itself. I.e. one could benefit from setting

$$\hat{x} = \gamma \tilde{x},$$

for some $\gamma \leq 1$. To justify this, assume that (1) is the correct model. Then

$$\begin{aligned} \frac{1}{n} \|x - \hat{x}\|_2^2 &= \frac{1}{n} \sum_i ((1 - \gamma)x_i + \gamma e_i)^2 \\ &\approx (1 - \gamma)^2 \hat{\mathbb{E}}[X^2] + \gamma^2 \frac{\epsilon^2}{12}. \end{aligned}$$

Optimizing this over γ gives

$$\gamma^* \approx \frac{\hat{\mathbb{E}}[X^2]}{\hat{\mathbb{E}}[X^2] + \frac{\epsilon^2}{12}} = 1 - O(\epsilon^2),$$

where $\hat{\mathbb{E}}[X^2] = \frac{1}{n} \sum_i x_i^2$ is the empirical second moment. This somewhat surprising *shrinkage effect* is empirically demonstrated on Fig. 1b. We do want to emphasize that the above estimate of γ^* is only an approximation of the optimal shrinkage factor. For example, if $x_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[-\frac{1+\epsilon}{2}, \frac{1+\epsilon}{2}]$ then the optimal shrinkage factor is exactly $\gamma^* = 1$ despite the additive-noise estimate giving a value of around $1 - \epsilon^2$.

¹Recall also that our interest is in the quality of approximating the inner product of two vectors $x, y \in \mathbb{R}^n$ from the inner product of their quantized versions \tilde{x}, \tilde{y} . The error in approximating this inner product is a scalar that depends on the $2n$ per-coordinate quantization errors, and therefore it is indeed the ‘‘average’’ behavior of the quantization errors along the n coordinates that will dictate the performance. For the similar reason, inner-product quantization error, being a quadratic function of $2n$ random variables (with linear part dominant in the high-rate regime), will be normally distributed regardless of the precise statistics of each entry’s quantization noise.

Overall, the MSE improvements from applying shrinkage are on the order of $O(\epsilon^4) = O(2^{-4R})$. Thus, the second impact of the high-rate assumption is that we will be ignoring effects of shrinkage, and also dropping all terms of order below 2^{-2R} from discussion of fundamental limits.

II. QUANTIZED MATRIX MULTIPLICATION: THEORY

Let us introduce the problem formally. Consider two matrices $A \in \mathbb{R}^{n \times a}$ and $B \in \mathbb{R}^{n \times b}$. Our goal is to compute their product $A^\top B$, which is simply a matrix of ab inner products of columns of A and B . The problem is that the computation device is only able to receive information about matrices A and B with a constrained rate of R bits per entry. Because of this limitation, computing exact product $A^\top B$ is not possible and hence we need to accept a certain error. The goal of quantized matrix multiplication is to reduce this error.

More formally, a rate R quantization scheme \mathcal{S} for matrix multiplication $A^\top B$ where $A \in \mathbb{R}^{n \times a}$ and $B \in \mathbb{R}^{n \times b}$ consists of:

- 1) A random variable $\omega \in \Omega$ (shared common randomness, or ‘‘public coin’’)
- 2) A pair of encoders $f_1 : \mathbb{R}^{n \times a} \times \Omega \rightarrow [2^{naR}]$, $f_2 : \mathbb{R}^{n \times b} \times \Omega \rightarrow [2^{nbR}]$. These encoders produce rate R descriptions of two matrices.
- 3) A decoder $g : [2^{naR}] \times [2^{nbR}] \times \Omega \rightarrow \mathbb{R}^{a \times b}$, which produces an estimate of the matrix product. We will simply write $\widehat{A^\top B} = g(f_1(A, \omega), f_2(B, \omega), \omega)$ to denote this approximation.

The distortion of the scheme on the instance A, B is defined as

$$D(A, B; \mathcal{S}) = \frac{1}{nab} \mathbb{E}_{\omega \sim P_\omega} [\|A^\top B - \widehat{A^\top B}\|_F^2].$$

To complete the problem setting we need to agree on what is known a priori about matrices A and B . We will consider two cases: worst-case (where nothing, except norm constraints, are assumed about A and B) and average-Gaussian.

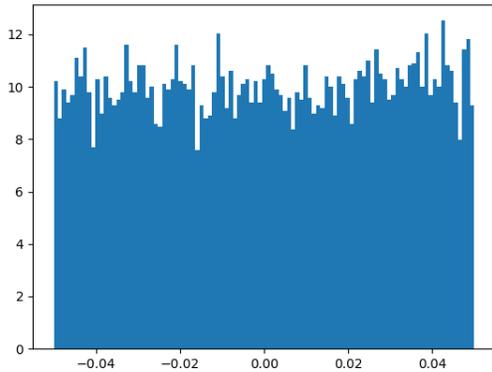
A. Quantization: worst-case setting

Let us consider the worst-case performance of a scheme, defined as

$$D(\mathcal{S}) = \sup_{A, B} D(A, B; \mathcal{S}),$$

with supremum over matrices A, B with all columns of ℓ_2 norm \sqrt{n} . Throughout this subsection we will assume this restriction on A, B unless noted otherwise.

The first and most natural scheme is to quantize each column of A and B using a rate- R vector quantizer over the sphere $\sqrt{n}\mathcal{S}^{n-1}$ and then set $\widehat{A^\top B} = \hat{A}^\top \hat{B}$. It is well



(a) Histogram of rounding errors

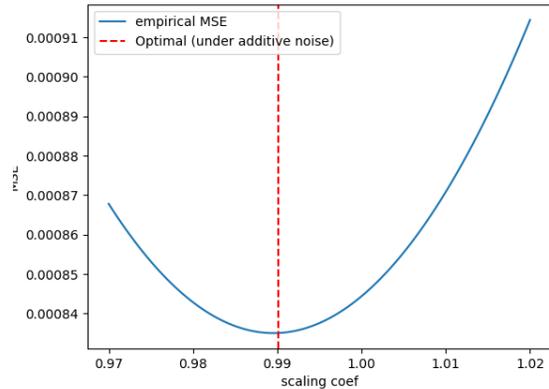
(b) MSE as a function of scaling coefficient γ

Fig. 1: Distribution of ϵ -quantization errors of a vector $\mathcal{N}(0, \sigma^2 I_{10000})$ with $\sigma^2 = 1/12$ and $\epsilon = 0.1$ and effects of shrinkage on MSE.

known that the best spherical quantizer is able to achieve approximation guarantee (uniformly over all A) s.t.

$$\|A - \hat{A}\|_F^2 \leq \|A\|_F^2 2^{-2R+o(1)} = na2^{-2R+o(1)}.$$

How do we bound worst-case distortion? Suppose that we were lucky and $\hat{B} = B$, then the best bound we can have is from Cauchy-Schwarz

$$\|\hat{A}^\top \hat{B} - A^\top B\|_F^2 = \|B^\top (\hat{A} - A)\|_F^2 \leq \|B\|_F^2 \|\hat{A} - A\|_F^2,$$

which is tight for the case when B and the quantization error $\hat{A} - A$ are co-aligned. Dropping the assumption of $\hat{B} = B$ and applying triangle inequality one can show using the same method the bound

$$D(\mathcal{S}) \leq n2 \cdot 2^{-2R+o(1)}.$$

This is not good since we were expecting to get bounded distortion (independent of n).

Can this adversarial alignment between B and quantization error of A be prevented by a better scheme? The answer is negative, at least for *any scheme that does not use common randomness* ω . Specifically, [1, Proposition 1] shows that any such non-randomized quantizer scheme \mathcal{S} for any finite rate $R > 0$ must satisfy

$$D(\mathcal{S}) \geq \delta(R)n,$$

where $\delta(R) > 0$ and is independent of \mathcal{S} .

The next idea is that of sketching [10], which is often used in the literature on approximate matrix multiplication. For simplicity, let us further assume $a = b = 1$. Then define $U \sim \mathcal{N}(0, I_n)$ and let it be shared between the encoders f_1 and f_2 . Let us then compute two scalar quantities $U_A \triangleq U^\top A$ and $U_B \triangleq U^\top B$. Notice that

$$\mathbb{E}[U_A U_B] = A^\top \mathbb{E}[U U^\top] B = A^\top B,$$

and hence the product $U_A U_B$ yields an unbiased estimate of the inner product. Now, since these two scalar quantities are with high probability bounded by $O(\sqrt{n})$ they can be easily quantized with exponentially small mean squared error given the budget of nR bits. Unfortunately, this idea also does not work, as a simple computation shows

$$\begin{aligned} \text{Var}[U_A U_B] &= \mathbb{E}[U_A^2 U_B^2] - (A^\top B)^2 \\ &= n^2 \mathbb{E}[U_1^2 (\rho U_1 + \sqrt{1 - \rho^2} U_2)^2] - n^2 \rho^2 \\ &= n^2 (1 + \rho^2), \end{aligned}$$

where $\rho = \frac{1}{n} A^\top B$ is the cosine of the angle between A and B . We see that even if U_A and U_B were provided to the decoder exactly, the error would still be $\Omega(n^2)$ due to variance of the estimate $U_A U_B$. Thus this kind of scheme again only attains guarantee

$$D(\mathcal{S}) \gtrsim n.$$

Of course, one could reduce variance by working with multiple sketches, however, to drop it all the way to $O(n)$ one would need order n sketches and then nR bits would need to be spent on quantizing an n dimensional vector, bringing us back to the original problem.

So far, we have seen that all non-randomized constructions as well as those based on sketching are only yielding $\Omega(n)$ distortion. Nevertheless, by combining randomization (in the form of random rotation and dither) with good lattice quantizers [1, Theorem 3] constructs a scheme attaining guarantee

$$D(\mathcal{S}) \leq \frac{2 \cdot 2^{2R} - 1}{(2^{2R} - 1)^2} + o(1). \quad (2)$$

In fact, the scheme achieves stronger guarantee even after dropping the assumption of normalized columns.

Specifically, for all matrices A and B with non-zero columns of norms upper- and lower-bounded by polynomials in n , we have

$$\begin{aligned} \forall i \in [a], j \in [b] : \\ \mathbb{E}_\omega \left((A^\top B)_{ij} - (\widehat{A^\top B})_{ij} \right)^2 \\ \leq \frac{\|a_i\|^2 \cdot \|b_j\|^2}{n} \left(\frac{2 \cdot 2^{2R} - 1}{(2^{2R} - 1)^2} + o(1) \right) \\ \approx \frac{\|a_i\|^2 \cdot \|b_j\|^2}{n} 2 \cdot (2^{-2R} + o(1)). \end{aligned} \quad (3)$$

In the last step we applied our high-rate assumption and thus dropping terms smaller than 2^{-2R} from consideration.

As will be seen from the next section, for Gaussian matrices A and B this upper bound is essentially tight. Thus, this performance guarantee cannot be significantly improved (except possibly by terms which are $o(2^{-2R})$ as $R \rightarrow \infty$).

Consequently, when designing rate R quantization schemes for generic matrices, with $R \gg 1$, the smallest distortion we can attain simultaneously for all $A \in \mathbb{R}^{n \times a}$, $B \in \mathbb{R}^{n \times b}$ is characterized by (3), and we therefore refer to

$$D_{ij}^* = \frac{\|a_i\|^2 \cdot \|b_j\|^2}{n} 2 \cdot 2^{-2R}, \quad i \in [a], j \in [b] \quad (4)$$

as the fundamental limit for high-resolution quantized generic matrix multiplication.

B. Quantization: the iid Gaussian case

So far we have only discussed the upper bounds. In order to prove lower bounds, we switch to studying the random-input case. Specifically, we define for any quantization scheme \mathcal{S}

$$\begin{aligned} D_{\text{Gaussian}}(\mathcal{S}) &= \mathbb{E}_{A,B} D(A, B; \mathcal{S}) \\ &= \frac{1}{nab} \mathbb{E}_{A,B,\omega} [\|A^\top B - \widehat{A^\top B}\|_F^2], \end{aligned}$$

where $A_{i,j} \perp\!\!\!\perp B_{i,k} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

[1, Theorem 2] shows that for any scheme \mathcal{S} of rate R we have (non-asymptotically, for all a, b, n)

$$D_{\text{Gaussian}}(\mathcal{S}) \geq \sigma^4 \Gamma(R),$$

where function $\Gamma(R)$ equals

$$\begin{cases} 1 - \left(1 - (2 \cdot 2^{-2R^*} - 2^{-4R^*})\right) \frac{R}{R^*} & R \leq R^* \\ 2 \cdot 2^{-2R^*} - 2^{-4R^*} & R > R^* \end{cases}, \quad (5)$$

where the critical rate is $R^* \approx 0.906$.

Furthermore, it was shown in [1, Theorem 1] that there exists a quantization scheme \mathcal{S} (which is a variation of the lattice-based construction behind (3)) attaining

$$D_{\text{Gaussian}}(\mathcal{S}) \leq \sigma^4 \Gamma(R) + o(1)$$

for all a, b and sufficiently large n . Therefore, the true fundamental limit was determined

$$\inf_{\mathcal{S}} D_{\text{Gaussian}}(\mathcal{S}) = \sigma^4 \Gamma(R) + o(1).$$

One interesting consequence of this result is demonstrating that (at least for Gaussian data and $R < R^*$) optimal strategy *must* combine sketching, also known as Johnson-Lindenstrauss dimensionality reduction, with traditional vector quantization. Empirically, this combination algorithm was investigated in [11], [12].

In this paper, as we agreed, we focus on the high-rate regime. In this regime, $R > R^*$ and it is further assumed large enough such that $2^{-4R} \ll 2 \cdot 2^{-2R}$ so that we can approximate $\Gamma(R) \approx 2 \cdot 2^{-2R}$. Overall, we see that on one hand, there exists a scheme (3) that simultaneously attains

$$\mathbb{E}[\|\widehat{A^\top B} - A^\top B\|_F^2] \lesssim \frac{\|A\|_F^2 \|B\|_F^2}{n} 2 \cdot 2^{-2R}$$

and that this bound is not improvable generally, since for iid Gaussian A, B we have left-hand side lower bounded by the right-hand side (upto $o(2^{-2R})$ terms in the high-rate regime).

Thus, overall, we have theoretically backed reasons to think of $2 \cdot 2^{-2R}$ as fundamental limit of quantized matrix product.

III. QUANTIZED GENERIC MATRIX MULTIPLICATION: PRACTICE

In this section we restrict attention to inner product $x^\top y$ for $x, y \in \mathbb{R}^n$. The matrix product $A^\top B$ of $A = [a_1 | \dots | a_a] \in \mathbb{R}^{a \times n}$ and $B = [b_1 | \dots | b_b] \in \mathbb{R}^{b \times n}$ consists of $a \times b$ inner products. In all schemes we discuss below, the approximate MatMul is computed by separately quantizing each column of each matrix, and then computing an approximation for each entry $(A^\top B)_{ij} \approx \hat{a}_i^\top \hat{b}_j$ by computing the inner product of quantized representations of $a_i \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^n$. Thus, analyzing the distortion of these schemes for quantized inner product immediately lends itself to analysis on the distortion of each entry in the matrix product.

We will assume throughout this section that per column scaling at perfect resolution is possible. Mathematically, this means that \hat{x} and \hat{y} are represented in the form of $\hat{x} = \gamma_x \tilde{x}$, $\hat{y} = \gamma_y \tilde{y}$, where γ_x and γ_y are infinite-resolution scalars, and \tilde{x}, \tilde{y} are high dimensional low precision vectors. In the AI literature this idea is sometimes called *per-channel* and *per-token* scaling [13]. In practice, these scalars are usually given in FP16/FP32 resolution, which contributes a miniscule rate overhead due to high dimensionality of the main vectors \tilde{x}, \tilde{y} .

A. INTM Multipliers

The constellation represented by an INTM data-type is usually defined as $\mathcal{F}_{\text{INTM}} = \mathbb{Z} \cap [-2^{M-1}, 2^{M-1}]$. Here, to simplify expressions, we will extend the constellation by one point and assume it consists of $\mathbb{Z} \cap [-2^{M-1}, 2^{M-1}]$. The most straightforward way to approximate $x^\top y$ using INTM multipliers is *absmax quantization* defined as follows:²

- Set $\gamma_x = 2^{-M-1}\|x\|_\infty$, $\gamma_y = 2^{-M-1}\|y\|_\infty$.
- Set $\hat{x} = \text{round}(x/\gamma_x)$, $\hat{y} = \text{round}(y/\gamma_y)$
- Set $\widehat{x^\top y} = \gamma_x \cdot \gamma_y \cdot \hat{x}^\top \hat{y}$.

Note that the definition of γ_x, γ_y ensures that all entries of the integer vectors \hat{x}, \hat{y} are in $[-2^{M-1}, 2^{M-1}]$ and can therefore indeed be represented by the INTM format. Let

$$e_x = \hat{x} - x/\gamma_x, \quad e_y = \hat{y} - y/\gamma_y. \quad (6)$$

For the analysis, we will make the simplifying assumption that $e_x, e_y \stackrel{iid}{\sim} [-1/2, 1/2]^n$ and are statistically independent of (x, y) . This assumption is certainly incorrect, but nevertheless, it results in a simple and quite accurate analysis for most pairs x, y encountered in real-worlds applications. Furthermore, the assumption can be made exact if one uses subtractive dithering [9]. We have

$$\begin{aligned} D_{\text{INTM}} &= \mathbb{E}(\widehat{x^\top y} - x^\top y)^2 \\ &= \mathbb{E}(\gamma_x e_x^\top y + \gamma_y e_y^\top x + \gamma_x \gamma_y e_x^\top e_y)^2 \\ &= \frac{\gamma_x^2}{12} \|y\|^2 + \frac{\gamma_y^2}{12} \|x\|^2 + \frac{n\gamma_x^2 \gamma_y^2}{12^2} \\ &= \frac{\|x\|^2 \cdot \|y\|^2}{n} \cdot \frac{2^{-2M}}{3} \\ &\quad \left[\frac{\|x\|_\infty^2}{\|x\|^2/n} + \frac{\|y\|_\infty^2}{\|y\|^2/n} + \frac{2^{-2M}}{3} \cdot \frac{\|x\|_\infty^2}{\|x\|^2/n} \cdot \frac{\|y\|_\infty^2}{\|y\|^2/n} \right]. \end{aligned}$$

In what follows we will assume M is large enough such that the last term above can be neglected (this is manifestation of our focus on a high-resolution analysis). With this approximation, we obtain

$$D_{\text{INTM}} \lesssim \frac{\|x\|^2 \cdot \|y\|^2}{n} \cdot 2^{-2M} \cdot \frac{\Delta_{\text{INT}}(x, y)}{3}, \quad (7)$$

where

$$\Delta_{\text{INT}}(x, y) = \frac{1}{2} \left(\frac{\|x\|_\infty^2}{\|x\|^2/n} + \frac{\|y\|_\infty^2}{\|y\|^2/n} \right). \quad (8)$$

We have that $\Delta_{\text{INT}}(x, y) \leq n$ Since $\frac{\|x\|_\infty^2}{\|x\|^2/n} \leq n$, and this is tight for the the natural basis vectors of \mathbb{R}^n . Thus, for some matrices, D_{INTM} exceeds the fundamental limit by a multiplicative factor of $\Omega(n)$. As we have

²We note that many openly available implementations of absmax, e.g. derivatives of GPTQ [14], implement absmax suboptimally, so that the smallest integer value of INTM is either unused (50% of time), or only used for one entry of the block (the other 50%).

seen before, see (2), deterministic quantizers generally must suffer from this kind of degradation.³

A popular way of preventing the situation of large $\Delta_{\text{INT}}(x, y)$ is applying a random orthogonal matrix $S \in O(n) \subset \mathbb{R}^{n \times n}$ to both vectors prior to quantization, such that $x \leftarrow Sx$ and $y \leftarrow Sy$. This has no effect on the inner product as $(Sx)^\top (Sy) = x^\top S^\top S y = x^\top y$, since $S^\top S = I_n$. On the other hand, since Sx as well as Sy become uniformly distributed on the n -dimensional sphere, one can show that⁴ that as long as $n \geq 27$

$$\mathbb{E}[\Delta_{\text{INT}}(x, y)] = \mathbb{E} \left[\frac{\|Sx\|_\infty^2}{\|Sx\|^2/n} \right] \leq 2 \ln n \quad \forall x, y \in \mathbb{R}^n.$$

We remark that applying random rotation (usually, as a random Hadamard transform) to remove alignment between data and quantization error, has been a standard technique in digital communication [15] and federated learning [16]. It has been recently applied to LLM quantization [3], [17], [18], and FP4-training [19]. Under random rotation, thus, we obtain

$$D_{\text{INTM,rotated}} \lesssim \frac{\|x\|^2 \cdot \|y\|^2}{n} \cdot 2^{-2R_{\text{eff}}(\text{INTM})}, \quad (9)$$

where $D_{\text{INTM,rotated}}$ is the *expected* distortion and⁵

$$R_{\text{eff}}(\text{INTM}) = M - \frac{1}{2} \log \left(\frac{2 \ln n}{3} \right). \quad (10)$$

For example, for $n = 4096$ we have that $R_{\text{eff}}(\text{INTM}) \approx M - 1.235$. In Section III-E, Table I it is numerically verified that the approximation (7) for the general case, as well as the approximation (9) for the rotated case, are remarkably accurate. Note that the $\frac{1}{2} \log \left(\frac{2 \ln n}{3} \right)$ rate penalty stems from using a single scale γ_x, γ_y for each vector. If we use group-scaling, that is, allocate a different scale to each sub-block of size $m < n$, the rate penalty will be reduced to $\frac{1}{2} \log \left(\frac{2 \ln m}{3} \right)$. However, encoding the group-scales incurs a rate penalty of $c \cdot \log m/m$ bits, where c is the number of bits used for describing the scale. Thus, one needs to optimize m in order to strike the right balance between the two effects.

For the Gaussian case where $x, y \sim \mathcal{N}(0, \sigma^2 I_n)$ and are statistically independent, we have that

³Though, we note, here the situation is more subtle: the quantization noise is assumed additive and independent of x , thus making quantizer non-deterministic. But the noise is nevertheless correlated with x through scaling factor.

⁴To see this, first observe $\mathbb{E} \left[\frac{\|Sx\|_\infty^2}{\|Sx\|^2/n} \right] = \mathbb{E} \|Z\|_\infty^2$ for $Z \sim \mathcal{N}(0, I_n)$. Next, we use $\mathbb{E} \|Z\|_\infty^2 = \int_0^\infty \Pr(\|Z\|_\infty^2 > t) dt$ along with the upper bound $\Pr(\|Z\|_\infty^2 > t) = 1 - (1 - 2Q(\sqrt{t}))^n \leq \min\{1, n\sqrt{\frac{2}{\pi t}} e^{-t/2}\}$. For any $t_0 > 0$ we can therefore upper bound the integral by $t_0 + n \int_{t_0}^\infty \sqrt{\frac{2}{\pi t}} e^{-t/2} dt \leq t_0 + n \sqrt{\frac{8}{\pi t_0}} e^{-t_0/2}$. For all $n \geq 3$ we may take $t_0 = 2 \ln n - \ln \ln n - \ln(\pi/2) > 0$, and the resulting bound is smaller than $2 \ln n$ for all $n \geq 27$.

⁵In this paper rate is measured in bits, and therefore log is always taken to base 2. We denote logarithm with the natural basis by \ln .

$\mathbb{E} \left[\frac{\|x\|^2 \cdot \|y\|^2}{n} \right] = \sigma^4 n$ and that $\mathbb{E} \left[\frac{\Delta_{\text{INT}}(x,y)}{3} \right] \leq \frac{2}{3} \ln n$. While these quantities are not uncorrelated, for large n their correlation vanishes, and we therefore treat them as uncorrelated.⁶ We consequently obtain that for Gaussian iid x, y and large n

$$\frac{1}{2n\sigma^4} \mathbb{E}(\widehat{x^\top y} - x^\top y)^2 \lesssim D_{\text{INTM, Gaussian}} \triangleq 2^{-2R_{\text{eff}}(\text{INTM})}. \quad (11)$$

B. Floating Point Multipliers

A signed floating point (FP) data-type consists of 1 sign bit s , \mathcal{M} mantissa bits $m_1, \dots, m_{\mathcal{M}}$ and \mathcal{E} exponent bits $e_1, \dots, e_{\mathcal{E}}$. It is further characterized by a fixed *exponent-bias* term $\mu = 2^{\mathcal{E}-1} - 1$. For example, the most popular FP8 format is E4M3, referring to $\mathcal{E} = 4$, $\mathcal{M} = 3$ and one sign bit. The constellation of points that can be represented by FP format contains the set

$$\mathcal{F}_{\text{FP}} = \left\{ (-1)^s \cdot 2^{E-\mu} \cdot (1 + 2^{-\mathcal{M}} \cdot M) : s \in \{0, 1\}, \right. \\ \left. E \in \{1, \dots, 2^{\mathcal{E}} - 2\}, M \in \{0, 1, \dots, 2^{\mathcal{M}} - 1\} \right\} \cup \{0\}.$$

In fact, one can represent more numbers called subnormal numbers when $E = 0$ and NaN and Inf when $E = 2^{\mathcal{E}} - 1$, but for simplicity we omitted those options from the definition of \mathcal{F}_{FP} above.

The FP quantization $Q_{\text{FP}}(z)$ of a number $z \in \mathbb{R}$ is as follows:

- Set $s_z \leftarrow \mathbb{1}\{z < 0\}$
- Set $E_z \leftarrow \mu + \lceil \log |z| \rceil$, and $\bar{z} = 2^{-(E_z - \mu)} \cdot |z|$. Note that $\bar{z} = 2^{-\lceil \log |z| \rceil} \cdot |z| \in [1, 2)$.
- Set $M_z \leftarrow \text{round}(2^{\mathcal{M}} \cdot (\bar{z} - 1))$
- If $M_z = 2^{\mathcal{M}}$, set $M_z \leftarrow 0$ and $E_z \leftarrow E_z + 1$

We then set $Q_{\text{FP}}(z) = (-1)^{s_z} \cdot 2^{E_z - \mu} \cdot (1 + 2^{-\mathcal{M}} \cdot M_z)$. This is possible if $E_z \in \{1, \dots, 2^{\mathcal{E}} - 2\}$. Otherwise overload occurs, and we output 0 if $E_z < 1$ or the signed largest value in the constellation (if $E_z > 2^{\mathcal{E}} - 1$).

We also define

$$e_{\bar{z}} \triangleq \bar{z} - 2^{-\mathcal{M}} \cdot \text{round}(2^{\mathcal{M}} \cdot \bar{z}), \quad \rho_z \triangleq \frac{2^{\lceil \log |z| \rceil}}{|z|} \quad (12)$$

and note that $\rho_z \in (\frac{1}{2}, 1]$, $e_{\bar{z}} \in 2^{-\mathcal{M}} [-\frac{1}{2}, \frac{1}{2})$, and that if overload did not occur we have that

$$Q_{\text{FP}}(z) = z(1 + \rho_z \cdot e_{\bar{z}}). \quad (13)$$

For the analysis that follows, we will adopt an approximate model which we call the *infinite-exponent independent noise* (IEIN) model. Under this model $Q_{\text{FP}}(z)$

⁶A more careful inspection of D_{INTM} for the Gaussian case shows that the expression in (11) is valid without any assumptions on the correlation between $\frac{\|x\|^2 \cdot \|y\|^2}{n}$ and $\Delta_{\text{INT}}(x, y)$.

is defined by equation (13), and $(e_{\bar{z}}, \rho_z, z)$ are mutually independent with $e_{\bar{z}} \sim \text{Uniform}(2^{-\mathcal{M}} [-\frac{1}{2}, \frac{1}{2}))$ and $\rho_z = 2^{-U}$ where $U \sim \text{Uniform}([0, 1])$. Both approximations are justified by the observation that while quantizing high-dimensional vectors, residuals of the rounding $[\cdot]$ will generally be equi-distributed in the $[-1/2, 1/2)$ interval, as we discussed in Section I-B.

The most straightforward way to approximate $x^\top y$ using FP multipliers is *absmax quantization*. Below we describe *dithered absmax quantization*, which is a simple variant of absmax quantization that is as easy to implement and will simplify the analysis:

- Let $U_x, U_y \sim \text{Uniform}([0, 1])$ be statistically independent. Let $\mathcal{E}_{\text{max-}} = 2^{\mathcal{E}} - 2 - (\mu - 1)$. Set $\gamma_x = 2^{U_x} 2^{-\mathcal{E}_{\text{max-}}} \|x\|_{\infty}$, $\gamma_y = 2^{U_y} 2^{-\mathcal{E}_{\text{max-}}} \|y\|_{\infty}$.
- Set $\tilde{x} = x/\gamma_x$ and $\tilde{y} = y/\gamma_y$
- Set $\hat{\tilde{x}} = Q_{\text{FP}}(\tilde{x})$, $\hat{\tilde{y}} = Q_{\text{FP}}(\tilde{y})$
- Set $\widehat{x^\top y} = \gamma_x \cdot \gamma_y \cdot \hat{\tilde{x}}^\top \hat{\tilde{y}}$.

When $U_x = U_y = 1$ with probability 1, this essentially reduces to standard absmax quantization.

Note that under dithered absmax quantization the IEIN model becomes quite realistic. In particular,

$$\rho_{\tilde{x}_i} = 2^{\lceil \log(\tilde{x}_i) \rceil - \log(\tilde{x}_i)} = 2^{-(\log(\tilde{x}_i 2^{-U_x}) - \lceil \log(\tilde{x}_i 2^{-U_x}) \rceil)},$$

where $\tilde{x}_i = 2^{\mathcal{E}_{\text{max-}}} x_i / \|x\|_{\infty}$. Since $U_x \sim \text{Uniform}([0, 1])$ is statistically independent of everything, we have that the IEIN assumption on $\rho_{\tilde{x}_i}$ holds exactly for all $i \in [n]$. Similarly it holds for $\rho_{\tilde{y}_i}$ for all $i \in [n]$, and $\rho_{\tilde{x}_i} \perp \rho_{\tilde{y}_i}$ and are also independent of x, y . Furthermore, the scaling by γ_x, γ_y prior to applying $Q_{\text{FP}}(\cdot)$ guarantees that if overload did occur, it is only for entries whose magnitude is about $2^{-(2^{\mathcal{E}}-3)}$ times smaller than the largest magnitude entry. Thus, if $n \ll 2^{-2 \cdot (2^{\mathcal{E}}-3)}$, the contribution of these entries to the total MSE in quantizing \tilde{x} is negligible. The assumption that $e_{\tilde{x}_i}$ is uniform is similar to the uniform quantization noise assumption we have made when analyzing INTM constellations, and it becomes more accurate as the number of mantissa bits \mathcal{M} grows.

Under the IEIN model, the approximation error is

$$e_{\text{FP}} = \sum_{i=1}^n \gamma_x \gamma_y \hat{\tilde{x}}_i \hat{\tilde{y}}_i - x_i y_i \\ = \sum_{i=1}^n x_i y_i (1 + \rho_{\tilde{x}_i} e_{\tilde{x}_i}) (1 + \rho_{\tilde{y}_i} e_{\tilde{y}_i}) - x_i y_i \\ = \sum_{i=1}^n x_i y_i (\rho_{\tilde{x}_i} e_{\tilde{x}_i} + \rho_{\tilde{y}_i} e_{\tilde{y}_i} + \rho_{\tilde{x}_i} \rho_{\tilde{y}_i} e_{\tilde{x}_i} e_{\tilde{y}_i}).$$

By our assumptions that $(e_{\tilde{x}_i}, e_{\tilde{y}_i}, \rho_{\tilde{x}_i}, \rho_{\tilde{y}_i})$ are indepen-

dent, and further assuming $\{e_{\bar{x}_i}, e_{\bar{y}_i}\}$ are iid, we have

$$\mathbb{E}[e_{\text{FP}}^2] = \frac{2^{-2\mathcal{M}}}{12} \sum_{i=1}^n x_i^2 y_i^2 \left(2C_{\text{FP}} + (C_{\text{FP}})^2 \frac{2^{-2\mathcal{M}}}{12} \right), \quad (14)$$

where

$$C_{\text{FP}} = \mathbb{E}[2^{-2U}] \approx 0.541, \quad U \sim \text{Uniform}([0, 1]) \quad (15)$$

Neglecting the last term, we have

$$\begin{aligned} D_{\text{FP}, \mathcal{M}} &= \mathbb{E}[e_{\text{FP}}^2] \lesssim 2C_{\text{FP}} \cdot \frac{2^{-2\mathcal{M}}}{12} \sum_{i=1}^n x_i^2 y_i^2 \\ &= \frac{\|x\|^2 \cdot \|y\|^2}{n} 2 \cdot 2^{-2R_{\text{eff}}(\text{FP}, \mathcal{M})} \cdot \Delta_{\text{FP}}(x, y), \end{aligned} \quad (16)$$

where

$$R_{\text{eff}}(\text{FP}, \mathcal{M}) = \mathcal{M} + \frac{1}{2} \log \left(\frac{12}{C_{\text{FP}}} \right) \approx \mathcal{M} + 2.2356, \quad (17)$$

and

$$\Delta_{\text{FP}}(x, y) = n \sum_{i=1}^n \frac{x_i^2}{\|x\|^2} \cdot \frac{y_i^2}{\|y\|^2}. \quad (18)$$

In Section III-E, Table I it is numerically verified that the approximation (16) is remarkably accurate.

Evidently, the gap from the fundamental limit is dictated by $\Delta_{\text{FP}}(x, y)$. It is easy to see that $0 \leq \Delta_{\text{FP}}(x, y) \leq n$. If at least one of the vectors x or y is drawn from an iid distribution, independently of the other vector, it immediately follows that $\mathbb{E}[\Delta_{\text{FP}}(x, y)] = 1$. Furthermore, if we apply a random rotation prior to quantization, such that $x \leftarrow Sx$ and $y \leftarrow Sy$, using the Cauchy-Schwartz inequality we have that

$$\begin{aligned} \mathbb{E}[\Delta_{\text{FP}}(x, y)] &\leq n \sum_{i=1}^n \left(\mathbb{E} \left(\frac{x_i}{\|x\|} \right)^4 \cdot \mathbb{E} \left(\frac{y_i}{\|y\|} \right)^4 \right)^{1/2} \\ &= \frac{3n}{n+2} < 3, \end{aligned}$$

where we have used the fact [1, Appendix D] that if $U \sim \text{Uniform}(\alpha \mathbb{S}^{n-1})$, where $\mathbb{S}^{n-1} = \{z \in \mathbb{R}^n : \|z\| = 1\}$ is the unit sphere, then $\mathbb{E}(U_1^4) = \frac{\alpha^4}{n^2} \frac{3n}{n+2}$.

For the Gaussian case where $x, y \sim \mathcal{N}(0, \sigma^2 I_n)$ and are statistically independent, we have $\mathbb{E} \left[\frac{\|x\|^2 \cdot \|y\|^2}{n} \right] = \sigma^4 n$ and $\mathbb{E}[\Delta_{\text{FP}}(x, y)] = 1$ (this holds for any iid distribution with zero-mean and variance σ^2). Furthermore, we have that $\frac{\|x\|^2 \cdot \|y\|^2}{n}$ and $\Delta_{\text{FP}}(x, y)$ are statistically independent. We consequently obtain that for Gaussian iid x, y

$$\begin{aligned} &\frac{1}{2n\sigma^4} \mathbb{E}(\widehat{x^\top y} - x^\top y)^2 \\ &\lesssim D_{\text{FP}, \mathcal{M}, \text{Gaussian}} \triangleq 2^{-2R_{\text{eff}}(\text{FP}, \mathcal{M})}. \end{aligned} \quad (19)$$

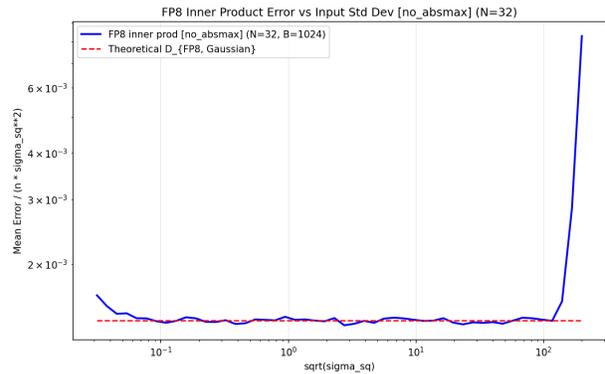


Fig. 2: Demonstrating accuracy of FP8 approximation (19). The figure plots average ratio (over 1024 pairs of vectors of dimension 32, each generated iid $\mathcal{N}(0, \sigma^2)$) of the normalized squared error of the inner-product (see left-hand side of (19)) against the simple theoretical approximation $2^{-2R_{\text{eff}}(\text{FP}, \mathcal{M})+1}$.

It turns out that this approximation is accurate even without dithered absmax scaling for a wide range of σ 's and even small dimensions n (see Fig. 2).

C. Multi-scaled INT and FP constellations

A current trend in LLMs inference is to use 4 bit multipliers whenever possible. Those include INT4 multipliers corresponding to entries in $\mathbb{Z} \cap [-8, 7]$, and FP4 multipliers corresponding to entries in

$$\mathcal{FP}_4 = \{0, \pm \frac{1}{2}, \pm 1, \pm \frac{3}{2}, \pm 2, \pm 4, \pm 6\}. \quad (20)$$

Recall that for INTM multipliers with absmax scaling, the effective rate given by (10) is $\frac{1}{2} \log \left(\frac{2n}{3} \right)$ bits smaller than M , even when random rotation is applied. For M as small as 4, and even moderate n , using absmax scaling therefore decreases $R_{\text{eff}}(\text{INTM})$ significantly with respect to M . The FP4 constellation suffers from a similar problem: since it uses only $E = 2$ exponent bits, overload errors are common under absmax scaling, and the attained distortion is far greater than that predicted by the IEIN model.

Consequently, those constellations are seldom used with absmax scaling. Instead, a multi-scaling procedure is applied. Here we describe the NV multi-scaling mechanism: In order to compute $x^\top y$, for $x, y \in \mathbb{R}^n$, each vector is first scaled by a global scale, as in absmax scaling, but afterwards it is split to consecutive sub-blocks of size 16. Then, each sub-block is scaled by an E4M3 FP8 number, such that that the absolute value of the maximal entry in the scaled sub-block is as close as

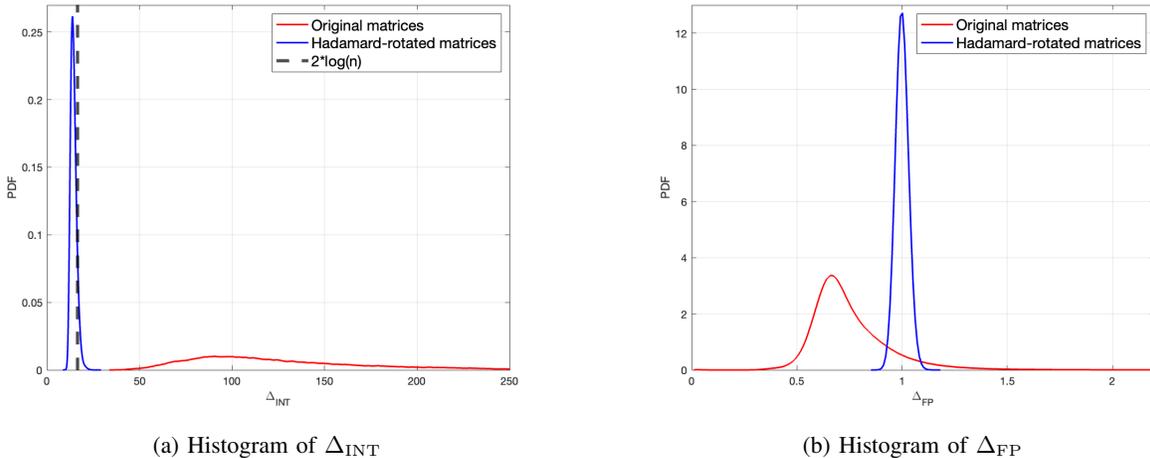


Fig. 3: Histograms of Δ_{FP} and Δ_{INT} for the setup described in Section III-E.

possible to 7 for INT4 or 6 for FP8.⁷ Consequently, both NVINT4 and NVFP4 actually have rate $4 + \frac{8}{16} = 4\frac{1}{2}$ bits per entry.

We now make the heuristic argument that the distortion attained by NVFP4 is typically upper bounded by the distortion predicted by the IEIN model for 1 mantissa bit (which has $R_{\text{eff}}(\text{FPM}) = 3.2356$). Let us ignore the finite resolution of the E4M3 scale assigned to each vector $\tilde{x} \in \mathbb{R}^{16}$, and assume for the analysis below that this scale is given in full resolution. Then, the input to the FP4 quantizer is $6 \frac{\tilde{x}}{\|\tilde{x}\|_{\infty}}$. Assume without loss of generality that $\|\tilde{x}\|_{\infty} = |\tilde{x}_1|$, and define the set of coordinates $\mathcal{OL} = \{i \in \{2, \dots, 16\} : 6 \frac{|\tilde{x}_i|}{\|\tilde{x}\|_{\infty}} < 1\}$. For all the coordinates in $[16] \setminus \{1 \cup \mathcal{OL}\}$ there is no exponent overload, and the IEIN model is a good approximation. In contrast, for $i = 1$ and $i \in \mathcal{OL}$ the IEIN model is not a good approximation. For $i = 1$, we have that \tilde{x}_1 is scaled exactly to a point $\pm 6 \in \mathcal{FP}_4$, such that it suffers no quantization error. On the other hand, for all $i \in \mathcal{OL}$, we have that the FP4 representation of $6\tilde{x}_i/\|\tilde{x}\|_{\infty}$ is in overload (its corresponding exponent is too small), and the IEIN model is overly optimistic. Since each one of these entries is quantized to either $\{0, \pm 1/2, \pm 1\}$, their total contribution to the distortion (after rescaling by $\|\tilde{x}\|_{\infty}/6$) is upper bounded by $|\mathcal{OL}| \cdot \left(\frac{1}{4}\right)^2 \left(\frac{\|\tilde{x}\|_{\infty}}{6}\right)^2$. Whenever $|\mathcal{OL}| \leq 6 < \frac{576}{48/C_{\text{FP}}}$, this is smaller than $\frac{1}{4} \frac{\|\tilde{x}\|_{\infty}^2}{12} C_{\text{FP}}$ which was “saved” (with respect to the IEIN model) by perfectly aligning the largest entry to the \mathcal{FP}_4 constellation. Thus, for vectors in \mathbb{R}^{16} with $|\mathcal{OL}| \leq 6$,

⁷Alternatively, for each sub-block one often chooses the minimal scale such that the largest value is at most 7 (INT) or 6 (FP). Also, while NVFP4 is a popular data-type supported in Nvidia’s modern GPUs, NVINT4 is far less popular and is described here only for comparison purposes.

the distortion for NVFP4 quantization is smaller than that predicted by the IEIN model with $\mathcal{M} = 1$.

D. NestQuant

The INTM constellation admits very simple quantization and de-quantization procedures. However, it suffers from two shortcomings: 1)The distortion in quantization of a vector $x \in \mathbb{R}^n$ depends on $\|x\|_{\infty}$ rather than on $\|x\|_2$; 2)The induced quantization cells are cubic, and therefore suffer from highly sub-optimal volume/second-moment tradeoff.⁸ The NestQuant framework [1], [22], which is the topic of this subsection, provides significant improvements with only a slight increase in the quantization/de-quantization complexity.

We review some basic lattice definitions. See [9] for a comprehensive treatment of lattices in information theory. For a lattice $L \subset \mathbb{R}^d$ we define the nearest neighbor quantizer $Q_L : \mathbb{R}^d \rightarrow L$ as

$$Q_L(x) = \underset{\lambda \in L}{\operatorname{argmin}} \|x - \lambda\|, \quad (21)$$

where ties are broken arbitrarily, but in systematic manner. The Voronoi region \mathcal{V}_L is defined as the set of all points in \mathbb{R}^n that are closer to 0 than to any other lattice point

$$\mathcal{V}_L = \{x \in \mathbb{R}^d : Q_L(x) = 0\}. \quad (22)$$

Any lattice $L \subset \mathbb{R}^d$ has a (non-unique) generating matrix $G \in \mathbb{R}^{d \times d}$ such that $L = G\mathbb{Z}^d$. The covolume of the lattice L , denoted $\operatorname{covol}(L)$, is the volume of its Voronoi region (or any other fundamental cell of L), which is also equal to $|G|$. The point density of a lattice is $\gamma(L) =$

⁸In fact, the strong slicing conjecture postulates that among all unit-volume convex sets at isotropic position, the cube has the largest second moment. See [20, Section 6.1] and [21]

$\text{covol}^{-1}(L) = |G|^{-1}$. We define the second moment of the lattice L as

$$\sigma^2(L) = \frac{1}{d} \mathbb{E} \|Z\|^2, \quad (23)$$

where $Z \sim \text{Uniform}(\mathcal{V}_L)$ is a random vector uniformly distributed over the Voronoi region of L .

For a lattice $L \subset \mathbb{R}^d$ and an integer $q \geq 2$ we have that $qL \subset L$ forms a *self-similar* nested lattice pair. The lattice qL is referred to as the *coarse lattice* and it forms a partition of L to q^d cosets

$$L = \bigcup_{i=1}^{q^d} (x_i + qL),$$

where $x_i \in L$, $i = 1, \dots, q^d$ are coset representatives. Note that any point in the coset $x_i + L$ can be chosen as the representative of this coset. Any choice of coset representatives induces a nested lattice codebook L/qL consisting of q^d points. A particularly useful choice is Voronoi codes [23], introduced by Conway and Sloane, where for each $i = 1, \dots, q^d$ the coset representative is chosen as the minimum energy member of this coset. In particular, the obtained codebook is of the form $L \cap q\mathcal{V}_L$. Encoding of x to $[q]^d$ amounts to computing $v = \text{Enc}(x) = [G^{-1}Q_L(x)] \bmod q$, where here $\bmod q$ denotes component-wise modulo q reduction. Decoding amounts to computing $\text{Dec}(v) = Gv - q \cdot Q_L(Gv/q)$.⁹ Thus, if we have an efficient implementation of the nearest neighbor lattice quantizer $Q_L(\cdot)$, we can also implement the encoder and decoder of Voronoi codes efficiently. Notably, *the complexity of encoding and decoding does not grow with q* (thus, high-rate and low-rate quantization have the same computational cost). Furthermore, whenever $Q_L(x) \in q\mathcal{V}_L$ we have that $\text{Enc}(\text{Dec}(x)) = Q_L(x)$, and the quantization error is in \mathcal{V}_L . We refer to the event $Q_L(x) \notin q\mathcal{V}_L$ as an *overload* event.

Note that the INTM constellation is in fact a Voronoi code for the nested lattice pair $(\beta\mathbb{Z}^n/2^M\beta\mathbb{Z}^n)$ for $\beta > 0$. In order to avoid overload when quantizing $x \in \mathbb{R}^n$, the absmax scaling sets $\beta = 2^{-(M-1)}\|x\|_\infty$, and the corresponding second moment is $\beta^2\sigma^2(\mathbb{Z})$, where $\sigma^2(\mathbb{Z}) = \frac{1}{12}$.

NestQuant applies a random rotation S (in practice, randomized Hadamard transform) to both vectors $x, y \in \mathbb{R}^n$ prior to quantization. Then, it splits each rotated vector to chunks of size d , and quantizes each one of them separately. It improves over INTM by using:

- 1) **Better lattice:** Replacing \mathbb{Z}^d with a lattice $L \subset \mathbb{R}^d$ of the same covolume, that admits efficient nearest neighbor decoding $Q_L(\cdot)$ and has smaller second

⁹We ignore dithering here, for simplicity, though it is needed for preventing undesired boundary effects [24, Section II].

moment $\sigma^2(L) < \sigma^2(\mathbb{Z}) = 1/12$ and smaller overload probability

$$\Pr(Q_{\beta L}(X) \notin q\beta\mathcal{V}_L) < \Pr(Q_{\beta\mathbb{Z}^d}(X) \notin q\beta\mathcal{V}_{\mathbb{Z}^d}),$$

for $X \sim \mathcal{N}(0, I_d)$;

- 2) **Multi-scaling:** Using a bank of K different scales $\{\beta_1, \dots, \beta_K\}$ and using the scale that results in the smallest squared error. The effective codebook is therefore

$$\mathcal{C} = \bigcup_{k=1}^K \beta_k(L \cap q\mathcal{V}_L).$$

In [1] it was shown that for $d = n$ with n large enough, there exist lattices L for which NestQuant attains the optimal distortion-rate tradeoff for matrix multiplication. However, such lattices typically do not admit efficient $Q_L(\cdot)$. In [22] it was shown that by using $d = 8$ with the lattice $L = E_8$, which has a very fast $Q_L(\cdot)$, one obtains excellent performance, significantly outperforming other LLM quantization strategies at a similar bit-rate.

E. Numerical results

The goal of this section is to give numerical validation to our analysis of INT and FP constellations, which relied on some approximations, and compare performance of various quantizers in a realistic setup. To this end we focus on one particular linear operation within an LLM, and consider the $W_v \in \mathbb{R}^{n \times a}$ matrix in the 15th layer of Llama3-8B, along with its corresponding activation vectors. Here, $n = 4096$, $a = 1024$, and the matrix $X \in \mathbb{R}^{b \times n}$ consists of $b = 10,000$ rows, which were sampled by running 5 different prompts on the model and taking about 2048 consecutive activation vectors for each prompt.

For these matrices we define 3 matrices of size $b \times a$:

$$\begin{aligned} K(i, j) &= 2 \frac{\|X(i, :)\|^2 \|W(:, j)\|^2}{n} \quad i \in [b], j \in [a] \\ \Delta_{\text{INT}}(i, j) &= \Delta_{\text{INT}}(X(i, :), W(:, j)) \quad i \in [b], j \in [a] \\ \Delta_{\text{FP}}(i, j) &= \Delta_{\text{FP}}(X(i, :), W(:, j)) \quad i \in [b], j \in [a], \end{aligned}$$

where Δ_{INT} and Δ_{FP} are defined in (8) and (18), respectively.

First, we explore the accuracy of our approximations for D_{INTM} and D_{FPM} , as given in (7) and (16). To this end we quantize X and W using INT8 absmax quantization and using FP8 (E4M3) dithered absmax quantization. We denote the approximation error of XW using these methods by $e_{\text{INT8}} \in \mathbb{R}^{b \times a}$ and $e_{\text{FP8}} \in \mathbb{R}^{b \times a}$, respectively. Table I provides the root mean squared error (RMSE) of the ab error entries with and without

Hadamard rotation, under different normalizations. We denote

$$\bar{e}_{\text{INT8}}(i, j) = \frac{e_{\text{INT8}}(i, j)}{\sqrt{K(i, j)\Delta_{\text{INT}}(i, j)/3}}, \quad i \in [b], j \in [a]$$

$$\bar{e}_{\text{FP8}}(i, j) = \frac{e_{\text{FP8}}(i, j)}{\sqrt{K(i, j)\Delta_{\text{FP}}(i, j)}} \quad i \in [b], j \in [a].$$

If our expressions for $D_{\text{INT}M}$ and $D_{\text{FP}M}$ are accurate, then for $I \sim \text{Uniform}([b])$ and $J \sim \text{Uniform}([a])$ it holds that

$$\mathbb{E}(\bar{e}_{\text{INT8}}^2(I, J)) = 2^{-2M}, \quad \mathbb{E}(\bar{e}_{\text{FP8}}^2(I, J)) = 2^{-2R_{\text{eff}}(\text{FP})}.$$

The first row of Table I shows that those approximations are indeed remarkably accurate.

The second row of Table I provides the RMS of the quantization errors normalized by $\sqrt{K(i, j)}$. Recall from (4) that the fundamental limit for generic rate R quantized MatMul is $D_{ij} = K(i, j) \cdot 2^{-2R}$, assuming $R \gg 1$. Thus, if the optimal high-dimensional rate $R = 8$ quantizers from [1] were used instead of the various INT8/FP8 schemes investigated, the normalized RMSE would have been 2^{-8} . The second row in Table I therefore shows how many of the 8 bits each of the scheme is using are “effective”.

The third row of Table I provides the RMS of the quantization errors normalized by $\sqrt{2n}$ for the case where X, W are of the same shape as above, but their entries are drawn iid $\mathcal{N}(0, 1)$. The results verify the accuracy of our expressions for $D_{\text{INT}M, \text{Gaussian}}$ and $D_{\text{FP}M, \text{Gaussian}}$ in (11) and (19), respectively. Note that, as expected, the effective rate for rotated INTM quantization is the same as that of INTM quantization for Gaussian matrices.

In order to obtain a deeper understanding of where the numbers in the second row in Table I come from, we explore the distribution of $\{\Delta_{\text{INT}}(i, j)\}_{i \in [b], j \in [a]}$ and $\{\Delta_{\text{FP}}(i, j)\}_{i \in [b], j \in [a]}$. In Figure 3a we plot the histogram of the $a \cdot b$ entries of the matrix Δ_{INT} with and without applying Hadamard rotation on both matrices. We see that for the original matrices Δ_{INT} can take very large values, whereas for Hadamard rotated matrices the entries of Δ_{INT} concentrate below $2 \ln(n)$. Thus, rotating the matrices prior to INT quantization is crucial (at least for Llama3-8B).

In Figure 3b we plot the histogram of the $a \cdot b$ entries of the matrix Δ_{FP} with and without applying Hadamard rotation on both matrices. Here, the situation is very different. While Hadamard rotation makes the distribution concentrated around 1, we see that without rotation the entries of Δ_{FP} are typically smaller than 1. This shows that not only is full-vector rotation unnecessary for FP quantization, it is actually harmful! Of course, this conclusion is specific for the matrices we used for this experiment. However, prior work [25,

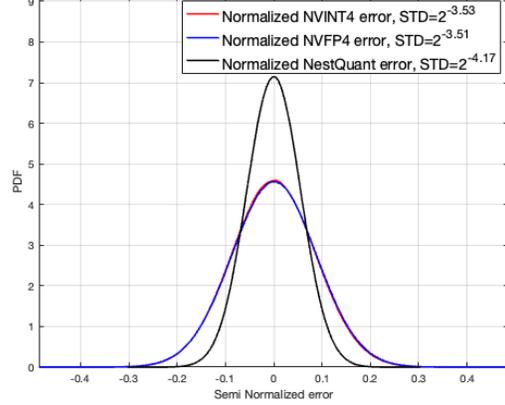


Fig. 4: Performance of several quantization schemes with $R = 4.5$.

Table 1], [26] have already observed that rotation is often harmful for FP LLM quantization, and we argue that the behavior of Δ_{FP} provides a rigorous explanation for this phenomenon.

Figure 4 compares the approximation error distribution of 3 quantization schemes with rate $R = 4.5$ bits per entry: NVINT4 after Hadamard rotation, NVFP4 without Hadamard rotation, and NestQuant with a bank of $K = 16$ different scales. For all schemes, we normalize the approximation error of the ij th entry by $\sqrt{K(i, j)}$. The performance of NVINT4 and NVFP4 is seen to be quite similar, while that of NestQuant is significantly better and attains ≈ 0.6 more bits of accuracy with respect to the former schemes.

IV. WEIGHT QUANTIZATION: THEORY

Modern LLMs often include a huge number of parameters, and cannot be stored in full-precision in the high bandwidth memory (HBM) of a single GPU. This problem can often be solved by weight quantization to R bits per parameter where R is the ratio between the HBM capacity and the model’s size. Note that only the weight matrices need to be quantized in this scenario, as the activations are computed online and do not need to be stored.¹⁰

Let us restrict attention to a particular linear layer in the network with weight matrix $W \in \mathbb{R}^{n \times a}$. The linear layer operates by taking (a vector of) input activations X and producing

$$Y = X^T W.$$

For that, W needs to be loaded from memory and as we discussed above this results in the crucial bottleneck

¹⁰We ignore the capacity required for the KV cache in this discussion.

	INT8	Hadamard INT8	FP8	Hadamard FP8
RMS $(\bar{e}_{\text{INT8/FP8}}(i, j))$	$2^{-7.9994}$	$2^{-8.0012}$	$2^{-5.2346}$	$2^{-5.2372}$
RMS $\left(\frac{e_{\text{INT8/FP8}}(i, j)}{\sqrt{K(i, j)}}\right)$	$2^{-5.2495}$	$2^{-6.8664}$	$2^{-5.4378}$	$2^{-5.2370}$
RMS $\left(\frac{e_{\text{INT8/FP8}}(i, j)}{\sqrt{2n}}\right)$ iid $\mathcal{N}(0, 1)$ entries	$2^{-6.8619}$	$2^{-6.8645}$	$2^{-5.2395}$	$2^{-5.2383}$

TABLE I: RMS of normalized errors

during generation/reasoning part of the LLM operation. Thus, our job is to replace W with quantized version \hat{W} , which can be loaded much faster, with the target of minimizing *expected* distortion over random inputs X , that is we aim to minimize

$$D = \frac{1}{n} \mathbb{E} \|X^\top (W - \hat{W})\|_F^2 = \frac{1}{n} \text{tr}(W - \hat{W})^\top \Sigma_X (W - \hat{W}), \quad (24)$$

where $\Sigma_X = \mathbb{E}[XX^\top]$ (This objective is by far the most popular one, though, others exist [27], [28].)

Since only the second order statistics Σ_X affect this objective, in practice one uses a set of samples (known as *calibration data*) from the underlying distribution in order to estimate it. With the second order statistics at hand, the problem of quantizing W becomes a weighted mean squared error (WMSE) quantization problem. We note that calibration data about X is also useful for quantization of both weights and activations (i.e. the setting of Section II). See e.g. [13], [4, Section 4.5], [29], [30].

Weight matrices start life as iid Gaussian and evolve during training. However, in many ways their statistics are still largely very similar to iid Gaussian and that is how we will model them in this (theoretical) section. Some (mostly older) LLMs have peculiar aspects of weight matrices, such as existence of outliers and rank deficiencies, that need to be exploited/mitigated in practical algorithms, though.

Notice also that objective (24) treats distortion across columns of W in an additive manner. Consequently, analyzing the case of $a = 1$ (single column, or single output neuron) can be done without loss of generality, at least as long as $n \gg 1$ and full advantage of vector quantization can be exploited already in dimension n , without needing to do joint vector quantization across all na dimensions.

So in summary, the goal of this section is to understand theoretically the problem of mapping $W \sim \mathcal{N}(0, I_n)$ to \hat{W} belonging to the universe of 2^{nR} possibilities, with the goal of minimizing

$$D = \frac{1}{n} (W - \hat{W})^\top \Sigma_X (W - \hat{W}). \quad (25)$$

One important aspect of the weight only quantization problem worth pointing out from the outset is the asymmetry between encoding and decoding. The encoding

(quantization) is done offline, once for each model, and may therefore be computationally demanding. The decoding (de-quantization), on the other hand, is executed online every time a weight matrix is used, and must therefore be highly efficient. (This is in contrast with activation quantization, which is done online and has to be extremely computationally efficient.) In particular, while information about Σ_X is available to the encoder (recall that it operates offline and can be quite slow), the decoder's online operation requires all information about Σ_X be packaged inside the rate-constrained description of W . Thus, assuming decoder knows Σ_X or its SVD basis (but not its spectrum) is not reflecting the practical constraints adequately.

Nevertheless, below we will start by treating the impractical *fully informed case*, which will provide a firm lower bound to the more practical version of *un-informed/oblivious decoder case*. We initiate our discussion with a simple heuristic demonstration of the main points on the example of scalar quantization.

A. Scalar quantization and waterfilling

Before discussing information theoretic results, let us consider a very simple case of X with uncorrelated coordinates, i.e. $\Sigma_X = \text{diag}\{\lambda_1, \dots, \lambda_n\}$. Suppose, in addition, that weights $W_i \in [-1/2, 1/2]$.

Consider, first, the case of usual uniform quantization using ϵ -grid, where $\epsilon = 2^{-R}$. The resulting distortion (in accordance with additive noise approximation (1)) will be

$$D_1(R) = \frac{1}{n} \sum_{i=1}^n \lambda_i \frac{\epsilon^2}{12} = \frac{\bar{\lambda}_A}{12} 2^{-2R},$$

where $\bar{\lambda}_A = \frac{1}{n} \sum_i \lambda_i$ is the arithmetic mean of λ 's.

Now, given that different coordinates of W contribute differently to D , one may naturally try to allocate rate more judiciously. Specifically, let us solve the problem of

$$D_2 = \min \frac{1}{12n} \sum_{i=1}^n \lambda_i \epsilon_i^2,$$

subject to rate constraint $\prod \frac{1}{\epsilon_i} \leq 2^{nR}$. Using Lagrange multipliers we can easily find a parametric formula for

the optimizer, given in terms of parameter $\tau > 0$:

$$D_2 = \frac{1}{n} \sum_{i=1}^n \min(\tau, \lambda_i/12)$$

$$R = \frac{1}{2n} \sum_{i=1}^n \log \max\left(1, \frac{\lambda_i}{12\tau}\right),$$

where the grid spacings $\epsilon_i = \min(\sqrt{\frac{12\tau}{\lambda_i}}, 1)$. This kind of allocation is traditionally called (reverse) *waterfilling solution*, due to interpretation of τ as water level that reduces i 'th coordinate baseline distortion (equal to $\lambda_i/12$) down to τ . Those coordinates that are so insignificant ("below waterlevel") get zero rate allocation and are quantized to zero.

In accordance with our focus on high-rate case, we can assume that $\tau < \min_i \lambda_i$ and get a simplified expression:

$$D_2(R) = \frac{\bar{\lambda}_G}{12} 2^{-2R},$$

where $\bar{\lambda}_G = (\prod_i \lambda_i)^{1/n}$ is the geometric mean of λ_i 's.

In all, we conclude that when quantizing vectors under weighted quadratic metric, one should choose quantization grid spacing $\epsilon_i \propto \frac{1}{\sqrt{\lambda_i}}$ with coefficient of proportionality chosen depending on the target required rate R . How much does one win from this optimization is given by the AM-GM inequality $\bar{\lambda}_G < \bar{\lambda}_A$. Another way to say is that optimizing per-coordinate quantizers one wins about $\frac{1}{2} \log \frac{\bar{\lambda}_A}{\bar{\lambda}_G}$ bits of rate.

We will return to this idea below, when we describe a practical WaterSIC algorithm for weight-only quantization in Section V-D. There the role of λ_i 's is played by the diagonal elements of the Cholesky decomposition of (non-diagonal) Σ_X .

Fig. 5 illustrates rate-advantage for input activations to various linear layers of Llama-3-8B.

B. WMSE Quantization: information theoretic limit

The waterfilling solution derived heuristically in the previous section can be in fact made rigorous, at least for the case of Gaussian weights $W \sim \mathcal{N}(0, \sigma_W^2 I_n)$, which we consider in this subsection. Another assumption we make is that the covariance matrix $\Sigma_X \in \mathcal{P}\mathcal{S}\mathcal{D}_n$ (the set of all $n \times n$ PSD matrices) is fixed and known to both encoder and decoder.

The formal problem is this: given $W \sim \mathcal{N}(0, \sigma_W^2 I_n)$, encoder $f : \mathbb{R}^n \rightarrow [2^{nR}]$ produces a rate- R description of W . Subsequently, decoder $g : [2^{nR}] \rightarrow \mathbb{R}^n$ converts this description into the estimate \hat{W} . The information-theoretic question is to determine the value of the smallest distortion attainable by the best pair (f, g) , i.e.

$$D^*(\Sigma_X, R) = \min_{f, g} \frac{1}{n} \mathbb{E}[(W - \hat{W})^\top \Sigma_X (W - \hat{W})].$$

The naive (suboptimal) solution is to use a standard isotropic Gaussian codebook of rate R , which results in error covariance given by $\mathbb{E}[(W - \hat{W})(W - \hat{W})^\top] = \sigma_W^2 2^{-2R} I_n$. Note that the previous result is simple if we understand \mathbb{E} as averaging over the Gaussian codebook, but is rather non-trivial if we want to demonstrate existence of a codebook with nearly-white covariance matrix of errors. This was the main technical difficulty behind the results of [1], cf. Theorem 13 there. Nevertheless, under this assumption on the error distribution, we get

$$D_{iso}(R) = 2^{-2R} \frac{\sigma_W^2}{n} \text{tr} \Sigma_X. \quad (26)$$

Although this is generally very far from $D^*(\Sigma_X, R)$, such scheme works simultaneously for all Σ_X and does not require knowledge thereof.

Scheme that is allowed to fully exploit knowledge of Σ_X can first compute $W' = V^\top W$, where V is the orthogonal matrix in the SVD decomposition of $\Sigma_X = V^\top \Lambda V$. If decoder can estimate \hat{W}' , then it can also set $\hat{W} = V\hat{W}'$. Then, the distortion can be expressed in terms of (W', \hat{W}') as

$$D = \frac{1}{n} \sum_{i=1}^n \lambda_i \mathbb{E}[(\hat{W}'_i - W'_i)^2].$$

We see that after this change of coordinates, the problem indeed becomes that of *weighted* mean-squared error (WMSE), justifying the name.

To relate distortion and rate, we consider a standard data-processing argument, see [31, Section 23.4]:

$$nR \geq I(W'; \hat{W}') \geq \sum_{i=1}^n I(W'_i; \hat{W}'_i),$$

where the second inequality is due to independence of coordinates of $W' \sim \mathcal{N}(0, \sigma_W^2 I_n)$, cf. [31, Theorem 6.1]. Now, given value $D_i = \mathbb{E}[(\hat{W}'_i - W'_i)^2]$ the smallest $I(W'_i; \hat{W}'_i) = \frac{1}{2} \log \frac{\sigma_W^2}{D_i}$ is attained under Gaussian coupling, cf. [31, Section 26.1.2]. Overall, we get

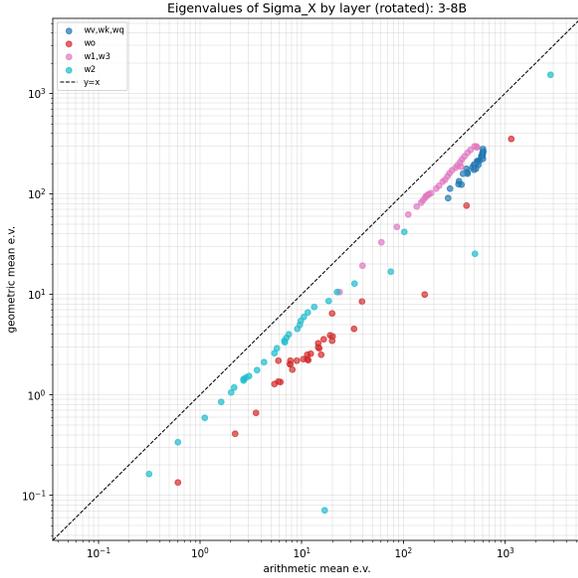
$$R \geq \frac{1}{2n} \sum_{i=1}^n \log \frac{\sigma_W^2}{D_i}$$

$$D = \frac{1}{n} \sum_{i=1}^n \lambda_i D_i.$$

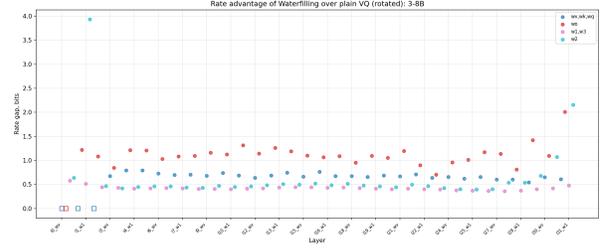
Minimizing the value of D given R can be done via Lagrange multipliers, resulting in the (reverse) *waterfilling solution* parameterized by $\tau > 0$:

$$D^*(\tau) = \frac{\sigma_W^2}{n} \sum_{i=1}^n \min\{\lambda_i, \tau\},$$

$$R^*(\tau) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \log \max\left\{1, \frac{\lambda_i}{\tau}\right\}. \quad (27)$$



(a) Scatter plot illustrating gap in AM-GM inequality for eigenvalues of Σ_X .



(b) Rate advantage (in high-rate regime) of waterfilling over D_{iso} (which corresponds to the case of Σ_X -oblivious encoder and decoder). Squares correspond to (skipped) layers with singular Σ_X .

Fig. 5: Illustrating Σ_X of activations entering various layers of Llama-3-8B when processing Wikitext-2 dataset. Note that this is an estimate of the rate advantage assumes weight matrices are well modeled by $\mathcal{N}(0, I_n)$. In particular, actual weight matrices were never used for this plot.

Note that, while our argument only gives a lower bound on the minimal distortion, it can be shown to be achievable asymptotically for large n . (This can either be done by grouping adjacent λ_i 's and jointly vector-quantizing corresponding groups of coordinates; or we can recall that the actual problem requires compressing $W \in \mathbb{R}^{n \times a}$ and we can do Gaussian vector quantization over a -long vectors as long as $a \gg 1$.) Thus, in the remainder of this paper, we will consider the value $D^*(R)$ given by waterfilling to be the actual fundamental limit, ignoring small non-asymptotic penalty.

Just like in the previous heuristic section, we see that optimal codebook allocates rate unequally: the PCA direction corresponding to high values of λ_i should get quantized more finely with $D_i \propto \frac{1}{\lambda_i}$. In the high-rate regime we get

$$D_{\text{High-Rate}}^*(R) = |\Sigma_X|^{1/n} \sigma_W^2 2^{-2R}, \quad (28)$$

and the advantage compared to uninformed isotropic coding is again given by the gap in the AM-GM inequality, or

$$\frac{1}{2} \log \frac{\frac{1}{n} \text{tr} \Sigma_X}{|\det \Sigma_X|^{1/n}}$$

bits in effective rate.

C. Σ_X -oblivious decoder

In the setup above, we have allowed the decoder to depend on Σ_X (in particular, to apply SVD decomposition). However, in LLM applications this would require communicating Σ_X (or the SVD basis matrix V) to GPU's compute devices, which would be very costly. Instead, we would like to use a very lean decoder, that only uses those bits provided with the description of W to produce its reconstruction, universally for all Σ_X .

We will therefore require that while the encoder $f(W, \Sigma_X)$ may depend on Σ_X , the decoder $g: [2^{nR}] \rightarrow \mathbb{R}^n$ depends only on the bits it receives from the encoder. With this constraint, the quantization problem consists of a codebook $\mathcal{C} \subset \mathbb{R}^n$ with $|\mathcal{C}| \leq 2^{nR}$ codewords, agreed upon by the encoder and decoder, where the encoder's job is to solve, or approximately solve, the problem

$$\hat{W}^* = \underset{c \in \mathcal{C}}{\text{argmin}} [(W - c)^\top \Sigma_X (W - c)]. \quad (29)$$

The encoder then describes \hat{W}^* using nR bits, and the decoder can reconstruct \hat{W}^* from those bits.

We know from the previous section that in order to achieve the informed (waterfilling) $D^*(R)$ optimal codebook needs to be anisotropic: the “grids” need to be denser along the PCA directions corresponding to higher values of λ_i . In the uninformed case, we are not able to do such adaptation (since Σ_X is unknown at the time of the codebook design). Thus, we are forced to

use isotropic codebook. In a forthcoming work [32] a somewhat surprising result is shown: if one generates \mathcal{C} iid from isotropic Gaussian distribution, then simultaneously for all Σ_X the adaptive encoder (29) achieves the following (parametric) rate-distortion:

$$D_{rc}(R) = \frac{\tau\sigma_W^2}{n} \sum_{i=1}^n \frac{\lambda_i}{\tau + \lambda_i}$$

$$R = \frac{1}{2n} \sum_{i=1}^n \log\left(1 + \frac{\lambda_i}{\tau}\right).$$

First, observe that in the high-rate regime $R \gg 1$ (i.e. $\tau \ll 1$), we get

$$D_{rc}(R) = |\Sigma_X|^{1/n} \sigma_W^2 2^{-2R} + O(2^{-4R}).$$

Comparing this to (28), we observe that in the high-rate regime this scheme attains the optimal waterfilling distortion (upto higher order corrections), *despite decoder not knowing* Σ_X . Furthermore, analysis in [32] shows that universally over all possible Σ_X the gap between waterfilling solution $D^*(R)$ and $D_{rc}(R)$ is never more than 0.1 bit.

This result demonstrates that the most significant part of the gap between (26) and (27) is not due to sub-optimal codebook design, but rather due to suboptimal rounding to that codebook. In other words, one must focus on implementing better rounding schemes rather than optimizing codebooks. Unfortunately, as we will see in the next section, even for the uniform quantization (\mathbb{Z}^n codebook), computing (29) is computationally hard. Thus, despite theoretical near-optimality of isotropic codebooks, the algorithmic difficulty makes Σ_X -oblivious quantization very interesting and rather underexplored. A practical low-complexity algorithm, known variously as successive-interference cancellation (SIC), Babai's algorithm, GPTQ or LDLQ, will be discussed in the next section.

V. WEIGHT QUANTIZATION: PRACTICE

In this section we discuss approximate solutions for the WMSE quantization problem, that can be computed efficiently. We will describe several variants of the GPTQ/LDLQ algorithms [2], [3], and propose novel improvements based on our information theoretic analysis. The schemes we discuss rely on a reformulation of the ‘‘optimal rounding’’ problem (29) using the Cholesky decomposition [33], [34]. We develop this reformulation in Subsection V-A. The analysis of the quantization schemes that follow requires some background in high-resolution quantization using lattices. We provide this background in Subsection V-B, and then in Subsection V-C we explain and analyze the SIC rounding algorithm. In its most basic form, this algorithm is equivalent to the canonical GPTQ/LDLQ [2], [3], as

was recently observed in [33], [34]. In light of the information theoretic analysis above, and the analysis of the SIC algorithm, it immediately becomes clear that a simple variation of GPTQ/LDLQ called *WaterSIC*, that we develop in Subsection V-D, provides an improved distortion and is provably quite close to the lower bound (28) we developed on the WMSE.

A. Reformulation of (29) via Cholesky Decomposition

Using the Cholesky decomposition, we can decompose Σ_X as $\Sigma_X = U^T U$ where $U \in \mathbb{R}^{n \times n}$ is an upper triangular matrix. Plugging this into (29) we obtain that given a fixed codebook \mathcal{C} the optimal quantization problem becomes

$$\hat{W}^* = \underset{c \in \mathcal{C}}{\operatorname{argmin}} \|Y - U \cdot c\|^2, \quad \text{where } Y = UW \in \mathbb{R}^n. \quad (30)$$

Denote

$$e_Y = Y - U\hat{W}^*, \quad e_W = W - \hat{W}^*, \quad (31)$$

and note that

$$\hat{W}^* = W + e_W = W + U^{-1}e_Y. \quad (32)$$

Recall that since $W \sim \mathcal{N}(0, \sigma_W^2 I_n)$, we have that $\Pr(W \notin \sqrt{n(1+\varepsilon)}\sigma_W^2 \mathcal{B}) \rightarrow 0$ for any $\varepsilon > 0$, as $n \rightarrow \infty$, where $\mathcal{B} = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$. For high-resolution quantization, where e_Y is so small that $U^{-1}e_Y$ can also be assumed small, this will imply that \hat{W}^* will typically also fall inside a ball with radius $\sqrt{n(1+\varepsilon_R)}\sigma_W^2 \mathcal{B}$, where ε_R vanishes with R (and n). This fact will be useful in the analysis below.

Before proceeding further, we note that Cholesky matrix' U diagonal entries will play here the same role as eigenvalues of Σ_X did in Section IV. In particular, we will see that algorithms whose codebooks have equal density in each direction (SIC/GPTQ) attain error

$$D_{\text{GPTQ}}(R) \approx \frac{2\pi e}{12} \sigma_W^2 \left(\frac{1}{n} \sum_i U_{i,i}^2\right) 2^{-2R}, \quad (33)$$

which depends on *arithmetic* mean of squared entries of diagonal of U (which is highly dependent on permutation of columns of Σ_X and application of random rotations). Our new scheme (‘‘WaterSIC’’) spends rate more judiciously (more bits to coordinates with higher $U_{i,i}$ and attains

$$D_{\text{WaterSIC}}(R) \approx \frac{2\pi e}{12} \sigma_W^2 \left(\prod_i U_{i,i}^2\right)^{1/n} 2^{-2R}. \quad (34)$$

On a first sight, we only get the same AM-GM improvement as with / without waterfilling. However, what is remarkable and much deeper is the fact that since $|\Sigma_X| = |U|^2$, we have $(\prod_i U_{i,i}^2)^{1/n} = |\Sigma_X|^{1/n}$, thus recovering (within factor $\frac{2\pi e}{12}$, which corresponds to Koshelev's famous 0.25 bit gap [31, Section 24.1.5]) the information-theoretic optimal waterfilling rate $D_{\text{High-Rate}}^*$, see (28).

B. Preliminaries on High-resolution quantization via shaped/entropy coded lattice quantizers

Designing a quantizer for the WMSE problem with Σ_X -oblivious decoder entails choosing the codebook \mathcal{C} , which consists of 2^{nR} vectors in \mathbb{R}^n . One way to construct such a codebook is to start with an infinite constellation, in our case a lattice $L \subset \mathbb{R}^n$, and then choose from L only 2^{nR} points [9]. This can be done by choosing a shaping region $\mathcal{S} \subset \mathbb{R}^n$, e.g. $\mathcal{S} = r\mathcal{B}$ for some $r > 0$, and taking $\mathcal{C} = L \cap \mathcal{S}$ where \mathcal{S} is dilated such that $|L \cap \mathcal{S}| = 2^{nR}$. Correspondingly, this way of thinking about vector quantization splits the problem into design of good infinite constellations (with prescribed point density) and shaping regions that chop off a finite part of it.

Many practical solutions for weight-only quantization that were considered in the literature fall under the shaped lattice quantization paradigm: GPTQ with INT constellations [2] as well as LDLQ with E_8 -based codebooks as in QuIP# [3], [35], and many more.

Assume the point $\hat{W}^* \in L$ that minimizes (30) with respect to the entire lattice L (rather than $L \cap \mathcal{S}$) is within the shaping region \mathcal{S} . In this case, $e_Y \in \mathcal{V}_{UL}$ and if we further assume $e_Y \sim \text{Uniform}(\mathcal{V}_{UL})$ the distortion is $\sigma^2(UL)$ (see lattice definitions in Section III-D). However, whenever the *overload event* $\hat{W}^* \notin \mathcal{S}$ occurs, the squared error may significantly exceed $\sigma^2(UL)$. In high-resolution quantization, we may assume $\hat{W}^* \approx W$, and the overload probability is well approximated by $\Pr(W \notin \mathcal{S})$. As mentioned above, taking $\mathcal{S} = \sqrt{n(1+\varepsilon)}\sigma_W^2\mathcal{B}$, with some small $\varepsilon > 0$ (or any other \mathcal{S} that contains this ball) suffices for achieving vanishing overload probability. Thus, for such choice of \mathcal{S} we have that $D \approx \sigma^2(UL)$.

If L is ‘‘sufficiently dense’’ with respect to \mathcal{S} , it holds that $|L \cap \mathcal{S}| \approx \gamma(L) \cdot \text{Vol}(\mathcal{S})$ where $\gamma(L)$ is the point density of L (see Section III-D for definition of $\gamma(L)$ and [36, Lemma 3.3] for precise bounds on $|L \cap \mathcal{S}|$). This approximation becomes accurate in the high-resolution limit (as high-resolution corresponds to large $\gamma(L)$). We can therefore approximate the rate as

$$R = \frac{1}{n} \log |L \cap \mathcal{S}| \approx \frac{1}{n} \log(\text{Vol}(\mathcal{S})) + \frac{1}{n} \log \gamma(L). \quad (35)$$

It follows that in the high-resolution regime the quantization rate R and the normalized log point-density $\frac{1}{n} \log \gamma(L)$ are equal up to a constant. Furthermore, recalling that $D \approx \sigma^2(UL)$ provided that $\Pr(W \notin \mathcal{S})$ is sufficiently small, we can express D in the form familiar

to us, specifically, we get from (35)

$$\begin{aligned} D &\approx \left(\sigma^2(UL) \cdot \gamma^{\frac{2}{n}}(UL) \right) |U|^{\frac{2}{n}} \cdot \text{Vol}^{\frac{2}{n}}(\mathcal{S}) \cdot 2^{-2R} \\ &= \left(\sigma^2(UL) \cdot \gamma^{\frac{2}{n}}(UL) \right) |\Sigma_X|^{\frac{1}{n}} \cdot \text{Vol}^{\frac{2}{n}}(\mathcal{S}) \cdot 2^{-2R}, \end{aligned} \quad (36)$$

where we also used the fact $\gamma(UL) = |U|^{-1}\gamma(L)$. The term in (\cdot) is a famous and extremely well-studied quantity [9] known as the normalized second moment (NSM) of UL , defined as $\sigma^2(UL) \cdot \gamma^{\frac{2}{n}}(UL)$. From (36) we see that the tradeoff between rate and distortion is determined by: 1)The volume of the shaping region (which is required to satisfy $\Pr(W \notin \mathcal{S}) \ll 1$); 2)The NSM of UL .

We stress that $\sigma^2(UL)$ is the WMSE distortion when the $\text{argmin}_{c \in L}$ in (30) is solved exactly. As we discuss below, finding the exact solution to this problem is generally infeasible for large n , and one must resort to approximate solutions. In the case where a sub-optimal quantizer $q : \mathbb{R}^n \rightarrow L$ is used, the NSM is replaced with $\frac{1}{n} \mathbb{E} \|Y - U \cdot q(Y)\|^2 \cdot \gamma^{\frac{2}{n}}(UL)$ in (36).

An alternative approach for shaping is entropy coding (EC). If quantization with variable rate is allowed, and only the *expected* quantization rate is constrained to be at most R , one can quantize the source to the point $\hat{W}^* \in L$ that minimizes (30) (assuming $\mathcal{C} = L$), and then describe the resulting point in bits using EC.

A sequence of classic works [7], [37], [38], [39] have considered the case of high-resolution entropy coded quantization with infinite constellation, and subsequent work restricted attention to the case where the infinite constellation is taken as a lattice [40], [41], [9]. Those works considered the MSE case with $\Sigma_X = I_n$, but their conclusion adapted to the WMSE case is that for high-resolution quantization and large n the tradeoff between distortion and average quantization rate also obeys (36) with $\text{Vol}^{\frac{2}{n}}(\mathcal{S})$ replaced by $2^{2h(W)} = 2\pi e \sigma_W^2$, where $h(\cdot)$ denotes differential entropy. Namely, for entropy coded high-resolution lattice quantization

$$D \approx \left(\sigma^2(UL) \cdot \gamma^{\frac{2}{n}}(UL) \right) 2\pi e \cdot |\Sigma_X|^{\frac{1}{n}} \sigma_W^2 \cdot 2^{-2R}. \quad (37)$$

Since rate and normalized lattice point density are equivalent under either shaped or entropy coded lattice quantization, it suffices to consider the tradeoff between WMSE distortion and $\gamma(L)$.

To that end, we now provide a lower bound on $\sigma^2(UL)$ that depends only on $|U|$ and on $\gamma(L)$. This bound is due to Zador [38] (See also [42, eq. (82)]). It follows from the fact that $\sigma^2(UL)$ is the power of a random vector uniformly distributed over $\mathcal{V}_{UL} \subset \mathbb{R}^n$, which is a convex body of volume $V = |U| \cdot \gamma^{-1}(L)$. Among all bodies in \mathbb{R}^n of volume V , the power of a

uniform random vector is minimized when the body is a ℓ_2 ball.¹¹

Proposition 1 (Zador): For any full-rank lattice $L \subset \mathbb{R}^n$ and any full-rank matrix $U \in \mathbb{R}^{n \times n}$

$$\sigma^2(UL) \geq \frac{\Gamma(\frac{n}{2} + 1)}{(n+2)\pi} \gamma(L)^{-\frac{2}{n}} |U|_{\frac{2}{n}} \approx \frac{\gamma(L)^{-\frac{2}{n}} |U|_{\frac{2}{n}}}{2\pi e}. \quad (38)$$

In (38), $\Gamma(\cdot)$ is the gamma function (not to be confused with $\Gamma(R)$ from (5)). The last approximation can be replaced with a lower bound at the expense of multiplying the right-hand side term by a factor of $\frac{n}{n+2}$. Note that substituting (38) in (37), gives $D \gtrsim D_{\text{High-Rate}}^*(R)$, where $D_{\text{High-Rate}}^*(R)$ is given in (28).

Even without relying on the validity of (37), Proposition 1 provides a simple lower bound on the WMSE distortion of any high-resolution quantization scheme for which the reconstruction satisfies $\hat{W} \in L$, for a lattice $L \subset \mathbb{R}^n$. This follows since for any such quantization scheme we have (due to (30))

$$D \geq \frac{1}{n} \min_{c \in L} \mathbb{E} \|Y - U \cdot c\|^2 = \frac{1}{n} \|Y - Q_{UL}(Y)\|^2,$$

where $Q_{UL} : \mathbb{R}^n \rightarrow UL$ is the nearest neighbor (optimal) quantizer for the lattice $UL \subset \mathbb{R}^n$. Under the high-resolution assumption $e_Y = Y - Q_{UL}(Y) \sim \text{Uniform}(\mathcal{V}_{UL})$ and consequently

$$D \geq \sigma^2(UL) \gtrsim \frac{\gamma(L)^{-\frac{2}{n}} |U|_{\frac{2}{n}}}{2\pi e} = |\Sigma_X|^{\frac{1}{n}} \frac{\gamma(L)^{-\frac{2}{n}}}{2\pi e}, \quad (39)$$

where we have used Proposition 1 and the fact that $|U|^2 = |\Sigma_X|$.

For example, in GPTQ/LLM it is common to take $L = \alpha\mathbb{Z}^n$ as the base lattice, whose density is $\gamma(L) = \alpha^{-n}$. From (39) we see that its attained distortion must satisfy $D \gtrsim \alpha^2 \frac{|\Sigma_X|^{1/n}}{2\pi e}$.

Note that (39) relies on the assumption $e = UW - Q_{UL}(UW) \sim \text{Uniform}(\mathcal{V}_{UL})$. This holds asymptotically in the limit of high resolution ($\gamma(L) \rightarrow \infty$) for all full-rank matrices $U \in \mathbb{R}^{n \times n}$. If one uses dithered lattice quantization [9], this assumption holds exactly for all $\gamma(L)$. However, in the case of dithered lattice quantization the resulting estimate \hat{W}^* would not be in L . Instead, under dithered lattice quantization we have $\hat{W}^* = Y + e$ where $e \sim \text{Uniform}(\mathcal{V}_{UL})$, $e \perp Y$.

¹¹In fact, it is known [43] that for almost all lattices (with respect to the natural measure on the space of lattices) $\sigma^2(L)$ is only $(1 + O(1/n))$ greater than the second moment of the corresponding ℓ_2 ball. Thus, this lower bound is asymptotically attained by a ‘‘typical’’ lattice.

Consequently, one can benefit by setting $\hat{W} = \beta \hat{W}^*$ with $\beta < 1$ as the estimate for W , as in this case

$$\begin{aligned} \frac{1}{n} \mathbb{E} (X^T (W - \hat{W}))^2 &= \mathbb{E} ((1 - \beta) X^T W - \beta e)^2 \\ &= (1 - \beta)^2 \sigma_W^2 \frac{\text{tr}(\Sigma_X)}{n} + \beta^2 \sigma^2(UL). \end{aligned} \quad (40)$$

This is the same shrinkage effect that we discussed in Section I-B and Fig. 1b. Note that $\beta \rightarrow 1$ as $\sigma^2(UL)$ decreases, which is the case for high-resolution quantization.

C. Product Codebooks, Codebook Spacing, and Successive Cancellation

This section discusses efficient quantization schemes via successive interference cancellation (SIC). The discussion is not restricted to lattice quantizers, and we will see that SIC can be applied for any product code. Afterwards, we will specialize the discussion to lattice quantizers.

In Section IV we argued that the relevant setup for LLMs is the uninformed/ Σ_X -oblivious decoder case. We introduced there the problem of quantizing a single column of the weight matrix. In practice, however, a weight matrix will typically have $a \gg 1$ column vectors, and the WMSE matrix Σ_X is common to all of them (as they all operate on the same activations). If the number of column vectors is of the same order as n , it is possible for the decoder to send an $O(n)$ bits description of Σ_X to the uninformed decoder with only a negligible effect on the quantization rate. In particular, the encoder can apply a diagonal matrix $\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^{n \times n}$, whose spacing coefficients $\{\alpha_i\}$ are chosen based on Σ_X , and quantize W using the codebook $\mathcal{A} \cdot \mathcal{C}$ rather than the codebook \mathcal{C} . The cost of reporting $(\alpha_1, \dots, \alpha_n)$ to the decoder, is amortized over the a columns of the weight matrix, which makes it negligible provided that a is not too small. In fact, spacing the codebook \mathcal{C} used for the reconstruction of W to the codebook $\mathcal{A} \cdot \mathcal{C}$ is equivalent to changing X^T to $X^T \mathcal{A}$. Consequently, in LLMs it is sometimes the case that \mathcal{A} can be absorbed in the activations using the layer-norm with no additional cost associated with describing the scales [13].

Our focus is on obtaining efficient approximate solution to the optimization in equation (29), which becomes equivalent to the optimization in (30) when $\Sigma_X = U^T U$ is decomposed using the Cholesky decomposition. Assume the codebook $\mathcal{C} \subset \mathbb{R}^n$ is a *product codebook* of the form

$$\mathcal{C} = \mathcal{C}_1 \times \dots \times \mathcal{C}_n, \quad \text{where } \mathcal{C}_i \subset \mathbb{R} \ \forall i \in [n]. \quad (41)$$

Product codebooks are a common practical choice as it allows for fast (and parallel) decoding. Under the MSE

criterion (corresponding to $\Sigma_X = I_n$) they also result in fast optimal encoding, as in this case $U = I_n$ and (30) becomes

$$\begin{aligned} \hat{W}_{\text{MSE}} &= \underset{c_1 \in \mathcal{C}_1, \dots, c_n \in \mathcal{C}_n}{\operatorname{argmin}} \sum_{i=1}^n (W_i - c_i)^2 \\ \implies \hat{W}_{\text{MSE},i} &= \underset{c_i \in \mathcal{C}_i}{\operatorname{argmin}} (W_i - c_i)^2, \quad \forall i \in [n]. \end{aligned} \quad (42)$$

Under the WMSE criterion with non-diagonal Σ_X , product codebooks do not in general lend themselves to fast optimal encoding algorithms. In particular, if $\mathcal{C} = \mathbb{Z}^n = \mathbb{Z} \times \dots \times \mathbb{Z}$, then exactly solving the optimization in (30) requires solving the closest vector problem (CVP) in the lattice $\tilde{L} = U\mathbb{Z}^n$, as observed in [33], [34]. This problem is known to be NP-hard. Consequently, one must resort to sub-optimal algorithms.

Here, we restrict attention to successive interference cancellation (SIC). The most general form of this algorithm is provided in Algorithm 1 and illustrated in Figure 6. In *generalSIC* the n codebooks $\mathcal{C}_1, \dots, \mathcal{C}_n \subset \mathbb{R}$ whose product is $\mathcal{C} \subset \mathbb{R}^n$ can be arbitrary, and each of them is further scaled by the corresponding α_i . This scaling can in general be absorbed in the codebooks definition. However, since we allow \mathcal{A} to depend on Σ_X through the matrix $U \in \mathbb{R}^{n \times n}$ while the codebooks $\mathcal{C}_1, \dots, \mathcal{C}_n$ are not allowed to depend on Σ_X , we do not absorb \mathcal{A} into the codebooks. In *generalSIC*, we first filter the vector Y using the feedforward filter $F \in \mathbb{R}^{n \times n}$. Any matrix $F \in \mathbb{R}^{n \times n}$ can be used here. Then, the vector \hat{W} is generated sequentially, starting from \hat{W}_n up to \hat{W}_1 . At every step i the scalar $(FY)_i$ is fed to the quantizer to generate \hat{W}_i , and then the vector FY is updated to $FY - \hat{W}_i B_{:,i}$. Due to the strictly lower triangular structure of the feedback matrix $B \in \mathbb{R}^{n \times n}$, we have that $\hat{W} = \mathcal{A}Q\left(\mathcal{A}^{-1}(FY - B\hat{W})\right)$. If the quantizer is modeled as adding independent quantization noise Z , we therefore obtain

$$\hat{W} = (I_n + B)^{-1}FY + (I_n + B)^{-1}AZ. \quad (43)$$

Finally, the estimate \hat{W} is further scaled by $\beta > 0$.

Let us temporarily fix the spacing matrix \mathcal{A} . The choices of F and B and β , should strike a balance between two quantities. First, we want the ℓ_2 norm of

$$\begin{aligned} e_Y &= Y - \beta U \hat{W} \\ &= (I_n - \beta U(I_n + B)^{-1}F)Y - \beta U(I_n + B)^{-1}AZ \end{aligned} \quad (44)$$

to be as small as possible. On the other hand, we also want \hat{W} to have differential entropy as small as possible, as this will dictate the volume of the shaping region required for capturing it (or equivalently, the average rate of encoding the quantizer's output using entropy coding).

However, in the limit of large quantization rate, where the energy of Z vanishes, the optimal

choice of F, B, β tends to the solution for which $(I_n - \beta U(I_n + B)^{-1}F) = 0$. This corresponds to choosing

$$\begin{aligned} \beta_{\text{High-Rate}} &= 1, \quad F_{\text{High-Rate}} = (\operatorname{diag}(U))^{-1}, \\ B_{\text{High-Rate}} &= (\operatorname{diag}(U))^{-1}U - I_n. \end{aligned} \quad (45)$$

If we use the *generalSIC* algorithm with F, B, β from (45), and use the same scalar codebook \mathcal{C}_0 for quantizing all coordinates, the algorithm simplifies to Algorithm 2 which we simply call the SIC algorithm.

As we shall see below, the choice of \mathcal{A} has an important impact on the scheme's performance. In the special case where $\mathcal{A} = \alpha I_n$ for some $\alpha > 0$, the SIC algorithm becomes completely equivalent to the canonical GPTQ algorithm (which in turn, is equivalent to the LDLQ algorithm [3]). This equivalence was recently shown in [33], [34]. Thus, in the sequel, we refer to SIC with $\mathcal{A} = \alpha I_n$ as the GPTQ algorithm. Note however that the GPTQ algorithm was originally presented in [2] from a *noise-shaping* point of view, where the quantization noise is filtered and fed to the next quantizer. In the SIC point of view, it is the signal Y , rather than the quantization noise, that is filtered and fed to the next quantizer.

We find the SIC point of view of [33], [34] more intuitive than the original noise-filtering perspective. Furthermore, under the SIC point of view, the problem is also similar to V-BLAST decoding for the Gaussian MIMO channel [44].

In practice, the sub-codebook \mathcal{C}_0 is often taken as the constellations corresponding to data-type FP8, FP4, INT8, etc., resulting in \hat{W} that can be used in fast MatMul hardware. For the remainder of our discussion we will assume $\mathcal{C}_0 = \mathbb{Z}$, such that the product codebook is a lattice. In this case, the SIC algorithm is merely a low-complexity sub-optimal algorithm for the closest vector problem, which is often also referred to as Babai's nearest plane algorithm [45]. For small n the sphere-decoder [46], [47] can be used, and for large n other sub-optimal algorithms can equally be used instead of SIC.

We now analyze the distortion attained by the SIC algorithm, with $\mathcal{C}_0 = \mathbb{Z}$. The resulting \hat{W} is in the lattice $\mathcal{A}\mathbb{Z}^n$, whose cardinality is unbounded. The description of \hat{W} in bits will be handled via shaping or entropy coding, and for now we only constrain the point density of $L = (\alpha_1\mathbb{Z}) \times \dots \times (\alpha_n\mathbb{Z})$ to $\gamma(L)^{-1} = (\prod_{i=1}^n \alpha_i) = \alpha^n$. Recall that from Proposition 1 and (39) we have that even if the optimal solution to (30) is found, the distortion must satisfy $D \gtrsim \alpha^2 \frac{|\Sigma_X|^{2/n}}{2\pi e}$. Note that for $\mathcal{C}_0 = \mathbb{Z}$, the quantizer Q used in the SIC algorithm takes the simple form $Q(x) = \operatorname{round}(x)$ so that the execution of the SIC algorithm is particularly simple. The next

Algorithm 1 generalSIC

Inputs: $Y \in \mathbb{R}^n$, feed-forward filter $F \in \mathbb{R}^{n \times n}$, strictly upper triangular feedback filter $B \in \mathbb{R}^{n \times n}$, diagonal spacing matrix $\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_n) \in \mathbb{R}_+^{n \times n}$, scaling coefficient $\beta > 0$ and codebooks $\mathcal{C}_1, \dots, \mathcal{C}_n \subset \mathbb{R}$

Outputs: $\hat{W} \in \mathcal{C}_1 \times \dots \times \mathcal{C}_n$.

$Y \leftarrow FY$

for $i = n : 1$ **do**

$\hat{W}_i \leftarrow \alpha_i Q_i \left(\frac{Y_i}{\alpha_i} \right)$

$\triangleright Q_i(x) = \text{argmin}_{c_i \in \mathcal{C}_i} (x - c_i)^2$

$Y \leftarrow Y - \hat{W}_i \cdot B_{:,i}$

$\triangleright B_{:,i}$ is the i th column of B

end for

$\hat{W} \leftarrow \beta \hat{W}$

Algorithm 2 SIC

Inputs: $Y \in \mathbb{R}^n$, upper triangular $U \in \mathbb{R}^{n \times n}$, diagonal matrix $\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_n) \in \mathbb{R}_+^{n \times n}$ and scalar codebook $\mathcal{C}_0 \subset \mathbb{R}$

Outputs: $c_{\text{SIC}} \in \mathcal{C}_0^{\otimes n}$

for $i = n : 1$ **do**

$c_{\text{SIC},i} \leftarrow \alpha_i Q \left(\frac{Y_i}{\alpha_i U_{i,i}} \right)$

$\triangleright Q(x) = \text{argmin}_{c \in \mathcal{C}_0} (x - c)^2$

$Y \leftarrow Y - c_{\text{SIC},i} \cdot U_{:,i}$

$\triangleright U_{:,i}$ is the i th column of U

end for

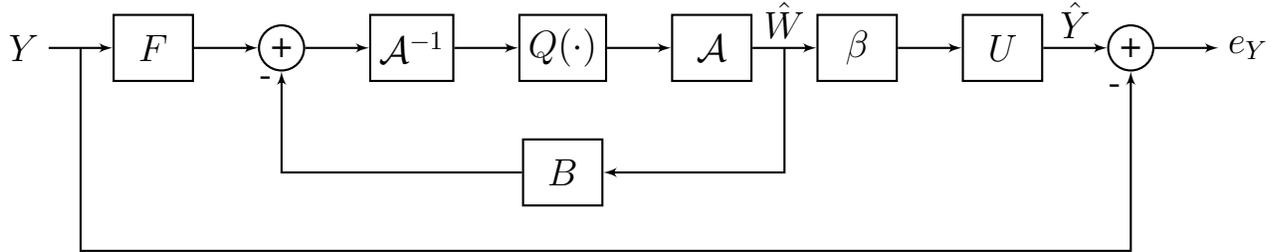


Fig. 6: Illustration of the generalSIC quantization algorithm. The matrix $F \in \mathbb{R}^{n \times n}$ is unrestricted, whereas the matrix $B \in \mathbb{R}^{n \times n}$ must be strictly upper triangular due to causality. The matrix $\mathcal{A} \in \mathbb{R}^{n \times n}$ is positive diagonal, and determines the spacing of the quantizer in each coordinate. The parameter $\beta > 0$ is a scaling parameter. The GPTQ algorithm is obtained as a special case for B, F, β taken as in (45), and $\mathcal{A} = \alpha I_n$. The WaterSIC algorithm is obtained with the same F, B, β , but with $\mathcal{A} = \text{diag}(\alpha_1^{\text{Water}}, \dots, \alpha_n^{\text{Water}})$ and α_i^{Water} are given in (49)

result can be deduced from [45], and we bring the simple proof for completeness. that

Lemma 1: Assume we apply the SIC Algorithm with $y \in \mathbb{R}^n$, upper triangular $U \in \mathbb{R}^{n \times n}$, and $\mathcal{C}_i = \alpha_i \mathbb{Z}$ for all $i \in [n]$. Then

$$e_{\text{SIC}} = y - U \cdot c_{\text{SIC}} \in \mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n},$$

where

$$\mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n} = \prod_{i=1}^n \left[-\frac{|\alpha_i U_{i,i}|}{2}, \frac{|\alpha_i U_{i,i}|}{2} \right).$$

If in addition $e_{\text{SIC}} \sim \text{Uniform}(\mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n})$ we have

$$\begin{aligned} D_{\text{SIC}} &\triangleq \frac{1}{n} \mathbb{E} \|Y - U \cdot c_{\text{SIC}}\|^2 \\ &= \frac{1}{12} \cdot \frac{1}{n} \sum_{i=1}^n (\alpha_i U_{i,i})^2. \end{aligned} \quad (46)$$

Proof. Let $c_{\text{SIC}}(y, U)$ be the result of applying the SIC algorithm with inputs $y \in \mathbb{R}^n$ and $U \in \mathbb{R}^{n \times n}$. Denote $\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^{n \times n}$. The lemma follows from combining the two following observations:

- 1) $\mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n} = \{y \in \mathbb{R}^n : c_{\text{SIC}}(y, U) = 0\}$;
- 2) For any $z \in \mathbb{Z}^n$ it holds that

$$c_{\text{SIC}}(y + U \mathcal{A} z, U) = \mathcal{A} z + c_{\text{SIC}}(y, U).$$

Thus the SIC algorithm induces the partition of space to decision regions

$$\mathcal{D}_z = U\mathcal{A}z + \mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n}, \quad \forall z \in \mathbb{Z}^n \quad (47)$$

and $c_{\text{SIC}}(y, U) = \mathcal{A}z$ iff $y \in \mathcal{D}_z$. Consequently, $e_{\text{SIC}} = y - U \cdot c_{\text{SIC}}(y, U) \in \mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n}$. The claim on D_{SIC} immediately follows from the product structure of $\mathcal{P}_{\{U_{i,i}, \alpha_i\}_{i=1}^n}$. ■

It follows from (46) that the distortion for the GPTQ algorithm followed by entropy coding is as given in (33).

By the arithmetic-mean geometric-mean inequality (AM-GM) and (46) we have

$$\begin{aligned} D_{\text{SIC}} &= \frac{1}{12} \cdot \frac{1}{n} \sum_{i=1}^n (\alpha_i U_{i,i})^2 \\ &\geq \frac{1}{12} \left(\prod_{i=1}^n \alpha_i \right)^{\frac{2}{n}} \left(\prod_{i=1}^n U_{i,i} \right)^{\frac{2}{n}} = \frac{\alpha^2}{12} |\Sigma_X|^{\frac{2}{n}}. \end{aligned} \quad (48)$$

The inequality above is achieved with equality iff the scaling factors are of the form

$$\alpha_i^{\text{Water}} \triangleq \alpha \cdot \frac{|U|_n^{\frac{1}{n}}}{|U_{i,i}|}, \quad \forall i \in [n]. \quad (49)$$

D. WaterSIC algorithm

Motivated by (48) and (49), we propose Algorithm 3: the WaterSIC weight-only quantization algorithm for a weight matrix $W \in \mathbb{R}^{n \times a}$. The properties of the resulting reconstruction \hat{W} are provided in Lemma 2.

Lemma 2: The reconstruction $\hat{W} = \text{diag}(\alpha_1, \dots, \alpha_n) Z_{\text{SIC}}$ produced by the WaterSIC weight-only quantization algorithm satisfies:

- 1) Any column $\hat{W}_{:,j}$ of \hat{W} belongs to the lattice $L = (\alpha_1 \mathbb{Z}) \times \dots \times (\alpha_n \mathbb{Z})$ whose density is $\gamma(L) = \alpha^{-n}$
- 2) $U(W - \hat{W}) \in \alpha |\Sigma_X|^{1/2n} \cdot \left[-\frac{1}{2}, \frac{1}{2}\right]^{n \times a}$
- 3) If we further assume $U(W - \hat{W}) \sim \text{Uniform}\left(\alpha |\Sigma_X|^{1/2n} \cdot \left[-\frac{1}{2}, \frac{1}{2}\right]^{n \times a}\right)$ then

$$D_{\text{WaterSIC}} = \frac{1}{an} \mathbb{E} \|X^\top (W - \hat{W})\|^2 = \frac{\alpha^2 |\Sigma_X|^{1/n}}{12}.$$

- 4) $\hat{W} \in W + \alpha |\Sigma_X|^{1/2n} \cdot U^{-1} \left[-\frac{1}{2}, \frac{1}{2}\right]^{n \times a}$
- 5) $Z_{\text{SIC}} \in \mathcal{A}^{-1} W + \tilde{U} \left[-\frac{1}{2}, \frac{1}{2}\right]^{n \times a}$ where $\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_n)$ and $\tilde{U} = \alpha |\Sigma_X|^{1/2n} \cdot (U\mathcal{A})^{-1}$ is an upper triangular matrix with unit determinant.

The proof is immediate in light of Lemma 1. Using (35), we see that if the resulting output Z_{SIC} is encoded to bits via ball-shaping or entropy coding, the relation between α and quantization rate R is

$$\alpha \approx \sqrt{2\pi e \sigma_W^2 2^{-2R}}. \quad (50)$$

Substituting this into item 3 of Lemma 2, gives (34). Remarkably, the waterfilling choice of scaling factors (49)

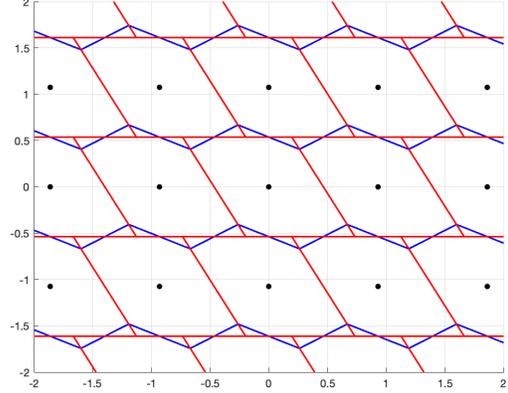


Fig. 7: Illustration of the quantization regions for the optimal lattice quantizer (blue) and for the WaterSIC lattice quantizer (red) for $\Sigma_X = V\Lambda V^T$ where $V = [1 \ 1; 1 \ -1]/\sqrt{2}$ and $\Lambda = \text{diag}(3, 1)$. The lattice used for quantization is $L = (\alpha_1 \mathbb{Z}) \times (\alpha_2 \mathbb{Z})$ where α_1, α_2 are determined via (49), with $\alpha = 1$.

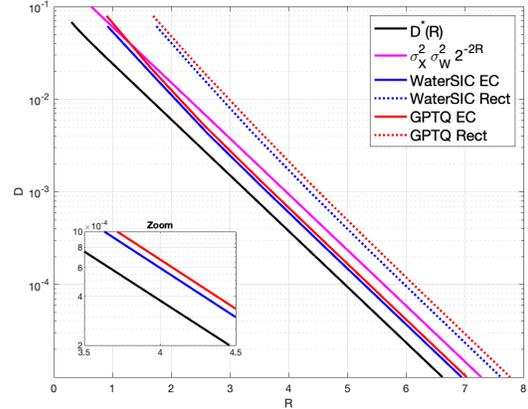


Fig. 8: Performance of several weight-only quantization schemes against benchmark. Here $W \sim \mathcal{N}(0, I_n)$ where $n = 4096$ and Σ_X is the empirical covariance matrix computed from activation samples corresponding to W_v in the 15th layer of Llama3-8B.

result in a lattice whose distortion under the very simple SIC algorithm is only a factor of $\frac{2\pi e}{12} \approx 1.4233$ from the lower bound (39). The loss of this factor is due to the fact that, while the weighted error $U(W - \hat{W})$ lies within a convex set in isotropic position, this set is a cube. The lower bound (39) on the other hand is achieved with equality only if the convex set supporting the weighted error is a ℓ_2 -ball. The decision regions for the WaterSIC algorithm, as well as for the optimal lattice quantizer corresponding to $L = (\alpha_1 \mathbb{Z}) \times (\alpha_2 \mathbb{Z})$ are illustrated in

Algorithm 3 WaterSIC weight-only quantization

Inputs: $W \in \mathbb{R}^{n \times a}$, PSD matrix Σ_X and point density $\alpha > 0$.

Outputs: $Z_{\text{SIC}} \in \mathbb{Z}^{n \times a}$ and $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}_+^n$ such that $\hat{W} = \text{diag}(\alpha_1, \dots, \alpha_n)Z_{\text{SIC}}$.

Compute upper-triangular $U \in \mathbb{R}^{n \times n}$ such that $\Sigma_X = U^\top U$

▷ Using the Cholesky decomposition

$Y \leftarrow UW$

$\alpha_i \leftarrow \alpha \cdot \frac{|U|_{ii}^{\frac{1}{n}}}{|U_{i,i}|}, \quad \forall i \in [n]$

$Z_{\text{SIC}} \leftarrow 0^{n \times a}$

▷ Initialize Z_{SIC} with zeros

for $i = n : 1$ **do**

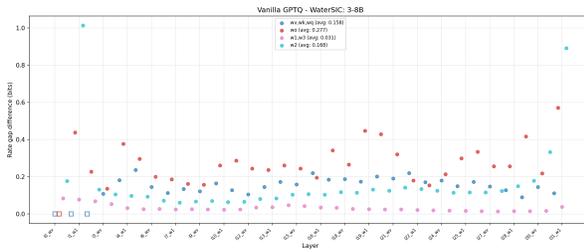
$Z_{\text{SIC},i,:} \leftarrow \text{round}\left(\frac{Y_{i,:}}{\alpha_i U_{i,i}}\right)$

▷ $Y_{i,:}$ and $Z_{\text{SIC},i,:}$ is the i th row of Y and Z_{SIC} , respectively

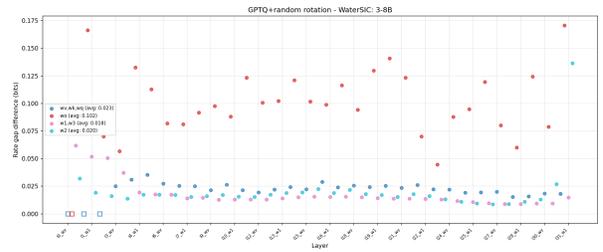
$W \leftarrow Y - \alpha_i U_{:,i} \cdot Z_{\text{SIC},i,:}$

▷ $U_{:,i}$ is the i th column of U

end for



(a) No random rotation.



(b) With random rotation.

Fig. 9: Illustrating rate advantage of WaterSIC over SIC for Σ_X of activations entering various layers of Llama-3-8B when processing Wikitext-2 dataset. Squares correspond to layers with singular Σ_X which we skip. Note that WaterSIC achieves (on random gaussian W) identical performance with or without rotation.

Figure 7.

Shaping: There are several ways to represent Z_{SIC} in bits. One way that is simple, but quite inefficient, is using the shaping region

$$\mathcal{S}_{\text{rect}} = [-q_1, q_1] \times \dots \times [-q_n, q_n] \quad \text{where } q_i = \|Z_{\text{SIC},i,:}\|_\infty, \quad \forall i \in [n]. \quad (51)$$

By definition, all entries of the i th row of Z_{SIC} are in $[-q_i, q_i]$, so overload never occurs. Furthermore, the encoder can describe the shaping region to the decoder by sending $\{q_i\}_{i=1}^n$, with negligible rate if the number of rows is large. And the quantization rate using $\tilde{\mathcal{S}}_{\text{rect}}$ is $R_{\text{rect}} = \frac{1}{n} \sum_{i=1}^n \log(1 + 2q_i)$. Here, we have used the same reconstruction alphabet for an entire row of W . In general, one can decide on different ways to group elements that are described using the same reconstruction alphabet. Such optimizations are called *fine-grained quantization* in the LLM literature. While using the optimal group size can significantly boost performance, here we do not explore this.

A better, but more complicated, way to perform shaping is via partitioning to cosets, much like in Forney’s trellis shaping [48], [40]. In particular, one can use a lattice $L_{\text{Shaping}} \subset \mathbb{Z}^n$, with $\text{covol}(L_{\text{Shaping}}) = 2^{nR}$, such that $|\mathbb{Z}^n / L_{\text{Shaping}}| = 2^{nR}$. In other words, L_{Shaping} partitions \mathbb{Z}^n to 2^{nR} cosets $\{z_i + L_{\text{Shaping}}\}$, where $z_1, \dots, z_{2^{nR}} \in$

\mathbb{Z}^n are coset representatives. In order to encode $Z_{\text{SIC},:,j}$ (the j th column of Z_{SIC}) the encoder describes the coset it belongs to, using nR bits. The decoder outputs the coset member $z \in \mathbb{Z}^n$ for which $\sum_{i=1}^n \alpha_i^2 z_i^2$ is minimal. It is easy to verify that the reconstruction constellation corresponding to this scheme is $\mathbb{Z}^n \cap \tilde{\mathcal{S}}_{\text{Lattice-Shaping}}$ where

$$\tilde{\mathcal{S}}_{\text{Lattice-Shaping}} = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^n \alpha_i^2 y_i^2 \leq \sum_{i=1}^n \alpha_i^2 (y_i - t_i)^2 \quad \forall t \in L_{\text{Shaping}} \right\}. \quad (52)$$

The difficulty here is that the decoder needs to find the coset member that minimizes the scaled energy, which may be computationally intensive unless L_{Shaping} admits special properties. There are many other options one can consider for shaping, which we did not list here.

Finally, the encoder may simply use entropy coding (implemented via any lossless compression package. Note that modern GPU already include dedicated hardware for fast lossless compression/decompression) to describe the entries of Z_{SIC} .

E. Quantizing Llama-3-8B

In Figure 8 we plot the rate-distortion curves attained by WaterSIC and GPTQ without tailored spacing (that

is, with $\alpha_i = \alpha$) under entropy coding or rectangular shaping corresponding to shaping with $\mathcal{S}_{\text{rect}}$ from (51). We see that for high rates the ratio between WaterSIC EC and the fundamental limit $D^*(R)$ from (27) is indeed quite close to $\frac{2\pi e}{12} \approx 1.4233$ (or 0.25 bit in rate). We also see that for this particular Σ_X , at high resolution WaterSIC EC offers only limited gain with respect to GPTQ EC. However, the former is provably close to $D^*(R)$ at high-resolution, whereas the latter is not, and therefore the gap between them may be bigger for other layers, other LLMs, other calibration sets etc.

Let us investigate difference between GPTQ and WaterSIC (entropy coded versions of both) further. Recall that in the high-rate regime, the rate advantage of WaterSIC over GPTQ is estimated as half logarithm of the ratio between arithmetic and geometric mean of squared diagonal entries of U , cf. (33) and (34). We illustrate this rate advantage on Fig. 9 for all layers of Llama-3-8B. We compare both the case of using (Cholesky decomposition of the) original Σ_X as well as $V^T \Sigma_X V$ for a random orthogonal V . Recall that WaterSIC's performance is independent of rotation, and hence identical in both figures, but the improvement over SIC is much smaller with random rotation. Let us consider implications of this finding.

Cholesky factor interpretation. Recall that k -th diagonal entry $U_{k,k}$ can be interpreted as follows. Think of activations (X_1, \dots, X_n) as a stochastic process. Then, each sample X_k can be decomposed as

$$X_k = X_k^{\parallel} + X_k^{\perp},$$

where X_k^{\parallel} is the component that is a linear combination of (X_1, \dots, X_{k-1}) (expressible in terms of $U_{k,j}$, $j = 1, \dots, k-1$) and X_k^{\perp} is the *orthogonal innovation* contained in X_k compared to previous coordinates. $U_{k,k}^2 = \mathbb{E}[(X_k^{\perp})^2]$ is simply the energy of this innovation. Now, applying trace to the identity $\Sigma_X = U^T U$ we get

$$\sum_{k,j} U_{k,j}^2 = \sum_{j=1}^n \lambda_j.$$

Therefore, the average $\frac{1}{n} \sum_k U_{k,k}^2$, that governs SIC's performance (33), equals $\frac{1}{n} \sum_j \lambda_j$ minus the sum of squares of off-diagonal entries of U . This implies, that GPTQ's performance is *the worst* in the PCA basis (where U becomes diagonal), which is counter-intuitive. We also see that the gap between D_{iso} (Σ_X -oblivious encoder and decoder) and the waterfilling D^* , which operates in the PCA-basis, is much higher than for GPTQ vs WaterSIC (compare gaps on Fig. 5 vs Fig. 9).

Privileged basis. On the other hand, as we see from comparing two figures on on Fig. 9 applying random rotation clearly improves performance of GPTQ, thus decreasing $\sum_k U_{k,k}^2$. This implies that the original basis is

“closer” to the diagonalizing PCA basis than a randomly rotated one. This suggests that the standard basis for X is somehow privileged among all other possible ones, an effect whose origins are most likely due to operation of Adam [49], which introduces axis aligned outliers in activations.

Cholesky in random basis. Can we estimate $\sum_k U_{k,k}^2$ from knowing the spectrum $\{\lambda_j\}$ of Σ_X ? The answer is affirmative if one considers applying random rotation. One such estimate was offered (for a restricted class of Σ_X) by [3, Lemma 2]. However, one can get a more precise information about this sum.

Indeed, when we take V to be a random orthogonal matrix and set $\Sigma_X = V^T \text{diag}\{\lambda_j\}V$, quantities $\{U_{k,k}\}$ become functions of a random matrix V . One can show that¹²

$$U_{k,k}^2 = \frac{|\Sigma_X^{(k)}|}{|\Sigma_X^{(k-1)}|},$$

where $|\Sigma_X^{(k)}|$ is a determinant of a $k \times k$ principal minor of Σ_X . In dimensions of interest, these random quantities concentrate quickly around their expectations, which can be computed as¹³

$$\mathbb{E}[|\Sigma_X^{(k)}|] = \frac{1}{\binom{n}{k}} \sum_{|S|=k} \prod_{i \in S} \lambda_i.$$

Thus, we can approximate:

$$U_{k,k}^2 \approx \frac{k}{n-k+1} \frac{\sum_{|S|=k} \prod_{i \in S} \lambda_i}{\sum_{|S|=k-1} \prod_{i \in S} \lambda_i}. \quad (53)$$

In particular, $U_{1,1}^2 \approx \frac{1}{n} \sum_{j=1}^n \lambda_j$ and $U_{n,n}^2 \approx \frac{n}{\sum_{j=1}^n \lambda_j^{-1}}$ and the intermediate values smoothly decrease from the arithmetic mean to harmonic mean. This fact is numerically illustrated on Fig. 10. Consequently, performance of GPTQ with random rotation can be accurately estimated from combining (33) and (53). It is an interesting open problem to estimate worst possible gap (over possible spectra $\lambda_j \geq \epsilon$) between GPTQ with rotation and WaterSIC (which in turn is 0.25-bit away from information-theoretically optimal waterfilling).

WaterSIC vs “universal codebook”. We note that theoretical results in Section IV-C demonstrated that solving (29) optimally essentially attains full waterfilling solution even in low rates, and furthermore does so with a fully Σ_X -oblivious codebook. Note that WaterSIC chooses per-coordinate scales $\alpha_i \propto U_{i,i}^{-1}$ (Cholesky decomposition of) Σ_X and, thus, requires large $a \gg 1$

¹²Simply notice that $\Sigma_X^{(k)} = U^{(k)T} U^{(k)}$.

¹³To see this, let \tilde{V} be a $n \times k$ matrix consisting of first k columns of V , then from Cauchy-Binet $|\Sigma_X^{(k)}| = |(\Lambda^{\frac{1}{2}} \tilde{V})^T (\Lambda^{\frac{1}{2}} \tilde{V})| = \sum_{|S|=k} \prod_{i \in S} \lambda_i |\tilde{V}_{S \times [k]}|^2$. Taking expectation over V we have from symmetry $\mathbb{E}[|\tilde{V}_{S \times [k]}|^2] = c$ for all subsets $S \subset [n]$. When all $\lambda_i = 1$ we must have $|\Sigma_X^{(k)}| = 1$ and hence $c = \binom{n}{k}^{-1}$.

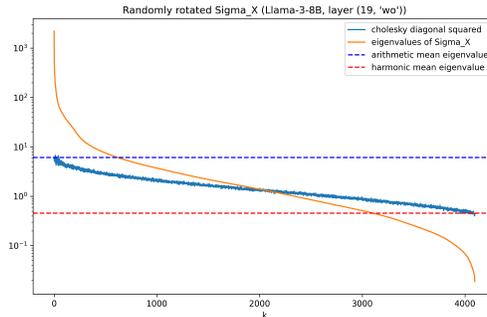


Fig. 10: Illustrating Cholesky diagonals $U_{k,k}^2$ for a randomly rotated $V^\top \Sigma_X V$ and accuracy of approximation (53) in terms of spectrum of Σ_X .

to amortize sending of α_i 's in high precision (usually BF16).

F. Future of weight-only quantization

The high-resolution approximations (36) and (37) in Subsection V-B, combined with the analysis of WaterSIC in Subsection V-D, show that when WaterSIC followed by EC or close-to-optimal spherical shaping is used, we attain the optimal information theoretic distortion from (28) up to a multiplicative gap of $\frac{2\pi e}{12}$. Figure 8 numerically confirms the accuracy of the high-resolution approximations. Furthermore, Figure 9 shows that with random rotation even the vanilla GPTQ algorithm, which is completely equivalent to the widely used GPTQ/LDLQ algorithms, achieves distortion only slightly greater than WaterSIC, and is therefore nearly optimal. In light of this, one may wonder whether the practical schemes are already so good that there is no further room for improvement. We claim that this is not the case, and list several important directions for improved quantization schemes:

First, in this survey paper we restricted attention to the high-rate regime. At small quantization rates (say 0.5–2 bits per entry) many other effects will become important: necessity of Johnson-Lindenstrauss dimensionality reduction, as shown in [1], shrinkage and other low-rate effects. Proper shaping and exploiting structure of W matrices will be paramount.

Second, even our analysis for the high-resolution regime relied on applying EC or near-optimal spherical shaping after quantization. While EC over \mathbb{Z} can be efficiently implemented, even on modern GPUs EC decompression is about an order of magnitude slower than loading the uncompressed entries. Thus, the usefulness of EC for running quantized models on a GPU is somewhat questionable. On the other hand, simple shaping methods that can be easily made efficient, e.g. rectangular shaping, may be highly suboptimal (see Figure 8).

The problem of shaping schemes that are amenable to fast GPU implementation is at its infancy [4].

Third, while the $\frac{2\pi e}{12}$ high-resolution gap-to-optimality of WaterSIC+EC is relatively small, reducing it further is of interest. This gap stems from the sub-optimality of the integer lattice \mathbb{Z}^n as a quantizer. Generally, we can decrease this gap by combining WaterSIC with jointly quantizing d entries from the same row of W to a lattice $L' \subset \mathbb{R}^d$, followed by EC. If we use the optimal lattice quantizer in \mathbb{R}^d we can reduce the $\frac{2\pi e}{12}$ gap to optimality to $2\pi e \cdot G_d$ where G_d is the optimal normalized second moment in dimension d . See [50, Table I] for a list of the best known NSM for $1 \leq d \leq 48$ as of 2023 (some improvements were reported in certain dimensions since then). The downside of replacing \mathbb{Z} with a higher-dimensional lattice is that EC of the quantizer's output becomes more challenging as its cardinality is much larger.

Fourth, implementation of WaterSIC relies on the fact that sending per-channel scales α_i can be amortized freely, which is only the case when one has many output neurons in a linear layer (large $a \gg 1$).

ACKNOWLEDGMENT

Results in Section IV-C are reproduced from a work in progress [32] with Alina Harbouzova (MIT EECS). We thank Egor Lifar and Semyon Savkin (MIT EECS) for numerous discussions and help with obtaining calibration matrices for Llama-3-8B.

REFERENCES

- [1] O. Ordentlich and Y. Polyanskiy, "Optimal quantization for matrix multiplication," *arXiv preprint arXiv:2410.13780*, 2024.
- [2] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "OPTQ: Accurate quantization for generative pre-trained transformers," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=tcbBPnfwxS>

- [3] J. Chee, Y. Cai, V. Kuleshov, and C. M. De Sa, “Quip: 2-bit quantization of large language models with guarantees,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 4396–4429, 2023.
- [4] S. Savkin, E. Porat, O. Ordentlich, and Y. Polyanskiy, “NestQuant: Nested lattice quantization for matrix products and LLMs,” *arXiv preprint arXiv:2502.09720*, 2025.
- [5] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, “Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 318–30 332, 2022.
- [6] B. Hassibi, D. G. Stork, and G. J. Wolff, “Optimal brain surgeon and general network pruning,” in *IEEE international conference on neural networks*. IEEE, 1993, pp. 293–299.
- [7] W. R. Bennett, “Spectra of quantized signals,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.
- [8] R. M. Gray, “Quantization noise spectra,” *IEEE Transactions on information theory*, vol. 36, no. 6, pp. 1220–1244, 2002.
- [9] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, 2014.
- [10] N. Alon, Y. Matias, and M. Szegedy, “The space complexity of approximating the frequency moments,” in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 20–29.
- [11] A. Zandieh, M. Daliri, and I. Han, “Qjl: 1-bit quantized jl transform for kv cache quantization with zero overhead,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25 805–25 813.
- [12] A. Zandieh, M. Daliri, M. Hadian, and V. Mirrokni, “Turboquant: Online vector quantization with near-optimal distortion rate,” *arXiv preprint arXiv:2504.19874*, 2025.
- [13] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, “Smoothquant: Accurate and efficient post-training quantization for large language models,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 38 087–38 099.
- [14] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “GPTQ: Code for accurate post-training quantization of generative pretrained transformers,” <https://github.com/IST-DASLab/gptq>, 2022.
- [15] M. Feder and A. Ingber, “Method, device and system of reduced peak-to-average-ratio communication,” Feb. 14 2012, uS Patent 8,116,695.
- [16] S. Vargaftik, R. B. Basat, A. Portnoy, G. Mendelson, Y. B. Itzhak, and M. Mitzenmacher, “Eden: Communication-efficient and robust distributed mean estimation for federated learning,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 21 984–22 014.
- [17] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman, “Quarot: Outlier-free 4-bit inference in rotated llms,” *arXiv preprint arXiv:2404.00456*, 2024.
- [18] Z. Liu, C. Zhao, I. Fedorov, B. Soran, D. Choudhary, R. Krishnamoorthi, V. Chandra, Y. Tian, and T. Blankevoort, “Spinquant: Llm quantization with learned rotations,” *arXiv preprint arXiv:2405.16406*, 2024.
- [19] B. Chmiel, M. Fishman, R. Banner, and D. Soudry, “Fp4 all the way: Fully quantized training of llms,” *arXiv preprint arXiv:2505.19115*, 2025.
- [20] O. Regev and N. Stephens-Davidowitz, “A reverse minkowski theorem,” *Annals of Mathematics*, vol. 199, no. 1, 2024.
- [21] B. Klartag and J. Lehec, “Affirmative resolution of bourgain’s slicing problem using guan’s bound,” *Geometric and Functional Analysis*, pp. 1–22, 2025.
- [22] S. Savkin, E. Porat, O. Ordentlich, and Y. Polyanskiy, “Nestquant: Nested lattice quantization for matrix products and llms,” *Proc. International Conference on Machine Learning (ICML)*, 2025.
- [23] J. Conway and N. Sloane, “Fast quantizing and decoding and algorithms for lattice quantizers and codes,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 227–232, 1982.
- [24] I. Kaplan and O. Ordentlich, “High-rate nested-lattice quantized matrix multiplication with small lookup tables,” in *Proc. ISIT 2025*, Ann Arbor, Michigan, June 2025.
- [25] J. Chen, V. Egiazarian, T. Hoefler, and D. Alistarh, “Wush: Near-optimal adaptive transforms for llm quantization,” *arXiv preprint arXiv:2512.00956*, 2025.
- [26] Y. Shao, P. Wang, Y. Chen, C. Xu, Z. Wei, and J. Cheng, “Block rotation is all you need for mxfp4 quantization,” *arXiv preprint arXiv:2511.04214*, 2025.
- [27] A. Tseng, Z. Sun, and C. De Sa, “Model-preserving adaptive rounding,” *arXiv preprint arXiv:2505.22988*, 2025.
- [28] H. Badri and A. Shaji, “Half-quadratic quantization of large machine learning models,” November 2023. [Online]. Available: https://mobiusml.github.io/hqq_blog/
- [29] S. Zhang, H. Zhang, I. Colbert, and R. Saab, “Qronos: Correcting the past by shaping the future... in post-training quantization,” *arXiv preprint arXiv:2505.11695*, 2025.
- [30] H. Zhang, S. Zhang, I. Colbert, and R. Saab, “Provable post-training quantization: Theoretical analysis of optq and qronos,” *arXiv preprint arXiv:2508.04853*, 2025.
- [31] Y. Polyanskiy and Y. Wu, *Information theory: From coding to learning*. Cambridge university press, 2024.
- [32] A. Harbuzova, O. Ordentlich, and Y. Polyanskiy, “(work in progress),” *arXiv*, 2026.
- [33] J. Chen, Y. Shabanzadeh, E. Crnčević, T. Hoefler, and D. Alistarh, “The geometry of llm quantization: Gptq as babai’s nearest plane algorithm,” *arXiv preprint arXiv:2507.18553*, 2025.
- [34] J. Birnack, “The lattice geometry of neural network quantization—a short equivalence proof of gptq and babai’s algorithm,” *arXiv preprint arXiv:2508.01077*, 2025.
- [35] A. Tseng, J. Chee, Q. Sun, V. Kuleshov, and C. De Sa, “Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks,” *arXiv preprint arXiv:2402.04396*, 2024.
- [36] O. Ordentlich, O. Regev, and B. Weiss, “Bounds on the density of smooth lattice coverings,” *arXiv preprint arXiv:2311.04644*, 2023.
- [37] P. Panter and W. Dite, “Quantization distortion in pulse-count modulation with nonuniform spacing of levels,” *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.
- [38] P. Zador, “Asymptotic quantization error of continuous signals and the quantization dimension,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 139–149, 1982.
- [39] A. Gersho, “Asymptotically optimal block quantization,” *IEEE Transactions on information theory*, vol. 25, no. 4, pp. 373–380, 1979.
- [40] M. V. Eyuboglu and G. D. Forney, “Lattice and trellis quantization with lattice-and trellis-bounded codebooks-high-rate theory for memoryless sources,” *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 46–59, 1993.
- [41] R. Zamir and M. Feder, “On lattice quantization noise,” *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1152–1159, 1996.
- [42] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., ser. Grundlehren der mathematischen Wissenschaften. New York: Springer-Verlag, 1999, vol. 290.
- [43] O. Ordentlich, “The Voronoi spherical cdf for lattices and linear codes: New bounds for quantization and coding,” *arXiv preprint arXiv:2506.19791*, 2025.
- [44] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, “V-blast: An architecture for realizing very high data rates over the rich-scattering wireless channel,” in *1998 URSI international symposium on signals, systems, and electronics. Conference proceedings (Cat. No. 98EX167)*. IEEE, 1998, pp. 295–300.
- [45] L. Babai, “On lovász’ lattice reduction and the nearest lattice point problem,” *Combinatorica*, vol. 6, no. 1, pp. 1–13, 1986.
- [46] U. Fincke and M. Pohst, “Improved methods for calculating vectors of short length in a lattice, including a complexity analysis,” *Mathematics of computation*, vol. 44, no. 170, pp. 463–471, 1985.

- [47] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE transactions on information theory*, vol. 48, no. 8, pp. 2201–2214, 2002.
- [48] G. D. Forney, "Trellis shaping," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 281–300, 1992.
- [49] N. Elhage, R. Lasenby, and C. Olah, "Privileged bases in the transformer residual stream," *Transformer Circuits Thread*, 2023. [Online]. Available: <https://transformer-circuits.pub/2023/privileged-basis/index.html>
- [50] E. Agrell and B. Allen, "On the best lattice quantizers," *IEEE Transactions on Information Theory*, vol. 69, no. 12, pp. 7650–7658, 2023.