Optimal Quantization for Matrix Multiplication

Or Ordentlich and Yury Polyanskiy

Abstract

Recent work in machine learning community proposed multiple methods for performing lossy compression (quantization) of large matrices. This quantization is important for accelerating matrix multiplication (main component of large language models), which is often bottlenecked by the speed of loading these matrices from memory. Unlike classical vector quantization and rate-distortion theory, the goal of these new compression algorithms is to be able to approximate not the matrices themselves, but their matrix product. Specifically, given a pair of real matrices A, B an encoder (compressor) is applied to each of them independently producing descriptions with R bits per entry. These representations subsequently are used by the decoder to estimate matrix product $A^\top B$. In this work, we provide a non-asymptotic lower bound on the mean squared error of this approximation (as a function of rate R) for the case of matrices A, B with iid Gaussian entries. Algorithmically, we construct a universal quantizer based on nested lattices with an explicit guarantee of approximation error for any (non-random) pair of matrices A, B in terms of only Frobenius norms $\|\bar{A}\|_F, \|\bar{B}\|_F$ and $\|\bar{A}^\top \bar{B}\|_F$, where $\bar{A}, \bar{B}$ are versions of A, B with zero-centered columns, respectively. For iid Gaussian matrices our quantizer achieves the lower bound and is, thus, asymptotically optimal. A practical low-complexity version of our quantizer achieves performance quite close to optimal. In addition, we derive rate-distortion function for matrix multiplication of iid Gaussian matrices, which exhibits an interesting phase-transition at $R \approx 0.906$ bit/entry, showing necessity of Johnson-Lindestrauss dimensionality reduction (sketching) in the low-rate regime.

CONTENTS

I Introduction and main results 2
- I-A Importance of quantization for modern applications . . . 5
- I-B Sketch of the proof . . . 6
- I-C Naive constructions: vector quantization, ϵ-nets and sketching . . . 8
- I-D Related work . . . 9
- I-E Paper organization . . . 11
- I-F Notation . . . 12

II Compression for Inner Product Computation: General Problem Setup and Simple Bounds 12
- II-A Optimal Decoder and Error Expressions . . . 13
- II-B Simple Lower Bounds . . . 14

III Compression for Inner Product Computation: The Symmetric Case 14
- III-A Upper Bound . . . 15
- III-B Lower Bound . . . 16
- III-C The Symmetric Gaussian case . . . 17

IV Compression for Matrix Multiplication 17
- IV-A Setup . . . 17
- IV-B Basic Properties and Bounds . . . 18
- IV-C Maximum Entropy Matrices . . . 19
- IV-D Fundamental Limits . . . 20
- IV-E The Symmetric Gaussian case . . . 21

V Lattice Quantization Scheme for Matrix Multiplication of Arbitrary Matrices 21

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. 1641/21. 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.

VI Nested Lattice Quantization for Inner Product Computation 25
VI-A Lattices . . . 25
VI-B Proof of Theorem 11 . . . 26
VI-B1 Dithered Nested Lattice Quantization for Inner Product . . . 26
VI-B2 Analysis . . . 27

VII Practical Implementation of Nested Lattice Quantizers 29

VIII Open problems 33

Appendix A: Convex envelope of $\Gamma_1(R)$ 35

Appendix B: Good Nested Lattices 35

Appendix C: Bounding the Effect of Overload Events 41

Appendix D: Projections of Random Uniform Orthogonal Vectors 41

Appendix E: Tail Probability of $U_{[d]}$ 43

References 44

I. Introduction and main results

Matrix multiplication is a key component of many numerical algorithms, and is often the dominant factor in the runtime of a program. With the surge of deep neural nets (DNNs) and large language models (LLMs), finding more efficient ways to perform matrix multiplication have become one of the most pressing challenges. Classical work in this field focused on minimizing the number of required operations [1], [2], [3], [4]. Specifics of contemporary problems, however, require rethinking this classical approach to matrix multiplication. First, in machine learning applications requirements for precision of computing matrix products are quite lax. Second, modern computational hardware is often bottlenecked by the memory bandwidth. A natural solution explored by many researchers is to apply lossy compression to matrices leading to deterioration in precision but improvement in the amount of data transferred between memory and computation cores.

We formalize this problem as follows. Consider a pair of matrices $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ which need to be described using R bits per entry (using separate compressors), such that a decoder that obtains bit descriptions of both matrices can estimate $\widehat{A^\top B}$. The metric for gauging quality of approximation that we will use is the squared error between ab entries of $\widehat{A^\top B}$ and $A^\top B$. Note that unlike classical vector quantization, we are requiring compression algorithms to be tailored to the special task of matrix multiplication. As a practical motivation, in Section I-A below we argue that reducing R down to a few bits/entry is necessary for LLMs to fully leverage modern matrix multiplication hardware.

Our main result shows existence of universal quantizers (based on lattices) which compress A and B to R bits/entry and come with explicit precision guarantees. Furthermore, we also show that these guarantees cannot be generally improved by proving a matching lower bound for the case of matrices A and B with iid Gaussian entries. We emphasize, though, that quantizers *are* universal and do not require Gaussian matrices. Indeed, following theoretical ideas laid out in this paper a practical implementation of quantizers [5] established state-of-the-art performance for quantizing weights and activations of various LLMs ranging from 3B to 70B parameters.

To introduce our main results, let us define the function

$$\Gamma(R) = \begin{cases} 1 - \left(1 - (2\cdot 2^{-2R^*} - 2^{-4R^*})\right)\frac{R}{R^*} & R < R^* \\ 2\cdot 2^{-2R} - 2^{-4R} & R \geq R^* \end{cases}. \tag{1}$$

where $R^* \approx 0.906$ is the solution to the fixed-point equation

$$R = \frac{1}{2}\log_2(1 + 4R\ln 2) \tag{2}$$

It will turn out that $\Gamma(R)$ is distortion-rate function for the matrix multiplication of iid Gaussian matrices.

We say that a matrix $A \in \mathbb{R}^{n\times m}$ has "M-bounded entries" if $|a_{i,j}| \in \{0\} \cup [M^{-1}, M]$ for all $i \in [n], j \in [m]$. Our results require the matrices A and B to have M-bounded entries, with $M = e^{o(n)}$. To be more concrete, throughout this paper we take $M = n^{10}2^{2000}$. In particular, this choice of M guarantees that matrices represented in FP64 format have bounded entries. Furthermore, Gaussian iid matrices with finite variance and large n have bounded entries with high probability. This extremely mild condition guarantees that we can describe the ℓ_2 norm of each column of A, B with small multiplicative error using $o(n)$ bits (see Section V). Let $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ be the all-ones vector. For a column vector $x \in \mathbb{R}^n$ we denote by $\bar{x} = x - \left(\frac{1}{n}\mathbf{1}^\top x\right)\mathbf{1}$ its zero-centered version. For a matrix $A = [a_1|\cdots|a_a] \in \mathbb{R}^{n\times a}$ we denote $\bar{A} = [\bar{a}_1|\ldots|\bar{a}_a]$. Our first result is the following.

Theorem 1: For any $\varepsilon > 0$ and sufficiently large n, there exist randomized encoders $f_1 : \mathbb{R}^{n\times a} \to [2^{naR}]$, $f_2 : \mathbb{R}^{n\times b} \to [2^{nbR}]$, and decoders $g : [2^{naR}] \times [2^{nbR}] \to \mathbb{R}^{a\times b}$ and $g_{1-\text{sided}} : [2^{naR}] \times \mathbb{R}^{n\times b} \to \mathbb{R}^{a\times b}$ such that for any $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ with bounded entries we have

1) Let $C = A^\top B$, $\tilde{C} = \bar{A}^\top \bar{B}$, and $\hat{C} = g(f_1(A), f_2(B))$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \le \tilde{C}_{i,j}^2 \cdot \left(\Gamma^2(R) + \varepsilon\right) + \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(\Gamma(R) - \Gamma^2(R) + \varepsilon\right) + n^{-8}, \tag{3}$$

and, in particular,

$$\mathbb{E}\|A^\top B - g(f_1(A), f_2(B))\|_F^2 < \|\bar{A}^\top \bar{B}\|_F^2 \cdot \left(\Gamma^2(R) + \varepsilon\right) + \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(\Gamma(R) - \Gamma^2(R) + \varepsilon\right) + a \cdot b \cdot n^{-8}. \tag{4}$$

2) Let $C = A^\top B$, $\tilde{C} = \bar{A}^\top \bar{B}$, and $\hat{C} = g_{1-\text{sided}}(f_1(A), B)$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \le \tilde{C}_{i,j}^2 \cdot \left(2^{-4R} + \varepsilon\right) + \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(2^{-2R} - 2^{-4R} + \varepsilon\right) + n^{-8}. \tag{5}$$

and, in particular,

$$\mathbb{E}\|A^\top B - g_{1-\text{sided}}(f_1(A), B)\|_F^2 < \|\bar{A}^\top \bar{B}\|_F^2 \cdot \left(2^{-4R} + \varepsilon\right) + \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(2^{-2R} - 2^{-4R} + \varepsilon\right) + a \cdot b \cdot n^{-8}. \tag{6}$$

Note that two parts simply describe the cases, where both or only one matrix needs to be compressed.[1] Our scheme operates by compressing each column of A and B using the same (randomized) nested-lattice quantizer $f_{\text{col}} : \mathbb{R}^n \to [2^{nR}]$, which is applied repeatedly to every column, whereas the decoder g simply estimates each column to get matrices $\hat{A}$ and $\hat{B}$ and computes their scaled matrix product; see Figs. 1 and 2. The parameter κ shown in Figures is used by the encoders for time-sharing/sparsification and is set to $\kappa = \min\{R/R^*, 1\}$ in the Theorem. In particular, for $R < R^*$ a fraction $1 - (\frac{R}{R^*})$ of coordinates are ignored (mapped to 0), corresponding to $\kappa = R/R^*$. As we will see shortly this dimensionality reduction (à la Johnson-Lindenstrauss) turns out to be necessary to achieve asymptotically optimal distortion.

To get a feel for Theorem 1 let us consider independent matrices A and B drawn iid Gaussian $\mathcal{N}(0, \sigma^2)$. For large n, such matrices have bounded entries and are also arbitrarily close to their centered version, with high probability. For the second part, where only A needs to be compressed, note that if B is the $n \times n$ identity matrix, the right hand sides of (5) and (6) read $\sigma^2(2^{-2R} + 2\varepsilon)$ and $na\sigma^2(2^{-2R} + 2\varepsilon)$, respectively, which are optimal, as they correspond to the Gaussian rate-distortion function. For the first part of the Theorem, we have that $\mathbb{E}\|A^\top B\|_F^2 = \frac{\mathbb{E}\|A\|_F^2\|B\|_F^2}{n} = \sigma^4 \cdot nab$ in this case and Theorem 1 shows estimate

$$\mathbb{E}[\|A^\top B - \widehat{A^\top B}\|_F^2] \le \sigma^4 nab(\Gamma(R) + \epsilon)\,.$$

It turns out that this is the best possible approximation (at this compression rate), as shown in our next result.

Theorem 2: Let $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ be independent random matrices, with iid $\mathcal{N}(0, \sigma^2)$ entries.

1) For any $n \ge 1$, and any pair of rate-R encoders $f_1 : \mathbb{R}^{n\times a} \to [2^{naR}]$, $f_2 : \mathbb{R}^{n\times b} \to [2^{nbR}]$ and decoder $g : [2^{naR}] \times [2^{nbR}] \to \mathbb{R}^{a\times b}$, we have

$$\mathbb{E}\|A^\top B - g(f_1(A), f_2(B))\|_F^2 \ge \sigma^4 \cdot nab \cdot \Gamma(R). \tag{7}$$

[1] corresponding to the case of "weights and attention" quantization and "weights-only" quantization in LLMs.

2) For any $n \geq 1$, and any rate-R encoder $f : \mathbb{R}^{n\times a} \to [2^{naR}]$ and decoder $g : [2^{naR}] \times \mathbb{R}^{n\times b} \to \mathbb{R}^{a\times b}$, we have

$$\mathbb{E}\|A^\top B - g(f(A), B)\|_F^2 \geq \sigma^4 \cdot nab \cdot 2^{-2R}. \tag{8}$$

In other words, the encoders f_1, f_2, g from Theorem 1 attain the lower bound from Theorem 2, and are therefore asymptotically optimal for this class of matrices.

We also show a simpler to use bound, based on our compression scheme applied with no "MMSE scaling" and no time-sharing - that is, with $\alpha = \kappa = 1$ in Figures 1, 2. The resulting bound does not meet the lower bound of Theorem 2 for Gaussian iid matrices. However, for moderate R it is never much worse than the bound from Theorem 1. For some matrices A, B it is significantly better than the bound from Theorem 1.

Theorem 3: For any $\varepsilon > 0$ and sufficiently large n, there exist randomized encoders $f_1 : \mathbb{R}^{n\times a} \to [2^{naR}]$, $f_2 : \mathbb{R}^{n\times b} \to [2^{nbR}]$, and decoders $g : [2^{naR}] \times [2^{nbR}] \to \mathbb{R}^{a\times b}$ and $g_{1-\text{sided}} : [2^{naR}] \times \mathbb{R}^{n\times b} \to \mathbb{R}^{a\times b}$ such that for any $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ with bounded entries we have

1) Let $C = A^\top B$, and $\hat{C} = g(f_1(A), f_2(B))$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \leq \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(\frac{2\cdot 2^{2R}-1}{(2^{2R}-1)^2} + \varepsilon\right) + n^{-8}, \tag{9}$$

and in particular

$$\mathbb{E}\|A^\top B - g(f_1(A), f_2(B))\|_F^2 < \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(\frac{2\cdot 2^{2R}-1}{(2^{2R}-1)^2} + \varepsilon\right) + a\cdot b\cdot n^{-8}. \tag{10}$$

2) Let $C = A^\top B$, and $\hat{C} = g_{1-\text{sided}}(f_1(A), B)$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \leq \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(\frac{1}{2^{2R}-1} + \varepsilon\right) + n^{-8}. \tag{11}$$

and in particular

$$\mathbb{E}\|A^\top B - g_{1-\text{sided}}(f_1(A), B)\|_F^2 < \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(\frac{1}{2^{2R}-1} + \varepsilon\right) + a\cdot b\cdot n^{-8}. \tag{12}$$

Note that the term $\|\bar{A}^\top\bar{B}\|_F^2$ does not appear at all in Theorem 3, and whenever $\|\bar{A}^\top\bar{B}\|^2 \gg \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}$ the error in Theorem 3 is significantly smaller than the error in Theorem 1.

To put Theorem 3 in context, we note that recent work in LLMs suggested to use random rotation of A and B and then quantize each column of the rotated matrices using sub-optimal lattice quantizers [6], [7]. uch rotations, first practically employed by [8], make each column vector Gaussian-like, see more in Sections I-B and I-D. A popular choice is to use the scalar quantizer, which is equivalent to quantizing to the lattice $\mathbb{Z}^n$ with a cubic shaping region. In practice, to apply the scalar quantizer on a vector $a_i \in \mathbb{R}^n$, the common approach in the DNN and LLM literature is to store $\|a_i\|_\infty$, then normalize to $\tilde{a}_i = a_i/\|a_i\|_\infty$ such that all entries of $\tilde{a}_i$ are in $[-1, 1]$, then use a R-bit scalar quantizer with dynamic range $[-1, 1]$, and finally rescale the result by $\|a_i\|_\infty$. See, e.g. [7]. If the vector a_i is uniform on $\sqrt{n}\mathbb{S}^{n-1}$ (the sphere with radius $\sqrt{n}$), then for large n we have that $\|a_i\|_\infty$ concentrates around $\sqrt{2\ln n}$. It follows that the expected squared quantization error this quantizer attains per entry is $\left(\frac{2}{3}\ln n\right)2^{-2R}$. Using this quantizer for matrix multiplication (after rotating each matrix by the same random rotation) therefore results in

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \leq \frac{\|a_i\|^2\|b_j\|^2}{n}\left(\frac{2}{3}\ln n\right)\left(\frac{2\cdot 2^{2R} + \frac{2}{3}\ln n}{(2^{2R})^2}\right), \quad \forall i \in [a], j \in [b]. \tag{13}$$

Thus, replacing the scalar quantizer $\mathbb{Z}^n$ with a high-dimensional pair of "good" nested lattices, as we do in the proof of Theorem 3 saves a factor of $\frac{2}{3}\ln n$ in the expected squared error for moderate R.

The scheme used for proving Theorem 3 is based on using high-dimensional nested lattices with some asymptotically optimal properties. Unfortunately, such lattices do not lend themselves to efficient implementation. Another key contribution of this paper, described in Section VII, is a simplified nested-lattice quantization scheme, based on Conway and Sloane's Voronoi codes [9], that is similar to the one used in the proofs of Theorem 1 and Theorem 3, but uses low-dimensional nested lattices. For such lattices, we suggest a fast implementation, whose computational efficiency does not depend on R. This simplified scheme attains performance fairly close to theoretical estimates

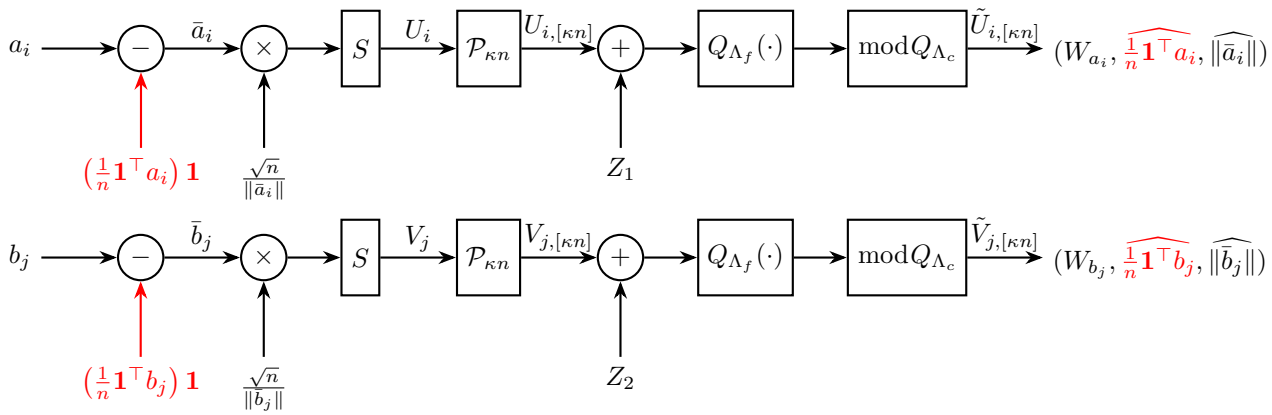

Fig. 1: Encoders for matrix multiplication. Each column of A is encoded by the same encoder, and each column of B is encoded by the same encoder. The encoder used for columns of A and that used for columns of B are also the same, except that for A we use the dither vector $Z_1 \in \mathbb{R}^{\kappa n}$, whereas for B we use the dither vector $Z_2 \in \mathbb{R}^{\kappa n}$. We illustrate the operation of the encoders on the ith column of A, $a_i \in \mathbb{R}^n$, and on the jth column of B, $b_j \in \mathbb{R}^n$. The block S corresponds to left multiplication by the rotation matrix $S \in \mathbb{R}^{n\times n}$, and the block $\mathcal{P}_{\kappa n}$ corresponds to projecting the vector $U_i \in \mathbb{R}^n$ (respectively $V_j \in \mathbb{R}^n$) to $\mathbb{R}^{\kappa n}$, $\kappa \in \frac{1}{n}\cdot\{0, 1, \ldots, n\}$, by keeping only its first κn coordinates. Here, κ is the time-sharing/sparsification parameter, determining the fraction of coordinates in each vector that are actually "described" to the decoder. The lattices $\Lambda_c \subset \Lambda_f \subset \mathbb{R}^{\kappa n}$ are nested. The component $Q_{\Lambda_f}(\cdot)$ is a lattice quantizer which maps a point in $\mathbb{R}^{\kappa n}$ to the closest lattice point in Λ_f. The component $\mathrm{mod}\,\Lambda_c$ maps a point $x \in \mathbb{R}^{\kappa n}$ to $x - Q_{\Lambda_c}(x) \in \mathcal{V}_c$, where $\mathcal{V}_c$ is the Voronoi region of Λ_c. The binary representation W_{a_i} (respectively W_{b_j}) is an encoding of $\tilde{U}_{i,[\kappa n]} \in (\Lambda_f \cap \mathcal{V}_c) \cong \Lambda_f/\Lambda_c$ (respectively $\tilde{V}_{j,[\kappa n]} \in \Lambda_f/\Lambda_c$) using $\log|\Lambda_f/\Lambda_c|$ bits. The scalars $\widehat{\frac{1}{n}\mathbf{1}^\top a_i}, \widehat{\|\bar{a}_i\|}$ (respectively, $\widehat{\frac{1}{n}\mathbf{1}^\top b_j}, \widehat{\|\bar{b}_j\|}$) are high-resolution descriptions of $\frac{1}{n}\mathbf{1}^\top a_i, \|\bar{a}_i\|$ (respectively, $\frac{1}{n}\mathbf{1}^\top b_j, \|\bar{b}_j\|$), which require only $O(\log n)$ bits. The dither vectors Z_1, Z_2 must be known to the decoder. They can be randomly drawn by the encoders and decoder and require sharing randomness between them (in practice, we just store random seed with the matrices). The matrix S need not be known by the decoder. The operations marked in red corresponds to zero-centering the column vectors, and may be avoided altogether. The effect of avoiding those operations on the performance is replacing $\bar{A}$ with A and $\bar{B}$ with B in the MSE upper bounds in Theorems 1, 3 and 12.

therein. Recent subsequent work [5] suggests that the resulting matrix quantization scheme may be a good candidate for practical application in LLMs and other algorithms relying on heavy matrix multiplication operations.

Additional contributions of this work include the following:

- We study the inner product case $a = b = 1$, in full generality, assuming the entries of A are drawn iid from distribution P, the entries of B are drawn iid from distribution Q, and the rates R_1 and R_2 are not necessarily equal. We derive several upper and lower bounds on the smallest attainable distortion in computing the inner product, and prove some results on the structure of the optimal encoders and decoder.
- For the matrix multiplication case, when the entries of A and B are drawn iid from a distribution P with zero mean and variance σ^2, we show that (7) continues to hold with $\Gamma(R)$ replaced by $\Gamma(R + D(P\|\mathcal{N}(0,\sigma^2))$.

Key ideas and proofs of these results are sketched in Section I-B. We proceed to motivation and review of the literature.

A. Importance of quantization for modern applications

To set the stage for the problem, let us estimate what level of quantization (in bits / entry) would be relevant for today's main consumer of matrix multiplications: the large language models (LLMs). For those, quantization is typically employed for accelerating inference. During inference LLM is busy computing many products $A^\top B$ of matrices with sizes $d \times a$ and $d \times b$ respectively. This requires $2abd$ FLOPs and $ad + bd + ab$ entries to load/store from memory. Ideally, we would want to quantize entries in such a way that all compute is fully utilized. For that we need to know the ratio ξ of available FLOPs to available memory bandwidth, a quantity known as "ops:bytes"

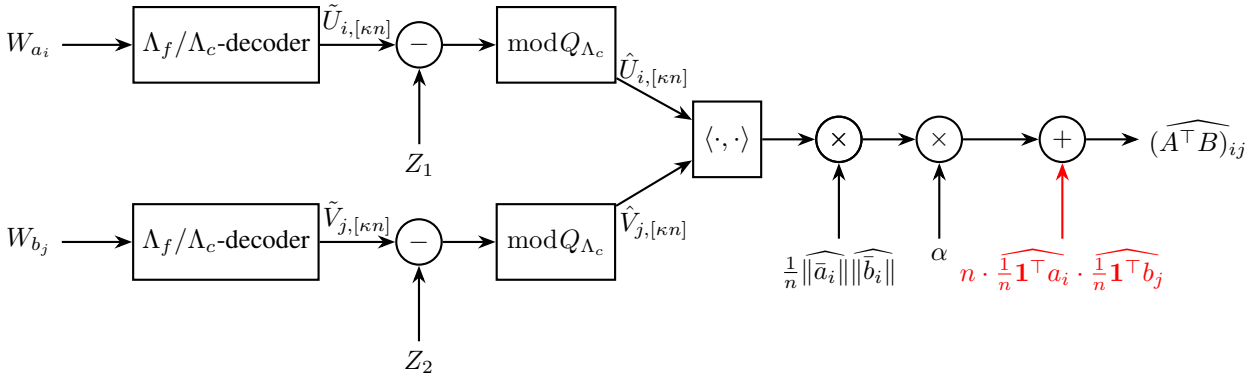

Fig. 2: Decoder for the matrix multiplication problem. We illustrate the estimation of $(A^\top B)_{ij}$. The component Λ_f/Λ_c-decoder maps $\log|\Lambda_f/\Lambda_c|$ bits to points in $\Lambda_f \cap \mathcal{V}_c \subset \mathbb{R}^{\kappa n}$, where $\mathcal{V}_c$ is the Voronoi region of the lattice Λ_c. The component $\langle\cdot,\cdot\rangle$ computes the inner product $\hat{U}_{i,[\kappa n]}^\top \hat{V}_{j,[\kappa n]}$, and $\alpha \in [0,1]$ is a (MMSE-like) scaling coefficient. The operation marked in red need only be implemented if the encoders implemented the corresponding zero-centering operations marked in red in Figure 1. Note that we can estimate the entire product $A^\top B$ by first decoding $\hat{\tilde{A}} = [\hat{U}_{1,[\kappa n]}|\cdots|\hat{U}_{a,[\kappa n]}]$ and $\hat{\tilde{B}} = [\hat{V}_{1,[\kappa n]}|\cdots|\hat{V}_{b,[\kappa n]}]$, computing the matrix $\alpha\hat{\tilde{A}}^\top\hat{\tilde{B}}$, and then computing its Kronecker product with the rank-1 matrix N whose ijth entry is $N_{ij} = \frac{1}{n}\widehat{\|\bar{a}_i\|}\widehat{\|\bar{b}_j\|}$, and adding to it the rank 1 matrix μ whose ijth entry is $\mu_{ij} = n \cdot \widehat{\frac{1}{n}\mathbf{1}^\top a_i} \cdot \widehat{\frac{1}{n}\mathbf{1}^\top b_j}$.

of a processor. It ranges from $\xi = 5\ldots20$ for modern CPUs (FP32 arithmetic via AVX512) to $\xi \approx 300$ for the fastest GPUs (FP16 on an H100 or B200). The quantization rate saturating compute should then be bounded (in bits/entry) as

$$R < \frac{16}{\xi}\frac{ab}{a+b+\frac{ab}{d}}\,. \tag{14}$$

It turns out that there are two stages of running inference with LLMs: the pre-fill (when the input prompt is processed) and the generation (when response tokens are sequentially generated). During the pre-fill LLM we have $a = d$ and $b = L$ (d is the so-called hidden dimension and L is the sequence length), while during the generation we have $a = L$ and $b = 1$ (the A matrix coming from KV-cache and B matrix being new token's embedding). Thus, to saturate the computation core, we need

$$R_{\text{pre-fill}} = \frac{16Ld}{\xi(d+2L)}\,, \qquad R_{\text{generate}} = \frac{16L}{\xi(L+1+L/d)} \approx \frac{16}{\xi}\,.$$

We can see that during generation phase, on CPUs we would want to approach 1-3 bits/entry, while on GPUs we will not be able to ever saturate compute (that is, a decrease in quantization rate translates proportionally to decrease in runtime). For the pre-fill phase, for large LLMs we get $R_{\text{generate}} > 16$ bit (that is, just storing plain FP16 is already good enough). Quantization during pre-fill might still be important for "small" LLMs running on fast GPUs: for example, for BERT [10] we have $L = 512$, $d = 768$ and $\xi = 300$ (for an H100), resulting in quantization rate $R \approx 11.7$ bit/entry.

Quantization is also important for reducing the power consumption of matrix multiplication. The energy cost of reading two bytes from DRAM is typically 2-3 orders of magnitude higher than that of performing a single INT8 addition/multiplication [11]. Consequently, for memory-limited matrix multiplication, quantization may also have a dramatic effect on the power consumption.

B. Sketch of the proof

This work started with the goal of trying to understand approximate matrix multiplication for two matrices A and B which are random, with iid Gaussian entries $\mathcal{N}(0,1)$. We started by trying to solve the case of $a = b = 1$ (Sections II and III), i.e. when $A^\top B$ is simply an inner product of two iid Gaussian vectors.

Recall that the Gaussian distortion-rate function is $D(R) = 2^{-2R}$, e.g. [12, Section 26.1.2]. A simple argument (Thm. 5) shows that compressing A to $\hat{A}$ and B to $\hat{B}$ via rate-R optimal Gaussian vector quantizer achieves error

$$\mathbb{E}[(\hat{A}^\top \hat{B} - A^\top B)^2] \le \phi(D(R)), \qquad \phi(x) := 2x - x^2 \,.$$

It turned out that the function $\phi(D(R))$ is monotonically decreasing but *not* convex. Thus, via time-sharing one can achieve a lower convex envelope of $\phi(D(R))$, which turns out to be the $\Gamma(R)$ function defined in (1).

We next proceed to lower bounds on distortion or, equivalently, to lower bounds on rate R required for the existence of encoders f_1, f_2 and decoder g satisfying

$$\mathbb{E}[(g(f_1(A), f_2(B)) - A^\top B)^2] \le nD \tag{15}$$

A simple oracle bound (by revealing B to the decoder) shows that rate R cannot be smaller than the standard Shannon rate-distortion function of A, see Theorem 4. However, this bound leaves a wide gap with the achievability bound given above. Next, by a standard data-processing argument (and observation that encoders for A and B can be without loss of generality be taken identical) in Section III-B we deduce that (15) requires rate

$$R \ge \limsup_{n\to\infty} \frac{1}{n} \inf_{\hat{A}} \{I(A; \hat{A}) : \frac{1}{n}\sum_{i=1}^n \phi(\lambda_i) \le D\} \,, \tag{16}$$

where $A \sim \mathcal{N}(0, I_n)$, infimum is over all $\mathbb{R}^n$-valued random variables $\hat{A}$ and $\{\lambda_i\}$ are the eigenvalues of $\mathrm{Cov}(A|\hat{A})$. This reduces inner product quantization to an optimization of a multi-letter mutual information. Notice that the distortion constraint is no longer separable, and hence the standard single-letterization (e.g. [12, Theorem 24.8]) does not work and the limit on the right-hand side is not possible to evaluate. For the special case of Gaussian distribution of entries of A we were able to single-letterize the expression on the right-hand side of (16), see Theorem 6, showing that left-hand side of (16) evaluates to $\Gamma^{-1}(D)$. Putting both upper and lower bounds together, we conclude that optimal compression rate for the iid Gaussian inner product problem is thus given by $\Gamma^{-1}(D)$, see Theorem 7.

We next proceed to solving the matrix case. Luckily, it turns out that for Gaussian iid matrices, again, the optimal compression for matrix multiplication of $A^\top B$ is asymptotically achieved by compressing each column separately via the use of optimal inner product quantizers, see Theorems 8 and 9.

Having solved the iid Gaussian case, we proceed to analyzing general (non-random) matrices and vectors. Specifically, for the inner product problem we first normalize each of the two vectors to have norm $\sqrt{n}$ and these norms are compressed using a separate high-resolution scalar quantizer. Next, normalized vectors are multiplied by a common random orthogonal matrix. This makes each resulting vector uniformly distributed on the sphere of radius $\sqrt{n}$, while their inner product is unchanged. As is well known a high-dimensional vector that is uniform on the sphere is very similar to an iid Gaussian vector (for example, in terms of joint distribution of small $O(\sqrt{n})$-sized subsets). Thus, we reduce the problem to (15) except this time $A_i, B_i \overset{iid}{\sim} \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, where $\rho = \frac{A^\top B}{\|A\|\cdot\|B\|}$. This slight change creates a crucial complication compared to the previous case of $\rho = 0$.

Indeed, suppose we are only tasked with quantizing B and A is given to the decoder undistorted. Because of dependence between two terms in the product $A^\top(B - \hat{B})$ we have to recourse to something like Cauchy-Schwarz, yielding

$$\mathbb{E}[(A^\top B - A^\top \hat{B})^2 \le \mathbb{E}[\|A\|^2 \|B - \hat{B}\|^2] = \Omega(n^2) \,.$$

Thus, using "black box" quantizers for A and B only yields n^2 performance guarantees violating (15). This is where *lattice quantization* comes in. Specifically, using the idea of dithering we can make a (randomized) quantizer whose quantization error $(B - \hat{B})$ becomes independent of B and A.

In order to guarantee finite quantization rate, we also need to "truncate" the infinite lattice, for which we use another key idea: a "good" nested lattice quantizer as in [13], [14], [15]. However, due to the nature of the problem we require construction of nested lattice pairs that satisfy stronger conditions than were known from prior work (see Theorem 13, whose proof builds upon heavy-lifting in a recent [16]). Overall, we construct quantizers for inner product problem of non-random vectors with a reconstruction error that depends only on the inner product between the vectors and their individual ℓ_2 norms, see Theorem 11. Since the performance bounds coincides with the lower bound for the iid Gaussian case, it turns out that the resulting quantizers are optimal and generally cannot be improved (except, possibly, in terms of finite-n performance). Together these steps complete proof of the main results quoted above.

Remark 1 (On zero-centering): One of the components in our quantization scheme is zero-centering, that is, subtracting from each vector its empirical mean prior to quantization. This is possible since the excess rate required for describing the empirical mean in high resolution is $O(\log n)$ bits and is negligible next to the cost of describing the entire vector which is $\Omega(n)$ bits. Describing the empirical mean in high resolution corresponds to describing the projection of the vector on $\frac{1}{n}\mathbf{1}$. In principle, we can actually choose in advance a subspace of $o\left(\frac{n}{\log n}\right)$ orthogonal directions and describe the projection of the vector to this subspace with high resolution and negligible cost for the quantization rate. However, it is not clear what choice of such subspaces will result in a good rate-distortion tradeoff for arbitrary matrices. In fact, our choice of describing the DC component of the vector is also quite arbitrary, and was only done to allow restricting attention to zero-mean data without loss of generality, as can be done in vector quantization under quadratic distortion.

C. Naive constructions: vector quantization, ϵ-nets and sketching

A crucial part of our construction is usage of common randomness between encoders and decoder. In order to demonstrate its importance, here we first analyze two obvious constructions and find out that they only achieve error of order $\Omega(n^2)$, while our Theorem 3 yields error $O(n)$.

The first natural question is to simply independently quantize A and B using a good vector quantizer (ϵ-net), to obtain $\hat{A}$ and $\hat{B}$, and then set $\widehat{A^\top B} = \hat{A}^\top \hat{B}$. This idea, unfortunately, does not work even for vectors on a (scaled) sphere $\mathbb{S}^{n-1}$ as the next proposition shows. We stress that in this setting the right-hand side of (9) is $O(n)$.

Proposition 1: Fix rate $R > 0$ and consider a special case of $A, B \in \sqrt{n}\mathbb{S}^{n-1}$ (inner product estimation). Then there exists $\delta = \delta(R) > 0$ and $n_0 = n_0(R)$ such that for any $n \geq n_0$ and any deterministic encoders f_1, f_2 (of rate R each) and decoder g, there exist a pair of vectors A, B with $\|A\| = \|B\| = \sqrt{n}$ such that

$$\|A^\top B - \widehat{A^\top B}\|^2 \geq \delta n^2 \,.$$

Proof. Classical results, e.g. [12, Corollary 27.4] show that there exists $\delta' > 0$ and a collection $\mathcal{C}$ of at least $2^{nR}+1$ points on $\sqrt{n}\mathbb{S}^{n-1}$ such that

$$\|A - \tilde{A}\| \geq \sqrt{n\delta'} \qquad \forall A \neq \tilde{A} \in \mathcal{C} \,.$$

By pigeonhole principle there should be such a pair with $f_1(A) = f_1(\tilde{A})$. Thus, for such a choice and any $B \in \sqrt{n}\mathbb{S}^{n-1}$ we have

$$\widehat{A^\top B} = \widehat{\tilde{A}^\top B} \,.$$

Choose $B = \sqrt{n}\frac{A-\tilde{A}}{\|A-\tilde{A}\|}$, then by the previous display we have

$$\begin{aligned}\sqrt{n\delta'} \leq \|A - \tilde{A}\| &= \frac{1}{\sqrt{n}}\|A^\top B - \tilde{A}^\top B\| \\ &= \frac{1}{\sqrt{n}}\|A^\top B - \widehat{A^\top B} + \widehat{\tilde{A}^\top B} - \tilde{A}^\top B\| \leq \frac{1}{\sqrt{n}}\left(\|A^\top B - \widehat{A^\top B}\| + \|\tilde{A}^\top B - \widehat{\tilde{A}^\top B}\|\right) .\end{aligned}$$

Thus, one of the terms in the parentheses should be at least $n\sqrt{\delta'}/2$, completing the proof. ■

So what makes deterministic quantizers fail and our lattice quantizer work? Note that in the former case, since A and B can be arbitrary the best we can do is apply a Cauchy-Schwarz estimate

$$(A^\top B - \hat{A}^\top B)^2 \leq \|A - \hat{A}\|^2\|B\|^2 \asymp n^2 \,.$$

Dithered quantizers prevent adversarial alignment of the error vector $A - \hat{A}$ and B. However, note that if we naively attempt to use dithered uniform quantizer $\epsilon\mathbb{Z}^n$, it would require rate $\log_2 \frac{\sqrt{n}}{\epsilon}$ due to $\|A\|_\infty$ possibly reaching out value of $\sqrt{n}$ over the $\sqrt{n}\mathbb{S}^{n-1}$, thus also failing to provide finite-rate and $O(n)$ MSE guarantee.

The next idea is that of sketching [17], which is often used in the literature on approximate matrix multiplication (see next section). Let $U \sim \mathcal{N}(0, I_n)$ and suppose it is 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

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

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 error given the budget of nR bits. Unfortunately, this idea also does not work, as a simple computation shows

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

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 $U_A U_B$. 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.

Despite both of these constructions yielding $\Omega(n^2)$ errors, our algorithm combines the two to obtain $O(n)$ error (and achieves the optimal constant): we first form the right number of sketches and then convert them to bits via dithered (nested) lattice quantizers.

At this point we stress that even after computing sketches of A and B via random rotation, using vector quantizers that achieve the optimal Gaussian RDF may not suffice. In fact, even a Gaussian RDF achieving dithered nested lattice quantizer may not suffice. The reason for this is that the error in approximating an inner product as the inner product between the two quantized vectors consists also of the inner product between the quantization errors. In order to make sure this term has small variance, one needs to control also the Frobenius norm of the quantization noise autocorrelation matrix, and not just its trace as in the standard Gaussian vector quantization setup. Our analysis show that dithered nested lattice quantizers that posses these required properties do exist.

D. Related work

Randomized linear algebra/sketching, and locality-sensitive hashing (LSH) are techniques widely used in practice for computing approximate inner products and approximate matrix multiplications, as well as other operations, in reduced dimensions. The figure of merit in these fields is typically the tradeoff between the reduced dimension and the approximation error. Since the dimension of the reduced matrix/vector is related to the number of bits required for storing it, this body of work is relevant to our study. However, the tradeoff between the number of bits per dimension and the total approximation error, and its dependence on the properties of A, B and $A^\top B$ is often subtle. Thus, there is no immediate translation between the required dimension of a sketch and the number of bits needed for representing it for obtaining the required accuracy.

Many algorithms have been developed for randomized linear algebra, see [18], [19] for a survey, and in particular for approximate matrix multiplication. A canonical example is the Monte-Carlo Matrix Multiplication (MCMM) algorithm [20] which randomly samples (the same) c rows from $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ and estimates $A^\top B$ as the (scaled) matrix multiplication of the sub-sampled matrices. Thus, each matrix is represented by ac (respectively bc real numbers), and the expected squared Frobenius norm of the approximation error is shown to scale like $O(\|A\|_F^2\|B\|_F^2/c)$. Similarly, LSH for cosine similarity or ℓ_2 distance also produce low-dimensional sketches of $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$, from which the inner product of $u^\top v$ can be approximated. Specifically, in [21] it is proposed to project the two vectors using c random projections (same random projections for both vectors) and only store the sign of each projection. The Hamming distance between the obtained vectors is distributed as $\mathrm{Binomial}\left(c, \frac{\theta(u,v)}{\pi}\right)$ where $\theta(u,v) = \cos\left(\frac{u^\top v}{\|u\|\cdot\|v\|}\right)$, such that the expected squared error in estimating $\theta(u,v)$ is $O(1/c)$. In [22] it is proposed to estimate $\|u-v\|$ (which is equivalent to estimating $u^\top v$ for u and v on the sphere) by computing Gaussian linear combinations of each of them, and using a (dithered) scalar quantizer for quantizing each of the entries of the linear combinations. Specifically, for a vector $G \in \mathbb{R}^n$, with iid $\mathcal{N}(0,1)$ entries, we have that $G^\top(u-v) \sim \mathcal{N}(0, \|u-v\|^2)$, and therefore the probability that $G^\top u$ and $G^\top v$ are quantized (after dithering) to the same value is a monotone function of $\|u-v\|$.

All the schemes mentioned above, as well as many other sketching/LSH schemes suffer from the same shortcoming: their relative error $\frac{\mathbb{E}\|\widehat{A^\top B} - A^\top B\|_F^2}{\|A^\top B\|_F^2}$ scales like $O\left(\frac{1}{c}\frac{\|A\|_F^2\|B\|_F^2}{\|A^\top B\|_F^2}\right)$, and typically these schemes are applied with constant c. When $\frac{\|A\|_F^2\|B\|_F^2}{\|A^\top B\|_F^2} = \Omega(1)$, these schemes perform remarkably well, despite the fact that c does not grow with n. However, when $\frac{\|A\|_F^2\|B\|_F^2}{\|A^\top B\|_F^2} = \omega(1)$, as is the case for random iid matrices, their relative error is very high. A notable exception is the algorithm proposed by Pagh in [23], which represents each matrix using $n\cdot\min\{m,a\}$ (respectively $n\cdot\min\{m,b\}$) real numbers, and produces an *unbiased* estimator for $A^\top B$ with expected error of $\mathbb{E}\left((\widehat{A^\top B})_{i,j} - (A^\top B)_{i,j}\right)^2 = O\left(\frac{\|A^\top B\|_F^2}{m}\right)$, for all i,j, and does so with runtime proportional to $n^2 + nm$

(ignoring logarithmic factors). When the product $A^\top B$ is known to be highly sparse, this allows to estimate the sparsity pattern with m proportional to the number of nonzero entries.

The topic of matrix quantization has received much attention in the last decade in the context of DNNs and LLMs. The goal here is to reduce the memory footprint of the weight matrices, allowing to load them to the main memory using less IOs, as well as speed up the multiplications and additions operations by moving from floating point numbers to small integers (and when possible, also sparsifying the matrices, saving some operations altogether). Roughly speaking, one can distinguish between two paradigms: *quantization-aware training*, where the training procedure is designed to output weight matrices with "cheap" representation [24], [25], and *post-training quantization*, where the training procedure is performed in high precision, and quantization of the weights is only performed after training has terminated (perhaps with some fine tuning afterwards) [26], [27], [28], [29], [30], [31], [6]. In order to further speed up matrix multiplication, and reduce the number of IOs needed for using KV-cache, some works also develop quantizers for the activations [27], [29], [30], [31], while other works assume the activations are kept in high precision [26], [6]. Quantization for DNNs and LLMs are typically evaluated according to the end-to-end performance of the quantized architecture, but often the Frobenius norm of the approximation error is considered as the intermediate optimization criterion for quantizing the weights at each layer [24], [32]. Some works rely on specific empirical observations on the distribution of weights and activations in LLMs. For example [28], [29], [30] exploit the fact that a small subset of entries in the activations have significantly larger magnitude than the majority of entries. Notably, in [31] it is observed that for large LLMs, quantizing all weights to $\{-1, 0, 1\}$ and the activations to 8 bits, hardly affects the overall performance. Among the work described above, the algorithm from [6] is closest to the scheme we use in the proof of our Theorem 1 and Theorem 3, as well as the practical adaptation of the scheme used in those proofs, which is described in Section VII. The work [6] develops an algorithm for quantizing the weight matrices (keeping the activations in high precision) that is based on random rotation using the randomized Hadamard transform (that can be performed in time $n \log n$) and then using the E_8 lattice for quantizing the rotated matrix. The mapping from lattice points to bits that was used in [6] required access to a lookup table, and was tailor-designed for $R = 2$, while using different rates requires to further use successive refinement (residual vector quantization). While our practical scheme in Section VII also uses product-lattice quantization, we use a nested lattice quantizer/Voronoi code [9], which results in a simple mapping from lattice points to bits, regardless of R. Furthermore, we quantize both matrices to be multiplied. In LLMs, when activations/KV-cache data is also compressed, quantization must occur in inference time, and the encoders are required to be fast. On the other hand, when weights-only quantization is performed, the encoding is done offline, and only decoding is required to be efficient. The work [33] leverages this asymmetry and builds a complicated encoder based on a trellis with a large number of states, while the decoder, on the other hand, is highly efficient. Such asymmetric schemes are not suitable for quantizing activations/KV-cache, whereas in the scheme we describe in Section VII both the encoders and the decoder are efficient, and can be both applied in inference time. In addition to reducing the limitations incurred by the memory bandwidth, an additional benefit of quantizing both matrices, is that one can replace the decoder with a lookup table, as in [34], [35], [36], [37], [38], resulting in very fast decoding in CPUs.

Following [6], the QuaRot [7] scheme also uses randomized Hadamard transform prior to quantization, followed by 4-bit scalar quantization of each entry in both rotated matrices. Our implementation in Section VII quantizes the entries of the rotated matrices using nested-lattice codes, which come much closer to the optimal rate-distortion tradeoff than scalar quantizers, with essentially the same complexity (provided that the base lattice has an efficient nearest-neighbor decoder, as is the case for essentially all "good" lattices in dimensions $2, 3, 4$ and 8).

To the best of our knowledge, there was very little work on distributed compression for inner product/matrix multiplication in the information theory literature. Recently, Malak [39] studied the problem of *lossless* distributed compression of binary random matrices for computing their product, and derived non-trivial bounds under stringent assumptions on the joint distribution. Some prior work considered the problem of distributed compression of random vectors with the goal of approximately computing a linear function of those vectors [40], [41]. In those works, the goal was to estimate, say, the difference between the two vectors in $\mathbb{R}^n$, which is itself a vector in $\mathbb{R}^n$. While the inner product of these vectors, which is a scalar in $\mathbb{R}$, can be computed from their difference (assuming their individual norms were encoded in high resolution), it seems, a-priory, that distributed compression for inner product computation is an easier task. Our results show that this is, in fact, hardly the case. Another line of related work in the information theory literature, is that of Ingber et al. [42] that considered the fundamental limits of lossy compression of a database in order to support approximate nearest neighbor search (see also [43] for a practical implementation). We note in passing that much recent work focused on coding for speeding up distributed matrix

multiplication by introducing redundancy for mitigating the effect of "slow workers" (stragglers), see, e.g., [44]. This line of work is not directly related to approximate matrix multiplication via compression, studied in this paper.

Finally, one may wonder if approximating matrix product in mean squared error (MSE) metric is the right distortion metric. Indeed, it was shown in [45] that if the high-dimensional vectors to be compressed are probability distributions and the metric is KL divergence (reasonable assumptions for attention matrices in LLMs), the optimal quantization algorithms become quite different from the MSE ones. We leave these extensions for future work.

To summarize, the main innovations of this work with respect to prior work are:

a. Our work provides, for the first time, the fundamental limits of quantization for matrix multiplication. We derive a non-asymptotic lower bound on the error of any quantization algorithm for the case of Gaussian iid matrices, and develop a "robust" quantization algorithm (that makes no assumptions on the matrices A, B) that asymptotically attains it. This gives a useful benchmark for evaluating the performance of any robust quantization algorithm.
b. On the algorithmic side, we introduce several new components that were not used in previous work on quantization for matrix multiplication: sparsification/time-sharing, MMSE scaling, *nested* lattice quantization. Those components, together with randomization in the form of rotation and dithering are required for attaining the optimal performance. For the analysis, we also prove new results on the existence of high-dimensional lattices with the required properties for quantized matrix multiplication.
c. We develop a low-complexity framework for approaching our theoretic lower bounds. Our framework is based on Voronoi codes in low dimensions, but together with an overload avoidance mechanism it nevertheless performs quite close to the asymptotic limits. It allows for fast encoding and decoding, and works for any $R > \log q$, where q is an integer. Unlike prior work on lattice-based weights-only quantization, the same scheme can be used for any such rate, and the encoding/decoding complexity is invariant to the quantization rate R. We note that based on these ideas a follow-up paper [5] developed a practical algorithm (NestQuant), which attains state of the art in 4-bit weight+activation quantization of LLMs.
d. Our lower and upper bounds give a theoretic justification for the widely used idea of applying the same random rotation to both matrices A and B prior to quantization. In particular, the schemes used in the proofs of Theorem 1 and Theorem 3 are based on random rotation followed by quantizers based on "good" high-dimensional nested lattices. Our analysis reveals that using ℓ_∞ normalization followed by quantization to $\mathbb{Z}^n$ on the rotated vectors (e.g. [7]) is highly sub-optimal, and using "good" nested lattices instead, leads to a multiplicative reduction by of factor $O(\log n)$ in the resulting distortion, see (13).

E. Paper organization

We begin our study with the special case where $a = b = 1$, so that matrix multiplication becomes an inner product. The reason is twofold: First, it is easier to gain intuition for this problem, and all techniques for proving converse (impossibility) results for the inner product case, easily extend to the matrix multiplication case. The second reason is that our achievability results are based on compression of each column of A separately and compression of each column of B separately, and estimating each inner product forming $C_{ij} = (A^\top B)_{i,j} = a_i^\top b_j$ separately. In Section II we formally define the compression for inner product computation problem, identify the structure of the optimal decoder, and give simple expressions on the attained distortion as a function of the encoders f_1 and f_2, as well as a simple lower bound on the distortion in terms of the "standard" distortion-rate function. In Section III we restrict attention to the symmetric case where the two vectors have the same distribution, and both encoders have the same rate. We prove lower and upper bounds on the smallest attainable distortion in this case, which coincide in the Gaussian case. In Section IV we generalize the inner product results for matrix multiplication $A^\top B$ of $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$, for general a and b. Building on the bounds developed for the inner product case of $a = b = 1$, we prove lower and upper bound on the smallest expected squared Frobenius norm of the error. In the special case where all entries in both matrices are iid Gaussian, the lower bound results in Theorem 2. In Section V we develop a quantization scheme, based on randomly rotating both A and B by the same rotation matrix, and then using nested-lattice quantizers for separately quantizing each column of the rotated matrices, for qunatization of arbitrary matrices $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$. The expected squared Frobenius norm of the approximation error attained by this scheme is upper bounded in Theorem 12. Our main results stated above, Theorem 1 and Theorem 3, are obtained as simple corollaries of Theorem 12. The upper bound depends on A and B only through $\|\bar{A}\|_F, \|\bar{B}\|_F, \|\bar{A}^\top \bar{B}\|_F$, and meets the lower bound from Theorem 2 for the case where A and B have Gaussian iid entries. The main component in the proof of Theorem 12 is a nested lattice quantization scheme for inner product computation of

two arbitrarily correlated vectors, each uniformly distributed on the sphere. This coding scheme is presented and analyzed in Section VI. For the analysis we also prove new lattice existence results, stated in Theorem 13. Finally, in Section VII we introduce a practical implementation of the quantization scheme from Theorem 12 for matrix multiplication of arbitrary matrices. In these scheme, we describe several compromises in the choice of lattices used for quantization, as well as in the rotation matrix used for rotating both A and B. With these compromises the quantization scheme and the decoder become extremely simple and fast. Some numerical evidence show that, nevertheless, the resulting approximation error is quite close to the lower bound from Theorem 2. We conclude the paper with stating several key open problems in Section VIII.

F. Notation

For $x > 0$ we denote by $\log(x)$ the logarithm of x taken to base 2, and by $\ln(x)$ the natural logarithm. We denote the Euclidean (ℓ_2) norm of a vector $x \in \mathbb{R}^n$ by $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ and its ℓ_1 norm by $\|x\|_1 = \sum_{i=1}^n |x_i|$. For a matrix $A \in \mathbb{R}^{n\times m}$ we denote the trace operation as $\mathrm{tr}(A) = \sum_{i=1}^n A_{ii}$, and the Frobenius norm is $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)}$. The multiset of eigenvalues of a square matrix $A \in \mathbb{R}^{n\times n}$ is denote by $\mathrm{eig}(A) = (\lambda_1, \ldots, \lambda_n)$. For $y > 1$ we denote $[y] = \{1, \ldots, \lfloor y \rfloor\}$. We denote the all ones vector in $\mathbb{R}^n$ by $\mathbf{1} = [1, \ldots, 1]^\top$. For a column vector $x \in \mathbb{R}^n$ we denote by $\bar{x} = x - \left(\frac{1}{n}\mathbf{1}^\top x\right)\mathbf{1}$ its zero-centered version. For a matrix $A = [a_1|\cdots|a_a] \in \mathbb{R}^{n\times a}$ with columns $\{a_i\}$ we denote $\bar{A} = [\bar{a}_1|\ldots|\bar{a}_a]$ its column-centered version. The differential entropy, mutual information, and KL divergence are denoted by $h(\cdot)$, $I(\cdot\,;\,\cdot)$, and $D(\cdot\|\cdot)$. The Gaussian distribution in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ is denoted $\mathcal{N}(\mu, \Sigma)$. For two random variables X, Y the notation $X \perp\!\!\!\perp Y$ means that they are statistically independent. For a distribution P on alphabet $\mathcal{X}$, the distribution $P^{\otimes n}$ is its n-product (iid) distribution on $\mathcal{X}^n$.

II. Compression for Inner Product Computation: General Problem Setup and Simple Bounds

Let P and Q be distributions on $\mathbb{R}$ with zero mean and unit variance, and let $U \sim P^{\otimes n}$ and $V \sim Q^{\otimes n}$ be statistically independent. As we argue below, the unit variance assumption is without loss of generality, and the zero mean assumption is essentially without loss of generality, as the effect of non-zero mean on the performance can be made negligible for large n. We consider the problem of quantizing U and V in order to compute their inner product $U^\top V$. In particular, an (n, R_1, R_2, D) code consists of mappings

$$f_1 : \mathbb{R}^n \to [2^{nR_1}] \tag{17}$$

$$f_2 : \mathbb{R}^n \to [2^{nR_2}] \tag{18}$$

$$g : [2^{nR_1}] \times [2^{nR_2}] \to \mathbb{R}, \tag{19}$$

with

$$D = D^{\mathrm{IP}} = \frac{1}{n}\mathbb{E}\left(U^\top V - g(f_1(U), f_2(V))\right)^2. \tag{20}$$

We define

$$D_n^{\mathrm{IP},*}(R_1, R_2) = D_n^{\mathrm{IP},*}(R_1, R_2, P, Q) = \inf\left\{D \;:\; \exists(n, R_1, R_2, D) - \text{code}\right\}. \tag{21}$$

We further define the asymptotic function

$$D^{\mathrm{IP}}(R_1, R_2) = D^{\mathrm{IP}}(R_1, R_2, P, Q) = \limsup_{n\to\infty} D_n^{\mathrm{IP},*}(R_1, R_2). \tag{22}$$

To see that the assumption that P and Q have unit variance is without loss of generality, assume that $\tilde{U} \sim \tilde{P}^{\otimes n}$ and $\tilde{V} \sim \tilde{Q}^{\otimes n}$, such that $\mathrm{Var}[\tilde{U}] = \sigma_1^2$ and $\mathrm{Var}[\tilde{V}] = \sigma_2^2$, and we would like to quantize $\tilde{U}$ and $\tilde{V}$ in order to estimate $\tilde{U}^\top\tilde{V}$. To that end we may define the unit-variance random variables $U = \frac{\tilde{U}}{\sigma_1}$ and $V = \frac{\tilde{V}}{\sigma_2}$ with corresponding distributions P and Q, compress them using $f_1(U)$ and $f_2(V)$, and estimate the inner product as

$$\widehat{\tilde{U}^\top\tilde{V}} = \sigma_1\sigma_2 \cdot g(f_1(U), f_2(V)), \tag{23}$$

where f_1, f_2, g attain $D_n^{\mathrm{IP},*}(R_1, R_2)$ for P and Q. This scheme will achieve

$$\mathbb{E}\left(\tilde{U}^\top\tilde{V} - \widehat{\tilde{U}^\top\tilde{V}}\right)^2 = \sigma_1^2\sigma_2^2 \cdot \mathbb{E}\left(U^\top V - g(f_1(U), f_2(V))\right)^2 = \sigma_1^2\sigma_2^2 \cdot D_n^{\mathrm{IP},*}(R_1, R_2). \tag{24}$$

This must be optimal, since otherwise we could have attained a smaller distortion for P and Q by first scaling U and V by σ_1 and σ_2, respectively, feeding them to the better inner product quantization system, and scaling the obtained inner product estimate by $\frac{1}{\sigma_1\sigma_2}$.

Next, let us address the zero-mean assumption. Let P and Q be zero-mean distributions, $U \sim P^{\otimes n}$, $V \sim Q^{\otimes n}$, and let $\tilde{U} = U + \mu_U \mathbf{1}$ and $\tilde{V} = V + \mu_V \mathbf{1}$ for some $\mu_U, \mu_V \in \mathbb{R}$. To encode $\tilde{U}$ and $\tilde{V}$ we may use the encoders f_1, f_2 designed for P, Q and send

$$\tilde{f}_1(\tilde{U}) = \left[f_1(\tilde{U} - \mu_U \mathbf{1}), \mathbf{1}^\top (\tilde{U} - \mu_U \mathbf{1})\right] = \left[f_1(U), \mathbf{1}^\top U\right], \tag{25}$$

$$\tilde{f}_2(\tilde{V}) = \left[f_2(\tilde{V} - \mu_V \mathbf{1}), \mathbf{1}^\top (\tilde{V} - \mu_V \mathbf{1})\right] = \left[f_2(V), \mathbf{1}^\top V\right], \tag{26}$$

and estimate the inner product $\tilde{U}^\top \tilde{V}$ as

$$\widehat{\tilde{U}^\top \tilde{V}} = g\left(f_1(U), f_2(V)\right) + n \cdot \mu_U \mu_V + \mu_U \mathbf{1}^\top V + \mu_V \mathbf{1}^\top U, \tag{27}$$

so that

$$\begin{aligned} \tilde{U}^\top \tilde{V} - \widehat{\tilde{U}^\top \tilde{V}} &= (U + \mu_U \mathbf{1})^\top (V + \mu_V \mathbf{1}) - \left[g(f_1(U), f_2(V)) + n \cdot \mu_U \mu_V + \mu_U \mathbf{1}^\top V + \mu_V \mathbf{1}^\top U\right] \\ &= U^\top V - g(f_1(U), f_2(V)). \end{aligned} \tag{28}$$

Thus, the error in the case of zero mean U, V and non-zero mean $\tilde{U} = U + \mu_U \mathbf{1}$, $\tilde{V} = V + \mu_V \mathbf{1}$ can be made the same, at the expense of also sending a description of $\mathbf{1}^T U$ and $\mathbf{1}^T V$. As those are scalars, they can be described to high resolution, say $O(n^{-2})$ using $O(\log n)$ bits. Thus, for large n and finite $R_1, R_2 > 0$, the cost of those descriptions is negligible.

Some of our bounds will rely on the distortion-rate function of $\mathbb{R}$-valued source under quadratic distortion. An (n, R, D) code for a source $U \sim P^{\otimes n}$ consists of an encoder $f : \mathbb{R}^n \to [2^{nR}]$ and a decoder $g : [2^{nR}] \to \mathbb{R}^n$ with $D = \frac{1}{n}\mathbb{E}\|U - g(f(U))\|^2$. We denote by $D_n^*(R) = D_n^*(R, P)$ the smallest distortion attained by any (n, R, D) code, and we denote the distortion-rate function by [12]

$$D_P(R) = \lim_{n\to\infty} D_n^*(R, P) = \min_{P_{Y|U}: I(U;Y) \le R} \mathbb{E}(U - Y)^2. \tag{29}$$

It is also well-know [12], that $D_n^*(R, P) \ge D_P(R)$ for any $n \ge 1$.

A. Optimal Decoder and Error Expressions

In the following, we assume f_1 and f_2 are fixed. We denote $W_U = f_1(U)$ and $W_V = f_2(V)$. Let $\hat{U} = \mathbb{E}[U|W_U]$ and $\hat{V} = \mathbb{E}[V|W_V]$.

Proposition 2: The optimal choice for g is $g^*(W_U, W_V) = \hat{U}^\top \hat{V}$.

Proof. The minimum mean squared error (MMSE) estimator of a random variable X from another random variable Y is $\hat{X} = \mathbb{E}[X|Y]$. Thus,

$$g^*(W_U, W_V) = \mathbb{E}[U^\top V | W_U, W_V] = \mathbb{E}[U^\top | W_U]\mathbb{E}[V|W_V] = \hat{U}^\top \hat{V}, \tag{30}$$

where the second equality follows since $(U, W_U) \perp\!\!\!\perp (V, W_V)$. ■

Let $e_U = U - \hat{U}$ and $\Sigma_{e_U} = \mathbb{E}[(U - \hat{U})(U - \hat{U})^\top]$. Similarly, let $e_V = V - \hat{V}$ and $\Sigma_{e_V} = \mathbb{E}[(V - \hat{V})(V - \hat{V})^\top]$. Recall that by the orthogonality principle[46, Chapter 4.2], it holds that $\mathbb{E}[\hat{U} e_U^\top] = 0$ and $\mathbb{E}[\hat{V} e_V^\top] = 0$.

Proposition 3: Assuming that entries of U and V have zero mean and unit variance, we have that the optimal decoder achieves

$$\mathbb{E}\left(U^\top V - g^*(W_U, W_V)\right)^2 = \frac{1}{n}\left[\mathrm{tr}(\Sigma_{e_V}) + \mathrm{tr}(\Sigma_{e_U}) - \mathrm{tr}(\Sigma_{e_U}\Sigma_{e_V})\right] \tag{31}$$

Proof. We have

$$D^{\mathrm{IP}} = \mathbb{E}\left((\hat{U} + e_U)^\top (\hat{V} + e_V) - \hat{U}^\top \hat{V}\right)^2 \tag{32}$$

$$= \mathbb{E}\left(\hat{U}^\top e_V + \hat{V}^\top e_U + e_U^\top e_V\right)^2 \tag{33}$$

$$= \mathbb{E}\left(\hat{U}^\top e_V\right)^2 + \mathbb{E}\left(\hat{V}^\top e_U\right)^2 + \mathbb{E}\left(e_U^\top e_V\right)^2, \tag{34}$$

where the last transition is due to the orthogonality principle and the statistical independence of U and V. We have that

$$\mathbb{E}\left(\hat{U}^\top e_V\right)^2 = \mathrm{tr}\left[\mathbb{E}[\hat{U}\hat{U}^\top e_V e_V^\top]\right] = \mathrm{tr}\left[\mathbb{E}[\hat{U}\hat{U^\top}]\Sigma_{e_V}\right] \tag{35}$$

Recalling that $\mathbb{E}[\hat{U}\hat{U^\top}] = I - \Sigma_{e_U}$, again, by the orthogonality principle, we obtain

$$\mathbb{E}\left(\hat{U}^\top e_V\right)^2 = \mathrm{tr}\left[(I - \Sigma_{e_U})\Sigma_{e_V}\right] = \mathrm{tr}(\Sigma_{e_V}) - \mathrm{tr}(\Sigma_{e_V}\Sigma_{e_U}), \tag{36}$$

Similarly,

$$\mathbb{E}\left(\hat{V}^\top e_U\right)^2 = \mathrm{tr}(\Sigma_{e_U}) - \mathrm{tr}(\Sigma_{e_V}\Sigma_{e_U}). \tag{37}$$

Finally,

$$\mathbb{E}\left(e_U^\top e_V\right)^2 = \mathrm{tr}\left[\mathbb{E}[e_U e_U^\top e_V e_V^\top]\right] = \mathrm{tr}(\Sigma_{e_U}\Sigma_{e_V}). \tag{38}$$

■

B. Simple Lower Bounds

We show that computing the inner product with mean squared error (MSE) of nD is necessarily harder than compressing each of the random vectors U and V with ℓ_2 norm of nD. Note that in the inner product quantization problem we are only interested in a single scalar in $\mathbb{R}$ while in the standard problem of quantizing a random vector we are interested in a vector in $\mathbb{R}^n$. Yet, the former problem is at least as hard as the latter.

Theorem 4: For any $n \geq 1$

$$D_n^{\mathrm{IP},*}(R_1, R_2, P, Q) \geq \max\left\{D_P(R_1), D_Q(R_2)\right\}, \tag{39}$$

and in particular

$$D^{\mathrm{IP}}(R_1, R_2, P, Q) \geq \max\left\{D_P(R_1), D_Q(R_2)\right\}. \tag{40}$$

Proof. From Proposition 3 we have that for any $f_1 : \mathbb{R}^n \to [2^{nR_1}]$ and $f_2 : \mathbb{R}^n \to [2^{nR_2}]$

$$\mathbb{E}\left(U^\top V - g^*(W_U, W_V)\right)^2 = \frac{1}{n}\left[\mathrm{tr}(\Sigma_{e_V}) + \mathrm{tr}(\Sigma_{e_U}) - \mathrm{tr}(\Sigma_{e_U}\Sigma_{e_V})\right] \tag{41}$$

$$= \frac{1}{n}\left[\mathrm{tr}(\Sigma_{e_U}) + \mathbb{E}(\hat{U}^\top e_V)^2\right] \tag{42}$$

$$\geq \frac{1}{n}\mathrm{tr}(\Sigma_{e_U}) \tag{43}$$

$$\geq D_n^*(R_1, P), \tag{44}$$

where (42) follows from (36), and the last inequality follows since W_U is an encoding of U with 2^{nR_1} codewords, which must incur distortion at least $D_n^*(R_1, P)$ by definition. Note that the inequality (43) holds with equality in the "single-sided" case where only U is quantized while $\hat{V} = V$, so that $e_V = 0$. The bound $D_n^*(R_1, R_2, P, Q) \geq D_n^*(R_2, Q)$ follows similarly. Our statement now follows since $D_n^*(R_1, P) \geq D_P(R_1)$ and $D_n^*(R_2, Q) \geq D_Q(R_2)$ for any $n \geq 1$. ■

III. Compression for Inner Product Computation: The Symmetric Case

In this section we assume $P = Q$, $R_1 = R_2 = R$, and define $D_n^{\mathrm{IP},*}(R, P) = D_n^*(R, R, P, P)$, and $D^{\mathrm{IP}}(R, P) = D^{\mathrm{IP}}(R, R, P, P)$. We first develop a simple upper bound based on using the same encoder for both vectors (that is $f = f_1 = f_2$), that time-shares between a "good" encoder for P under quadratic distortion, and a zero-rate encoder. We then develop a lower bound on the distortion of inner product compression, which shows that for the symmetric case, using the same encoder $f = f_1 = f_2$ for both U and V is optimal, and depends on the spectrum of the covariance matrix of $e_U = U - \mathbb{E}[U|f(U)]$. We then give some constraints on the error spectrum that can be attained by a rate R encoder. Using this characterization we obtain a general lower on $D^{\mathrm{IP}}(R, P)$. Thus, overall we show in this section that for any iid source $P = Q$ with $\mathbb{E}_{U_i \sim P}[U_i] = 0$, $\mathbb{E}_{U_i \sim P}[U_i^2] = 1$ we have

$$\Gamma(R + D(P\|\mathcal{N}(0,1))) \leq D^{\mathrm{IP}}(R, P) \leq \Gamma(R),$$

in particular showing that $D^{\mathrm{IP}}(R, \mathcal{N}(0,1)) = \Gamma(R)$.

A. Upper Bound

Define the function

$$\phi(x) = 2x - x^2. \tag{45}$$

and note that $x \mapsto \phi(x)$ is increasing and concave on $[0,1]$. We give a time-sharing upper bound on $D^{\mathrm{IP}}(R,P)$ in terms of $\phi(D_P(R))$.

Theorem 5: Assuming that P has zero mean and unit variance, we have

$$D^{\mathrm{IP}}(R,P) \le \min_{0\le\kappa\le 1} (1-\kappa) + \kappa\cdot\phi\left(D_P\left(\frac{R}{\kappa}\right)\right) \le \Gamma(R)\,, \tag{46}$$

where $\Gamma(R)$ is defined in (1).

Proof. Note that it is sufficient to prove the first inequality. Indeed, we know that $D_P(R) \le D_{\mathcal{N}(0,1)}(R) = 2^{-2R}$, e.g. [12, Theorem 26.3], and in Appendix A we show that

$$\Gamma(R) = \min_{0\le\kappa\le 1} (1-\kappa) + \kappa\cdot\phi\left(2^{-2\frac{R}{\kappa}}\right)\,.$$

In order to show the first inequality, we will prove that

$$D_n^{\mathrm{IP},*}(R,P) \le \min_{\kappa\in\frac{1}{n}\{0,1,\ldots,n\}} (1-\kappa) + \kappa\cdot\phi\left(D^*_{\kappa n}\left(\frac{R}{\kappa},P\right)\right) \tag{47}$$

from which the statement immediately follows. Let $\kappa \in \frac{1}{n}\{0,1,\ldots,n\}$, and consider a compressor for $P^{\otimes\kappa n}$ under quadratic distortion: $f:\mathbb{R}^{\kappa n}\to[2^{nR}=2^{n\kappa\frac{R}{\kappa}}]$ and corresponding optimal decoder $g:[2^{nR}=2^{n\kappa\frac{R}{\kappa}}]\to\mathbb{R}^{\kappa n}$, $g(w)=\hat{U}^{\kappa n}=\mathbb{E}[U^{\kappa n}|f(U^{\kappa n})=w]$, that attains

$$D = \frac{1}{\kappa n}\mathbb{E}\|U^{\kappa n}-\hat{U}^{\kappa n}\|^2 = \frac{1}{\kappa n}\operatorname{tr}(\Sigma_{e_{U^{\kappa n}}}). \tag{48}$$

We encode U by applying f on $U^{\kappa n}$ and do not describe the other coordinates. The resulting covariance error matrix is therefore block diagonal of the form

$$\Sigma_{e_U} = \begin{bmatrix} \Sigma_{e_{U^{\kappa n}}} & 0 \\ 0 & I_{(1-\kappa)n} \end{bmatrix}. \tag{49}$$

Consequently,

$$\operatorname{tr}(\Sigma_{e_U}) = \operatorname{tr}(\Sigma_{e_{U^{\kappa}{}_n}}) + \operatorname{tr}(I_{(1-\kappa)n}) = n\kappa D + n(1-\kappa) \tag{50}$$

$$\operatorname{tr}(\Sigma_{e_U}\Sigma_{e_U}) = \operatorname{tr}(\Sigma_{e_{U^{\kappa}{}_n}}\Sigma_{e_{U^{\kappa}{}_n}}) + \operatorname{tr}(I_{(1-\kappa)n}) = \|\Sigma_{e_{U^{\kappa}{}_n}}\|_F^2 + n(1-\kappa). \tag{51}$$

Recall that for a positive semi-definite matrix $A\in\mathbb{R}^{m\times m}$ it holds that $\|A\|_F^2 \ge \frac{1}{m}(\operatorname{tr}(A))^2$. This follows since the vector $\lambda = \operatorname{eig}(A)$ has non-negative entries, so that $\operatorname{tr}(A) = \|\lambda\|_1$, and therefore $\|A\|_F^2 = \|\lambda\|_2^2 \ge \frac{1}{m}\|\lambda\|_1^2 = \frac{1}{m}(\operatorname{tr}(A))^2$. Thus,

$$\|\Sigma_{e_{U^{\kappa}{}_n}}\|_F^2 \ge \frac{1}{\kappa n}(\operatorname{tr}(\Sigma_{e_{U^{\kappa}{}_n}}))^2 = \kappa n D^2, \tag{52}$$

and, by (51), we have

$$\operatorname{tr}(\Sigma_{e_U}\Sigma_{e_U}) \ge n\kappa D^2 + n(1-\kappa). \tag{53}$$

We use the same encoder also for encoding V, such that $\Sigma_{e_V} = \Sigma_{e_U}$, and use the optimal decoder g^* for estimating $U^\top V$. Applying Proposition 3, we obtain

$$D^{\mathrm{IP}} = \frac{1}{n}\left[\operatorname{tr}(\Sigma_{e_U}) + \operatorname{tr}(\Sigma_{e_V}) - \operatorname{tr}(\Sigma_{e_U}\Sigma_{e_V})\right] \tag{54}$$

$$= \frac{1}{n}\left[2\operatorname{tr}(\Sigma_{e_U}) - \|\Sigma_{e_U}\|_F^2\right] \tag{55}$$

$$\le (1-\kappa) + \kappa\cdot(2D - D^2) \tag{56}$$

$$= (1-\kappa) + \kappa\phi(D). \tag{57}$$

Taking the compressor f that attains $D^*_{\kappa n}\left(\frac{R}{\kappa},P\right)$, we obtain the claimed result. ■

B. Lower Bound

Lemma 1: For the symmetric case, assuming that P has zero mean and unit variance, there is no loss of optimality in taking $f_1 = f_2 = f$, and

$$D_n^{\mathrm{IP},*}(R,P) = \frac{1}{n}\inf_f \left[2\|\lambda(f)\|_1 - \|\lambda(f)\|_2^2\right] = \frac{1}{n}\inf_f \sum_{i=1}^n \phi\left(\lambda_i(f)\right), \tag{58}$$

where the infimum runs over all encoders $f : \mathbb{R}^n \to [2^{nR}]$, and

$$\lambda(f) = \mathrm{eig}\left(\Sigma_{e_U^f}\right), \tag{59}$$

where $e_U^f = U - \mathbb{E}[U|f(U)]$, $\Sigma_{e_U^f} = \mathbb{E}[e_U^f e_U^{f,\top}]$.

Proof. By Proposition 3, we have that for any two encoders $f_1 : \mathbb{R}^n \to [2^{nR}]$ and $f_2 : \mathbb{R}^n \to [2^{nR}]$, when the optimal decoder is used, it holds that

$$D^{\mathrm{IP}} = \frac{1}{n}\left[\mathrm{tr}(\Sigma_{e_U^{f_1}}) + \mathrm{tr}(\Sigma_{e_V^{f_2}}) - \mathrm{tr}(\Sigma_{e_U^{f_1}}\Sigma_{e_V^{f_2}})\right] \tag{60}$$

$$= \frac{1}{n}\left[\mathrm{tr}(\Sigma_{e_U^{f_1}}) + \mathrm{tr}(\Sigma_{e_U^{f_2}}) - \mathrm{tr}(\Sigma_{e_U^{f_1}}\Sigma_{e_U^{f_2}})\right], \tag{61}$$

where the last equality follows since $P = Q$, and therefore U and V have the same distribution. Rearranging (61), we obtain

$$n(1 - D^{\mathrm{IP}}) = \mathrm{tr}((I_n - \Sigma_{e_U^{f_1}})(I_n - \Sigma_{e_U^{f_2}})) \tag{62}$$

$$\leq \sqrt{\mathrm{tr}((I_n - \Sigma_{e_U^{f_1}})^2)\,\mathrm{tr}((I_n - \Sigma_{e_U^{f_2}})^2)} \tag{63}$$

$$\leq \max_{i\in\{1,2\}} \mathrm{tr}((I - \Sigma_{e_U^{f_i}})^2) \tag{64}$$

where (63) follows since $(C, D) \mapsto \mathrm{tr}(CD)$ defines inner product on symmetric matrices $C, D \in \mathbb{R}^{n\times n}$, and therefore the Cauchy-Schwartz inequality holds. We have therefore obtained

$$D^{\mathrm{IP}} \geq \frac{1}{n}\min_{i\in\{1,2\}} (2\,\mathrm{tr}\,\Sigma_{e_U^{f_i}} - \mathrm{tr}\,\Sigma_{e_U^{f_i}}^2) \tag{65}$$

$$\geq \frac{1}{n}\inf_f \left[2\|\lambda(f)\|_1 - \|\lambda(f)\|_2^2\right]. \tag{66}$$

Note that all inequalities in the derivation hold with equality if $f_1 = f_2 = f^*$, where f^* attains infimum above (or is a sequence of functions approaching this infimum). ■

The following Shannon-lower-bound-type lemma constrains the eigenvalues of an MSE matrix for estimating U from a 2^{nR}-level quantizer $f : \mathbb{R}^n \to [2^{nR}]$.

Lemma 2: Assume P has zero mean and unit variance. Let $f : \mathbb{R}^n \to [2^{nR}]$ be a 2^{nR}-level quantizer, and define $\lambda(f) = (\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ as in (59). Then

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\frac{1}{\lambda_i} \leq R + D(P\|\mathcal{N}(0,1)). \tag{67}$$

Proof. We may assume without loss of generality that $h(P) > -\infty$, as otherwise $D(P\|\mathcal{N}(0,1)) = \infty$ and the statement trivially holds. Let $e_U^f = U - \mathbb{E}[U|f(U)]$, $\Sigma_{e_U^f} = \mathbb{E}[e_U^f e_U^{f,\top}]$. Since the Gaussian distribution maximizes differential entropy under second moment constraints, we have that

$$h(U|f(U)) \leq \frac{1}{2}\log\det\left((2\pi e)\Sigma_{e_U^f}\right) = n\cdot\frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log(2\pi e\lambda_i). \tag{68}$$

Consequently,

$$nR \geq I(U; f(U)) = h(U) - h(U|f(U)) \geq h(U) - n \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log(2\pi e \lambda_i) \tag{69}$$

$$= h(\mathcal{N}^{\otimes n}(0,1)) - n \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log(2\pi e \lambda_i) + h(P^{\otimes n}) - h(\mathcal{N}^{\otimes n}(0,1)) \tag{70}$$

$$= n \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log \frac{1}{\lambda_i} - D(P \| \mathcal{N}(0,1)) \right), \tag{71}$$

which yields the claimed result. ■

Theorem 6: Assuming P has zero mean and unit variance, for any $n \geq 1$

$$D_n^{\mathrm{IP},*}(R, P) \geq \Gamma \left(R + D(P \| \mathcal{N}(0,1)) \right), \tag{72}$$

where $\Gamma(R)$ is defined in (1), and in particular

$$D^{\mathrm{IP}}(R, P) \geq \Gamma \left(R + D(P \| \mathcal{N}(0,1)) \right). \tag{73}$$

Proof. Let $f : \mathbb{R}^n \to [2^{nR}]$ be a 2^{nR}-level quantizer, and define $\lambda(f) = (\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ as in (59). Denote by $K = K_f$ the uniform distribution over (the multiset) $\lambda(f)$. By Lemma 1, we have that

$$D_n^{\mathrm{IP},*}(R, P) = \inf_f \mathbb{E}_{\lambda \sim K_f} \phi(\lambda) = \inf_f \mathbb{E}_{\lambda \sim K_f} \phi \left(2^{-2R_{\mathcal{N}}(\lambda)} \right), \tag{74}$$

where $R_{\mathcal{N}}(\lambda) = \frac{1}{2} \log \frac{1}{\lambda}$. Denote $\Gamma_1(R) = \phi(2^{-2R})$. In Appendix A we show that

$$\text{convex envelope of } \Gamma_1(R) = \Gamma(R). \tag{75}$$

It therefore follows that

$$D_n^{\mathrm{IP},*}(R, P) = \inf_f \mathbb{E}_{\lambda \sim K_f} \Gamma_1(R_{\mathcal{N}}(\lambda)) \tag{76}$$

$$\geq \inf_f \mathbb{E}_{\lambda \sim K_f} \Gamma \left(R_{\mathcal{N}}(\lambda) \right) \tag{77}$$

$$\geq \inf_f \Gamma \left(\mathbb{E}_{\lambda \sim K_f} R_{\mathcal{N}}(\lambda) \right) \tag{78}$$

$$\geq \Gamma \left(R + D(P \| \mathcal{N}(0,1)) \right), \tag{79}$$

where we have used Lemma 2 in the last inequality. ■

C. The Symmetric Gaussian case

Combining Theorem 5 and Theorem 6, we obtain a complete characterization for the Gaussian case.

Theorem 7:

$$D^{\mathrm{IP}}(R, \mathcal{N}(0,1)) = \Gamma(R) = \begin{cases} 1 - \left(1 - \phi(2^{-2R^*})\right) \frac{R}{R^*} & R < R^* \\ \phi(2^{-2R}) & R \geq R^* \end{cases}. \tag{80}$$

Proof. The upper bound follows from applying Theorem 5 with $\kappa = \min\{R/R^*, 1\}$, and recalling that $D_{\mathcal{N}(0,1)}(R) = 2^{-2R}$. The lower bound follows directly from Theorem 6. ■

IV. Compression for Matrix Multiplication

A. Setup

Let $A \in \mathbb{R}^{n \times a}$ be a matrix whose entries are drawn iid from the distribution P and $B \in \mathbb{R}^{n \times b}$ be a matrix, statistically independent of A, whose entries are drawn iid from the distribution Q. We assume both P and Q are

distributions with zero mean and unit variance. We consider the problem of quantizing A and B in order to compute their matrix multiplication $A^\top B$. In particular, an (n, a, b, R_1, R_2, D) code consists of mappings

$$f_1 : \mathbb{R}^{n\times a} \to [2^{naR_1}] \tag{81}$$
$$f_2 : \mathbb{R}^{n\times b} \to [2^{nbR_2}] \tag{82}$$
$$g : [2^{naR_1}] \times [2^{nbR_2}] \to \mathbb{R}^{a\times b}, \tag{83}$$

with

$$D = D^{\mathrm{MM}} = \frac{1}{n\cdot a\cdot b}\mathbb{E}\|A^\top B - g(f_1(A), f_2(B))\|_F^2. \tag{84}$$

We define

$$D^{\mathrm{MM},*}_{n,a,b}(R_1, R_2) = D^{\mathrm{MM},*}_{n,a,b}(R_1, R_2, P, Q) = \inf\left\{D \ : \ \exists(n, a, b, R_1, R_2, D) - \text{code}\right\}. \tag{85}$$

We further define the asymptotic function

$$D^{\mathrm{MM}}_{a,b}(R_1, R_2) = D^{\mathrm{MM}}_{a,b}(R_1, R_2, P, Q) = \limsup_{n\to\infty} D^*_{n,a,b}(R_1, R_2), \tag{86}$$

B. Basic Properties and Bounds

Denote $W_A = f_1(A)$ and $W_B = f_2(B)$ and further denote $\hat{A} = \mathbb{E}[A|W_A]$ and $\hat{B} = \mathbb{E}[B|W_B]$. Define $\Sigma_A = \mathbb{E}[(A-\hat{A})(A-\hat{A})^\top] \in \mathbb{R}^{n\times n}$ and $\bar{M}_A = \mathbb{E}[\hat{A}\hat{A}^\top] \in \mathbb{R}^{n\times n}$. Similarly, $\Sigma_B = \mathbb{E}[(B-\hat{B})(B-\hat{B})^\top] \in \mathbb{R}^{n\times n}$ and $\bar{M}_B = \mathbb{E}[\hat{B}\hat{B}^\top] \in \mathbb{R}^{n\times n}$. As in the scalar case, we still have the identities:

$$\Sigma_A + \bar{M}_A = aI_n \tag{87}$$
$$\Sigma_B + \bar{M}_B = bI_n. \tag{88}$$

The next theorem generalizes the basic bounds we derived above for the inner product case, to the matrix multiplication case. The proofs are similar to the statements above, and are therefore omitted.

Theorem 8: Assume P and Q have zero mean and unit variance. The following hold:

1) For fixed f_1, f_2, the optimal choice for g is $g^*(W_A, W_B) = \hat{A}^\top\hat{B}$, and the distortion is given by

$$\begin{aligned} D^{\mathrm{MM}} &= \frac{1}{n\cdot a\cdot b}\left[\mathrm{tr}(\Sigma_A\bar{M}_B) + \mathrm{tr}(\Sigma_B\bar{M}_A) + \mathrm{tr}(\Sigma_A\Sigma_B)\right] \\ &= \frac{1}{n}\left[\frac{1}{a}\mathrm{tr}(\Sigma_A) + \frac{1}{b}\mathrm{tr}(\Sigma_B) - \frac{1}{a\cdot b}\mathrm{tr}(\Sigma_A\Sigma_B)\right]. \end{aligned}$$

2) The oracle lower bound (taking $\hat{B} = B$ or $\hat{A} = A$) gives

$$D^{\mathrm{MM}} \geq \max\left\{\frac{1}{n\cdot a}\mathrm{tr}\,\Sigma_A, \frac{1}{n\cdot b}\mathrm{tr}\,\Sigma_B\right\},$$

and consequently for any $n \geq 1$

$$D^{\mathrm{MM},*}_{n,a,b}(R_1, R_2, P, Q) \geq \max\left\{D_P(R_1), D_Q(R_2)\right\},$$

and in particular

$$D^{\mathrm{MM}}_{a,b}(R_1, R_2, P, Q) \geq \max\left\{D_P(R_1), D_Q(R_2)\right\}.$$

3) For the symmetric case, where $R_1 = R_2 = R$ and $P = Q$, we have

$$D^{\mathrm{MM}}_{a,b}(R, P) \leq \min_{0\leq\kappa\leq 1}(1-\kappa) + \kappa\cdot\phi\left(D_P\left(\frac{R}{\kappa}\right)\right)$$

This is asymptotically attained by quantizing only the first κn coordinates of each column of A and each column of B.

4) For the symmetric case, where $R_1 = R_2 = R$ and $P = Q$, for any $n \geq 1$ we have

$$D^{\mathrm{MM},*}_{n,a,b}(R,P) \geq \frac{1}{n}\min\left\{\inf_{f_a}\left[2\|\lambda(f_a)\|_1 - \|\lambda(f_a)\|_2^2\right], \inf_{f_b}\left[2\|\lambda(f_b)\|_1 - \|\lambda(f_b)\|_2^2\right]\right\}$$
$$= \frac{1}{n}\min\left\{\inf_{f_a}\sum_{i=1}^{n}\phi\left(\lambda_i(f_a)\right), \inf_{f_b}\sum_{i=1}^{n}\phi\left(\lambda_i(f_b)\right)\right\}, \tag{89}$$

where the infima runs over all encoders $f_a : \mathbb{R}^{n\times a} \to [2^{naR}]$, $f_b : \mathbb{R}^{n\times b} \to [2^{nbR}]$, and

$$\lambda(f_a) = \mathrm{eig}\left(\frac{1}{a}\Sigma_{e_A^{f_a}}\right), \quad \lambda(f_b) = \mathrm{eig}\left(\frac{1}{b}\Sigma_{e_B^{f_b}}\right) \tag{90}$$

where $e_A^{f_a} = A - \mathbb{E}[A|f_a(A)]$, $\Sigma_{e_A^{f_a}} = \mathbb{E}[e_A^{f_a} e_A^{f_a,\top}]$, and $e_B^{f_b} = B - \mathbb{E}[B|f_b(B)]$, $\Sigma_{e_B^{f_b}} = \mathbb{E}[e_B^{f_b} e_B^{f_b,\top}]$.

C. Maximum Entropy Matrices

The fact that the Gaussian distribution maximizes the differential entropy of a vector, under second moment constraints, played a pivotal role in the derivation of our bounds for inner product quantization. For matrix multiplication quantization, the following lemma will play a similar role.

Lemma 3: Let $M \in \mathbb{R}^{n\times a}$ be a random matrix with $\mathbb{E}[M] = 0$, and $\mathbb{E}[MM^\top] = \Sigma$. Then

$$h(M) \leq \frac{a}{2}\log\det\left(2\pi e\frac{1}{a}\Sigma\right) = a\cdot\sum_{i=1}^{n}\frac{1}{2}\log(2\pi e\lambda_i), \tag{91}$$

where $\lambda = \mathrm{eig}\left(\frac{1}{a}\Sigma\right)$, and this is attained with equality if the columns of M are independent $\mathcal{N}\left(0, \frac{1}{a}\Sigma\right)$ random vectors.

Proof. Write $M = [m_1|m_2|\cdots|m_a]$, where $m_1, \ldots, m_a$ are zero-mean random vectors in $\mathbb{R}^n$. Denote the marginal distribution of m_i by P_i. Let $\Sigma_i = \mathbb{E}[m_i m_i^\top]$, and recall that

$$\Sigma = \mathbb{E}[MM^\top] = \sum_{i=1}^{a}\mathbb{E}[m_i m_i^\top] = \sum_{i=1}^{a}\Sigma_i. \tag{92}$$

We further have that

$$h(M) = h(m_1,\ldots,m_a) \leq \sum_{i=1}^{a}h(m_i) \leq a\cdot h\left(\frac{1}{a}\sum P_i\right), \tag{93}$$

where we have used sub-additivity and concavity of differential entropy in the inequalities above. Noting that the covariance matrix corresponding to the distribution $\frac{1}{a}\sum_{i=1}^{a}P_i$ is $\frac{1}{a}\sum_{i=1}^{a}\Sigma_i = \frac{1}{a}\Sigma$, we have

$$h(M) \leq a\cdot h\left(\mathcal{N}\left(0,\frac{1}{a}\Sigma\right)\right) = \frac{a}{2}\log\det\left(2\pi e\frac{1}{a}\Sigma\right). \tag{94}$$

All inequalities are attained with equality when $m_i \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{a}\Sigma\right)$, for $i = 1, \ldots, a$. ■

This immediately gives the following generalization of Lemma 2

Lemma 4: Assume the distribution P has zero mean and unit variance. Let $f_a : \mathbb{R}^{n\times a} \to [2^{naR}]$ be a 2^{naR}-level quantizer, and define $\lambda(f_a) = (\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ as in (90). Then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\frac{1}{\lambda_i} \leq R + D(P\|\mathcal{N}(0,1)). \tag{95}$$

Proof. Without loss of generality, we may assume $h(P) \geq -\infty$, as otherwise $D(P\|\mathcal{N}(0,1)) = \infty$ and the statement is trivial. Let $e_A^{f_a} = A - \mathbb{E}[A|f_a(A)]$, $\Sigma_{e_A^{f_a}} = \mathbb{E}[e_A^{f_a} e_A^{f_a,\top}]$. By Lemma 3, we have that

$$h(A|f_a(A)) \leq \frac{a}{2}\log\det\left((2\pi e)\frac{1}{a}\Sigma_{e_A^{f_a}}\right) = na\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log(2\pi e\lambda_i). \tag{96}$$

Consequently,

$$naR \geq I(A; f_a(A)) = h(A) - h(A|f_a(A)) \geq h(A) - na \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log(2\pi e \lambda_i) \tag{97}$$

$$= h(\mathcal{N}^{\otimes na}(0,1)) - na \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log(2\pi e \lambda_i) + h(P^{\otimes na}) - h(\mathcal{N}^{\otimes na}(0,1)) \tag{98}$$

$$= na\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log\frac{1}{\lambda_i} - D(P\|\mathcal{N}(0,1))\right), \tag{99}$$

which yields the claimed result. ■

D. Fundamental Limits

Using Theorem 8 and Lemma 4, we prove the following result for the symmetric matrix multiplication case.

Theorem 9: Assuming the distribution P has zero mean and unit variance, for any $n \geq 1$

$$D^{\mathrm{MM},*}_{n,a,b}(R,P) \geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right), \tag{100}$$

where $\Gamma(R)$ is defined in (1), and in particular

$$D^{\mathrm{MM}}_{a,b}(R,P) \geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right). \tag{101}$$

Proof. Let $f_a : \mathbb{R}^{n\times a} \to [2^{naR}]$ be a 2^{naR}-level quantizer, and define $\lambda(f_a) = (\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ as in (90). Denote by $K = K_{f_a}$ the uniform distribution over (the multiset) $\lambda(f_a)$, and $R_{\mathcal{N}}(\lambda) = \frac{1}{2}\log\frac{1}{\lambda}$, as in the proof of Theorem 6, we have that

$$\mathbb{E}_{\lambda\sim K_{f_a}}\phi(\lambda) = \mathbb{E}_{\lambda\sim\Gamma_{f_a}}\phi\left(2^{-2R_{\mathcal{N}}(\lambda)}\right). \tag{102}$$

Recalling from the proof of Theorem 6 that the $\Gamma_1(R) = \phi(2^{-2R}) \geq \Gamma(R)$ and the function $R \mapsto \Gamma(R)$ is convex and non-increasing, it follows that

$$\mathbb{E}_{\lambda\sim K_{f_a}}\phi\left(2^{-2R_{\mathcal{N}}(\lambda)}\right) \geq \mathbb{E}_{\lambda\sim K_f}\Gamma\left(R_{\mathcal{N}}(\lambda)\right) \tag{103}$$

$$\geq \Gamma\left(\mathbb{E}_{\lambda\sim K_f} R_{\mathcal{N}}(\lambda)\right) \tag{104}$$

$$\geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right), \tag{105}$$

where we have used Lemma 4 in the last inequality. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\lambda_i(f_a)) = \mathbb{E}_{\lambda\sim K_{f_a}}\phi(\lambda) \geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right). \tag{106}$$

Similarly, for any $f_b : \mathbb{R}^{n\times b} \to \left[2^{nbR}\right]$ we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\lambda_i(f_b)) = \mathbb{E}_{\lambda\sim K_{f_b}}\phi(\lambda) \geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right). \tag{107}$$

Thus, by (89) in Theorem 8, for any $n \geq 1$

$$D^{\mathrm{MM}}_{n,a,b}(R,P) \geq \min\left\{\min_{f_a}\frac{1}{n}\sum_{i=1}^{n}\phi(\lambda_i(f_a)), \min_{f_b}\frac{1}{n}\sum_{i=1}^{n}\phi(\lambda_i(f_b))\right\} \tag{108}$$

$$\geq \Gamma\left(R + D(P\|\mathcal{N}(0,1))\right), \tag{109}$$

as claimed. ■

Proof of Theorem 2. Part 1 follows immediately from Theorem 9. Part 2 follows from part 2 of Theorem 8, and recalling that $D_{\mathcal{N}(0,1)}(R) = 2^{-2R}$. ■

E. The Symmetric Gaussian case

Combining Theorem 8 and Theorem 9, we obtain a complete characterization for the Gaussian case.

Theorem 10:

$$D_{a,b}^{\mathrm{MM}}(R,\mathcal{N}(0,1)) = \Gamma(R). \tag{110}$$

Proof. The upper bound follows applying Part 3 of Theorem 8 with $\kappa = \min\{R/R^*, 1\}$, and recalling that $D_{\mathcal{N}(0,1)}(R) = 2^{-2R}$. The lower bound follows directly from Theorem 9. ■

V. Lattice Quantization Scheme for Matrix Multiplication of Arbitrary Matrices

Our theoretical analysis in Sections II -IV assumed the entries in the vectors/matrices to be multiplied are drawn iid from some known distribution. In this section, we drop this assumption, and, building on the observations from the analysis above, develop a robust scheme for compression for matrix multiplication. Our scheme is designed to attain the optimal distortion in the case where A and B have iid Gaussian entries, but the error it attains for arbitrary matrices can also be upper bounded.

We first develop encoders $f_1, f_2 : \mathbb{R}^n \to [2^{nR}]$ and a decoder $g : [2^{nR}] \times [2^{nR}] \to \mathbb{R}$ for estimating the inner product of $U, V \in \sqrt{n}\mathbb{S}^{n-1}$ where $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ is the unit sphere. We then show how these encoders and decoder can be leveraged for compression for matrix multiplication. Let $\mathrm{O}_n(\mathbb{R})$ be the orthogonal group, consisting of all orthonormal matrices in $\mathbb{R}^{n\times n}$. It will be useful to analyze the performance of f_1, f_2, g with respect to the following distribution on U, V.

Definition 1 (ρ-correlated spherically uniform random vectors): Let $S = [S_1|S_2|\cdots|S_n] \sim \mathrm{Uniform}(\mathrm{O}_n(\mathbb{R}))$ be a random matrix uniformly distributed over the group of orthogonal matrices in $\mathbb{R}^{n\times n}$ (that is, S is drawn from the Haar measure on $\mathrm{O}_n(\mathbb{R})$). We say that the random vectors $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ are ρ-correlated spherically uniform random vectors if $U = \sqrt{n}S_1$, $Z = \sqrt{n}S_2$ and

$$V = \rho U + \sqrt{1-\rho^2}Z. \tag{111}$$

Theorem 11: For any $\varepsilon > 0$, $0 < \kappa \le 1$ and sufficiently large n, there exist randomized encoders $f_1, f_2 : \mathbb{R}^n \to [2^{nR}]$ and decoders $g : [2^{nR}] \times [2^{nR}] \to \mathbb{R}$, and $g_{1-\mathrm{sided}} : [2^{nR}] \times \mathbb{R}^n \to \mathbb{R}$, such that if U, V are ρ-correlated spherically uniform

1) for every $-1 \le \rho \le 1$ and $0 \le \alpha \le 1$

$$\frac{1}{n}\mathbb{E}(U^\top V - \alpha g(f_1(U), f_2(V))^2 < \rho^2 n\left(1-\kappa\alpha\right)^2 + \kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{R}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{R}{\kappa}}\right)}\right) + \varepsilon(1+\rho^2 n), \tag{112}$$

   where $\phi(t) = 2t - t^2$.

2) for every $-1 \le \rho \le 1$ and $0 \le \alpha \le 1$

$$\frac{1}{n}\mathbb{E}(U^\top V - \alpha g_{1-\mathrm{sided}}(f_1(U), V))^2 < \rho^2 n\left(1-\kappa\alpha\right)^2 + \kappa\alpha^2\left(1-\kappa+\frac{1}{2^{2\frac{R}{\kappa}}-1}\right) + \varepsilon(1+\rho^2 n). \tag{113}$$

The proof is based on dithered nested lattice quantization, and is brought in Section VI.

Remark 2: The randomization required by the encoders and decoders above is in the form of *dithering*, as will become clear in the proof of Theorem 11. For the special case of $\kappa = \alpha = 1$, the MSE does not involve ρ, and therefore, there must exist fixed (deterministic) values for the dither vectors which attain the same MSE as above, or smaller. We believe that there also exist fixed values for the dithers, for which Theorem 11 holds, and that randomness is not required at all here. The technical challenge is that when $\kappa, \alpha \neq 1$ the MSE is bounded as a weighted sum of expectations, where the weights depend on ρ and α. Thus, showing the existence of "good" fixed dithers, requires showing that there are dither values for which all involved expectations are small. This requires establishing some (very weak form of) of concentration of involved random variables, and is left for future work. Note that the assumptions of Theorem 11 are that U, V are random vectors on the sphere. This assumption is enforced in the next Theorem by applying a random rotation on arbitrary matrices. While random dithering is likely to not be needed, we do believe that the random rotation is necessary for Theorem 12 to hold.

Remark 3: Our proof of Theorem 11 is based on using dithered "good" nested lattice quantizers. With these quantizers, the quantization noise is statistically independent of the input and is approximately white, a property

that is crucial for the analysis. Using unstructured random coding techniques for proving Theorem 11 does not seem straightforward. The results from [47] may be a good starting point for pursuing this direction.

Equipped with Theorem 11, we can now easily prove Theorem 12. Recall that for a column vector $x \in \mathbb{R}^n$ we denote by $\bar{x} = x - \left(\frac{1}{n}\mathbf{1}^\top x\right)\mathbf{1}$ its zero-centered version. For a matrix $A = [a_1|\cdots|a_a] \in \mathbb{R}^{n\times a}$ we denote $\bar{A} = [\bar{a}_1|\ldots|\bar{a}_a]$.

Theorem 12: For any $\varepsilon > 0$, $0 < \kappa \leq 1$ and sufficiently large n, there exist randomized encoders $f_1 : \mathbb{R}^{n\times a} \to [2^{naR}]$, $f_2 : \mathbb{R}^{n\times b} \to [2^{nbR}]$, and decoders $g : [0,1]\times[2^{naR}]\times[2^{nbR}] \to \mathbb{R}^{a\times b}$ and $g_{1-\text{sided}} : [0,1]\times[2^{naR}]\times\mathbb{R}^{n\times b} \to \mathbb{R}^{a\times b}$ such that for any $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ with bounded entries we have

1) Let $C = A^\top B$, $\tilde{C} = \bar{A}^\top\bar{B}$, and $\hat{C} = g(\alpha, f_1(A), f_2(B))$ for $0 < \alpha \leq 1$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \leq \tilde{C}_{i,j}^2 \cdot \left((1-\kappa\alpha)^2 + \varepsilon\right) + \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(\kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{R}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{R}{\kappa}}\right)}\right) + \varepsilon\right) + n^{-8}, \tag{114}$$

and in particular

$$\mathbb{E}\|A^\top B - g(\alpha, f_1(A), f_2(B))\|_F^2 < \|\bar{A}^\top\bar{B}\|_F^2 \cdot \left((1-\kappa\alpha)^2 + \varepsilon\right) + \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(\kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{R}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{R}{\kappa}}\right)}\right) + \varepsilon\right) + a\cdot b\cdot n^{-8}. \tag{115}$$

2) Let $C = A^\top B$, $\tilde{C} = \bar{A}^\top\bar{B}$, and $\hat{C} = g_{1-\text{sided}}(\alpha, f_1(A), B)$ for $0 < \alpha \leq 1$. Then, for any $i \in [a], j \in [b]$ we have

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \leq \tilde{C}_{i,j}^2 \cdot \left((1-\kappa\alpha)^2 + \varepsilon\right) + \frac{\|\bar{a}_i\|^2\|\bar{b}_j\|^2}{n}\left(\kappa\alpha^2\left(1-\kappa+\frac{1}{2^{2\frac{R}{\kappa}}-1}\right) + \varepsilon\right) + n^{-8}. \tag{116}$$

and in particular

$$\mathbb{E}\|A^\top B - g_{1-\text{sided}}(\alpha, f_1(A), B)\|_F^2 < \|\bar{A}^\top\bar{B}\|_F^2 \cdot \left((1-\kappa\alpha)^2 + \varepsilon\right) + \frac{\|\bar{A}\|_F^2\|\bar{B}\|_F^2}{n}\left(\kappa\alpha^2\left(1-\kappa+\frac{1}{2^{2\frac{R}{\kappa}}-1}\right) + \varepsilon\right) + a\cdot b\cdot n^{-8}. \tag{117}$$

Proof of Theorem 12. We only prove part 1. The proofs for part 2 is nearly identical, and we specify the required modifications in the end of the proof.

Recall that $M = n^{10}2^{2000}$ and let $\delta = M^{-5}$. Let f_1, f_2, g be the encoders and decoder from Theorem 11. Based on those f_1, f_2, g, we propose the following rate-R quantization scheme for quantization of matrices $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$ in order to estimate $C = A^\top B$:

1) Let $\mu_{a_i} = \frac{1}{n}\mathbf{1}^\top a_i$ for $i \in [a]$ (similarly, $\mu_{b_j} = \frac{1}{n}\mathbf{1}^\top b_j$ for $j \in [b]$). Since the matrices A and B have bounded entries, we have that $\mu_{a_i} \in [-M, M]$, $\forall i \in [a]$, and similarly for μ_{b_j}, $j \in [b]$. For each $i \in [a]$ we quantize μ_{a_i} to the nearest point in $\{k\cdot 2\delta\}_{k=-M/(2\delta)}^{k=M/(2\delta)}$, such that the quantized value $\hat{\mu}_{a_i}$ satisfies $\hat{\mu}_{a_i} = \mu_{a_i} \pm \delta$. This requires a total of $a\log(M/\delta)$ bits. Similarly, we quantize μ_{b_j} to $\hat{\mu}_{b_j}$ for each $j \in [b]$, which requires a total of $b\log(M/\delta)$ bits.
2) Let $\bar{a}_i = a_i - \mu_{a_i}\mathbf{1}$ for $i \in [a]$ (similarly, $\bar{b}_j = b_j - \mu_{b_j}\mathbf{1}$ for $j \in [b]$). Since the matrices A and B have bounded entries, we have that $\|\bar{a}_i\| \leq \|a_i\| \leq \sqrt{n}M$, $\forall i \in [a]$, and similarly for $\|\bar{b}_j\|$, $j \in [b]$. We quantize the each $\|\bar{a}_i\|$, $i \in [a]$, to the nearest point in the grid $\{0\}\cup\{M^{-4}(1+\delta)^k\}_{k=0}^T$, where $T = \frac{\lceil\log(\sqrt{n}M^5)\rceil}{\log(1+\delta)}$. This requires a total of $a\log(T+2) < a\left(\log(3+\log(\sqrt{n}M^5)) - \log\log(1+\delta)\right)$ bits. Note that if $\|\bar{a}_i\| \in \{0\}\cup[M^{-4}, \sqrt{n}M]$ we have that $\widehat{\|\bar{a}_i\|} = \|\bar{a}_i\|(1\pm\delta)$, and if $0 < \|\bar{a}_i\| < M^{-4}$, then $|\|\bar{a}_i\| - \widehat{\|\bar{a}_i\|}| < M^{-4}$. We quantize each $\|\bar{b}_j\|$, $j \in [b]$ to $\widehat{\|\bar{b}_j\|}$, in a similar manner, requiring a total of $b\log(T+2) < b\left(\log(3+\log(\sqrt{n}M^5)) - \log\log(1+\delta)\right)$ bits.
3) Draw $S \sim \text{Uniform}(\text{O}_n(\mathbb{R}))$ at both encoders (using common randomness), and compute $\tilde{A} = [\tilde{a}_1|\cdots|\tilde{a}_a] = S\bar{A}$, and $\tilde{B} = [\tilde{b}_1|\cdots|\tilde{b}_b] = S\bar{B}$, where $\bar{A} = [\bar{a}_1|\cdots|\bar{a}_a]$ and $\bar{B} = [\bar{b}_1|\cdots|\bar{b}_b]$.

4) Let

$$U_i = \sqrt{n}\frac{\tilde{a}_i}{\|\bar{a}_i\|} = \sqrt{n}S\frac{\bar{a}_i}{\|\bar{a}_i\|}, \quad i = 1, \ldots, a \tag{118}$$

$$V_j = \sqrt{n}\frac{\tilde{b}_j}{\|\bar{b}_i\|} = \sqrt{n}S\frac{\bar{b}_j}{\|\bar{b}_j\|}, \quad j = 1, \ldots, b. \tag{119}$$

Let

$$\varepsilon_0 = \frac{1}{n}\left[\log(M/\delta) + \log(3 + \log(\sqrt{n}M^5)) - \log\log(1+\delta)\right] \tag{120}$$

and note that ε_0 can be made arbitrarily small for n large enough. Apply $f_1 : \mathbb{R}^n \to [2^{n(R-\varepsilon_0)}]$ on U_i, for $i = 1, \ldots, a$, and $f_2 : \mathbb{R}^n \to [2^{n(R-\varepsilon_0)}]$ on V_j, for $j = 1, \ldots, b$.

5) Use $g : [2^{n(R-\varepsilon_0)}] \times [2^{n(R-\varepsilon_0)}] \to \mathbb{R}$, to estimate each entry of $C = A^\top B$ as

$$\hat{C}_{ij} = \alpha\frac{\widehat{\|\bar{a}_i\|}\ \widehat{\|\bar{b}_j\|}}{n}g(f_1(U_i), f_2(V_j)) + n\hat{\mu}_{a_i}\hat{\mu}_{b_j}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b. \tag{121}$$

To analyze the mean squared error $\mathbb{E}(C_{ij} - \hat{C}_{i,j})^2$, first note that

$$\bar{a}_i^\top \bar{b}_j = (a_i - \mu_{a_i}\mathbf{1})^\top \left(b_j - \mu_{b_j}\mathbf{1}\right) = a_i^\top b_j - n\mu_{a_i}\mu_{b_j}, \tag{122}$$

so that

$$\begin{aligned} C_{ij} = a_i^\top b_j &= \bar{a}_i^\top \bar{b}_j + n\mu_{a_i}\mu_{b_j} = \bar{a}_i^\top S^\top S\bar{b}_j + n\mu_{a_i}\mu_{b_j} = \tilde{a}_i^\top \tilde{b}_j + n\mu_{a_i}\mu_{b_j} \\ &= \frac{\|\bar{a}_i\|\ \|\bar{b}_j\|}{n}U_i^\top V_j + n\mu_{a_i}\mu_{b_j}. \end{aligned} \tag{123}$$

We therefore have that

$$C_{ij} - \hat{C}_{ij} = \frac{\|\bar{a}_i\|\ \|\bar{b}_j\|}{n}U_i^\top V_j - \alpha\frac{\widehat{\|\bar{a}_i\|}\ \widehat{\|\bar{b}_j\|}}{n}g(f_1(U_i), f_2(V_j)) + n\mu_{a_i}\mu_{b_j} - n\hat{\mu}_{a_i}\hat{\mu}_{b_j} \tag{124}$$

$$= e_{ij} + \Delta, \tag{125}$$

where

$$e_{ij} = \frac{\|\bar{a}_i\|\ \|\bar{b}_j\|}{n}\left(U_i^\top V_j - \alpha g(f_1(U_i), f_2(V_j))\right) \tag{126}$$

and $\Delta = \Delta_1 + \Delta_2$, where

$$\Delta_1 = n\mu_{a_i}\mu_{b_j} - n\hat{\mu}_{a_i}\hat{\mu}_{b_j} \tag{127}$$

and

$$\Delta_2 = \frac{\alpha}{n}g(f_1(U_i), f_2(V_j))\left(\widehat{\|\bar{a}_i\|}\ \widehat{\|\bar{b}_j\|} - \|\bar{a}_i\|\|\bar{b}_j\|\right). \tag{128}$$

We have that

$$|\Delta_1| \le n\delta(|\mu_a| + |\mu_b|) + n\delta^2 \le 3nM^{-4}. \tag{129}$$

To upper bound $|\Delta_2|$, first note that without loss of generality we can assume $|\alpha g(f_1(U_i), f_2(V_j)| \le n$ because the quantity $U_i^\top V_j$ it estimates is in $[-n, n]$. Furthermore,

$$\left|\widehat{\|\bar{a}_i\|}\ \widehat{\|\bar{b}_j\|} - \|\bar{a}_i\|\|\bar{b}_j\|\right| \le \begin{cases} \|\bar{a}_i\| \cdot \|\bar{b}_j\| \cdot 3\delta & \|\bar{a}_i\|, \|\bar{b}_j\| \in \{0\} \cup [M^{-4}, \sqrt{n}M] \\ M^{-8} & \|\bar{a}_i\|, \|\bar{b}_j\| < M^{-4} \\ 2\sqrt{n}M^{-3} & \text{otherwise} \end{cases}. \tag{130}$$

Thus (for $n \ge 4$),

$$|\Delta_2| \le 3M^{-5}\|\bar{a}_i\| \cdot \|\bar{b}_j\| + 2\sqrt{n}M^{-3} \le 4nM^{-3}. \tag{131}$$

We consequently have that

$$|\Delta| < \varepsilon_1 = 7nM^{-3}, \tag{132}$$

with probability 1. We have therefore obtained

$$\mathbb{E}(C_{ij}-\hat{C}_{ij})^2 \leq \mathbb{E}(e_{ij}^2)+\varepsilon_1^2+2\varepsilon_1\mathbb{E}|e_{ij}| \leq \mathbb{E}(e_{ij}^2)+\varepsilon_1^2+4\varepsilon_1\|\bar{a}_i\|\cdot\|\bar{b}_j\|, \tag{133}$$

where in the last inequality we have used the fact that both $|U_i^\top V_j| \leq n$ and $|\alpha g(f_1(U_i), f_2(V_j))| \leq n$.

We are therefore left with the task of upper bounding $\mathbb{E}(e_{ij}^2)$. To that end, let $\rho_{ij} = \frac{\bar{a}_i^\top \bar{b}_j}{\|\bar{a}_i\|\ \|\bar{b}_j\|}$. We claim that U_i, V_j are ρ_{ij}-correlated spherically uniform random vectors. To see this, note that due to the random rotation matrix S, we may assume without loss of generality that

$$\frac{\bar{a}_i}{\|\bar{a}_i\|} = [1|0|0|\cdots|0]^\top, \tag{134}$$

$$\frac{\bar{b}_j}{\|\bar{b}_j\|} = [\rho_{ij}|\sqrt{1-\rho_{ij}^2}|0|\cdots|0]^\top, \tag{135}$$

and this assumption will have no affect on the joint distribution of U_i, V_j. Writing $S = [S_1|S_2|\cdots|S_n]$, we therefore have that $U_i = \sqrt{n}S_1$ and $V_j = \rho_{ij}U_i + \sqrt{1-\rho_{ij}^2}Z$, with $Z = \sqrt{n}S_2$. Thus, if f_1, f_2, g are the encoders and decoder from Theorem 11, we therefore have that for any $\varepsilon' > 0$ and n large enough

$$\frac{1}{n}\mathbb{E}(U_i^\top V_j - \alpha g(f_1(U_i), f_2(V_j)))^2 < \rho_{ij}^2 n\left(1-\kappa\alpha\right)^2 + \kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}\right) + \varepsilon'(1+\rho_{ij}^2 n). \tag{136}$$

Consequently,

$$\begin{aligned}\mathbb{E}(e_{ij}^2) &= \mathbb{E}\left(\frac{\|\bar{a}_i\|\ \|\bar{b}_j\|}{n}\left(U_i^\top V_j - \alpha g(f_1(U_i), f_2(V_j))\right)\right)^2 \\ &< \frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\rho_{ij}^2 n\left(1-\kappa\alpha\right)^2 + \kappa\alpha^2\frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}\right) + \frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\varepsilon'(1+\rho_{ij}^2 n) \qquad (137)\\ &= (\bar{a}_i^\top \bar{v}_j)^2\left((1-\kappa\alpha)^2+\varepsilon'\right) + \frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\left(\kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}\right)+\varepsilon'\right). \qquad (138)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}(C_{ij}-\hat{C}_{ij})^2 \leq (\bar{a}_i^\top \bar{v}_j)^2\left((1-\kappa\alpha)^2+\varepsilon'\right) + \frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\left(\kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{(R-\varepsilon_0)}{\kappa}}\right)}\right)+\varepsilon'\right)\\ + 4\|\bar{a}_i\|\ \|\bar{b}_j\|\varepsilon_1 + \varepsilon_1^2. \qquad (139)\end{aligned}$$

Thus, recalling that $\|\bar{a}_i\|\ \|\bar{b}_j\| \leq nM^2$, for any $\varepsilon > 0$ and n large enough

$$\mathbb{E}(C_{ij}-\hat{C}_{ij})^2 \leq (\bar{a}_i^\top \bar{v}_j)^2\left((1-\kappa\alpha)^2+\varepsilon\right) + \frac{\|\bar{a}_i\|^2\ \|\bar{b}_j\|^2}{n}\left(\kappa\alpha^2\left(\frac{1-\kappa+\kappa\phi\left(2^{-2\frac{R}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{R}{\kappa}}\right)}\right)+\varepsilon\right) + n^{-8} \tag{140}$$

The proof of part 1 is complete, by noting that $\tilde{C}_{ij} = \bar{a}_i^\top \bar{b}_j$ and that

$$\begin{aligned}\frac{1}{n}\sum_{i,j}\|\bar{a}_i\|^2\|\bar{b}_j\|^2 = \frac{1}{n}\sum_{i=1}^{a}\|\bar{a}_i\|^2\sum_{j=1}^{b}\|\bar{b}_j\|^2 = \frac{\|\bar{A}\|_F^2\ \|\bar{B}\|_F^2}{n},\\ \sum_{i,j}\tilde{C}_{ij}^2 = \|\tilde{C}\|_F^2. \qquad (141)\end{aligned}$$

The proof for part 2 follows identically from part 2 of Theorem 11. ■

With Theorem 12 at hand, we easily obtain Theorem 1 and Theorem 3 as simple corollaries.

Proof of Theorem 1. For part 1, let $\alpha = \alpha_\kappa = \left(1 - \phi\left(2^{-2\frac{R}{\kappa}}\right)\right)$. Applying part 1 of Theorem 12 gives

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \le \tilde{C}_{i,j}^2 \cdot \left(G^2(R,\kappa) + \varepsilon\right) + \frac{\|\bar{a}_i\|^2 \|\bar{b}_j\|^2}{n}\left((1 - G(R,\kappa))G(R,\kappa) + \varepsilon\right) + n^{-8}, \tag{142}$$

where

$$G(R,\kappa) = 1 - \kappa + \kappa\phi\left(2^{-2\frac{R}{\kappa}}\right). \tag{143}$$

Choosing $\kappa = \min\{R/R^*, 1\}$, we get $G(R,\kappa) = \Gamma(R)$, establishing the claim.

For part 2, let $\alpha = \alpha_\kappa = 1 - 2^{-2\frac{R}{\kappa}} = \frac{2^{2\frac{R}{\kappa}} - 1}{2^{2\frac{R}{\kappa}}}$. Applying part 2 of Theorem 12 gives

$$\mathbb{E}(C_{i,j} - \hat{C}_{i,j})^2 \le \tilde{C}_{i,j}^2 \cdot \left(\tilde{G}^2(R,\kappa) + \varepsilon\right) + \frac{\|\bar{a}_i\|^2 \|\bar{b}_j\|^2}{n}\left((1 - \tilde{G}(R,\kappa))\tilde{G}(R,\kappa) + \varepsilon\right) + n^{-8}, \tag{144}$$

where

$$\tilde{G}(R,\kappa) = 1 - \kappa + \kappa 2^{-2\frac{R}{\kappa}}. \tag{145}$$

Choosing $\kappa = 1$, we get $\tilde{G}(R,\kappa) = 2^{-2R}$, establishing the claim. ■

Proof of Theorem 3. Follows by applying part 1 and part 2 of Theorem 12 with $\alpha = \kappa = 1$ (that is, no time-sharing and no MMSE scaling). Note that a straightforward application of Theorem 12 with $\alpha = \kappa = 1$ leaves an $\|\bar{A}^\top \bar{B}\|_F^2 \cdot \varepsilon$ term. However, a careful inspection of the proof of Theorem 11 shows that this term is not needed for the special case of $\alpha = \kappa = 1$. ■

VI. Nested Lattice Quantization for Inner Product Computation

A. Lattices

We review some basic lattice definitions. See [14] for a comprehensive treatment of lattices in information theory. For a lattice $L \subset \mathbb{R}^d$ we define the nearest neighbor quantizer $Q_L : \mathbb{R}^d \to L$ as

$$Q_L(x) = \operatorname*{argmin}_{\lambda \in L} \|x - \lambda\|, \tag{146}$$

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

$$\mathcal{V}_L = \left\{x \in \mathbb{R}^d \ : \ Q_L(x) = 0\right\}. \tag{147}$$

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 $\mathrm{covol}(L)$, is the volume of its Voronoi region (or any other fundamental cell of L), which is also equal to $|\det G|$. Let $\mathcal{B} = \{x \in \mathbb{R}^d \ : \ \|x\| \le 1\}$ be the unit ℓ_2 ball in $\mathbb{R}^d$, whose volume is

$$V_d = \frac{\pi^{d/2}}{\Gamma\left(1 + \frac{d}{2}\right)}, \tag{148}$$

where here Γ is Euler's Gamma function, not to be confused with $\Gamma(R)$ defined in (1). We denote by $r_{\mathrm{eff}}(L) = (\mathrm{covol}(L)/V_d)^{\frac{1}{d}}$ the effective radius of L, that is, the radius of a ℓ_2 ball in $\mathbb{R}^d$ whose volume $V_d r_{\mathrm{eff}}^d(L)$ equals $\mathrm{covol}(L)$. The covering radius of L is defined as

$$r_{\mathrm{cov}}(L) = \min\left\{r > 0 \ : \ L + r\mathcal{B} = \mathbb{R}^d\right\} = \max\left\{\|x\| \ : \ x \in \mathcal{V}_L\right\}. \tag{149}$$

Clearly, $r_{\mathrm{eff}}(L) \le r_{\mathrm{cov}}(L)$. Let $Z \sim \mathrm{Uniform}(\mathcal{V}_L)$ be a random vector uniformly distributed over the Voronoi region of L. We define the second moment of the lattice L as

$$\sigma^2(L) = \frac{1}{d}\mathbb{E}\|Z\|^2, \tag{150}$$

and the covariance matrix of L as

$$R(L) = \mathbb{E}[ZZ^\top]. \tag{151}$$

The modulo operation with respect to the lattice L, is defined in this paper as

$$[x] \bmod L = x - Q_L(x). \tag{152}$$

Note that $[x] \bmod L \in \mathcal{V}_L$.

The proof of Theorem 11 uses a nested-lattice quantizer [14], based on a pair of nested lattices $\Lambda_c \subset \Lambda_f$ in $\mathbb{R}^d$. A quantizer is constructed from such a pair by first quantizing each point in $\mathbb{R}^d$ to $Q_{\Lambda_f}(x)$, the nearest point in the lattice Λ_f. Since there is an infinite number of points in Λ_f, the encoder cannot describe $Q_{\Lambda_f}(x)$ using dR bits. Instead, it describes the *coset* of Λ_c in which $Q_{\Lambda_f}(x)$ lies. There are $|\Lambda_f/\Lambda_c| = \mathrm{covol}(\Lambda_c)/\mathrm{covol}(\Lambda_f)$ such cosets, and therefore, if $|\Lambda_f/\Lambda_c| \le 2^{dR}$ the encoder can indeed send that information with dR bits. When the decoder gets this information, it knows that $Q_{\Lambda_f}(x) \in Q_{\Lambda_f}(x) + \Lambda_c$, but does not know which point within this coset is $Q_{\Lambda_f}(x)$. Typically, the decoder will output the most likely member from the coset. For the case where X is an iid Gaussian vector in $\mathbb{R}^d$, this (approximately) corresponds to selecting $\hat{x}$ as the member with the smallest energy in $Q_{\Lambda_f}(x) + \Lambda_c$. Under this paradigm, the reconstruction points are $\Lambda_f \cap \mathcal{V}_{\Lambda_c}$.

Many works have established the existence of nested lattices that simultaneously posses many desired properties, namely, relatively large packing radius, small covering radius, small second moment, and resilience to noise [13], [48], [49], [50], [51], [52], [14], [53], [54], [15], [55], [56], [16], [57], [58], [59]. In the problem of quantization for inner product computation a new ingredient enters the picture, that was not previously needed. The inner product reconstruction error includes a term that consists of the inner product of the quantization errors of each one of the vectors. To control this term, we need the variance of the inner product $Z^\top \bar{Z}$ between two independent dither vectors $Z, \bar{Z} \sim \mathrm{Uniform}(\mathcal{V}_L)$ to be small. This in turn, requires the spectrum of the quantization error to be "nearly-white" in the Frobenius norm sense. Namely, a good lattice quantizer L for the inner product problem needs to satisfy $\frac{1}{d}\|R(L)\|_F^2 \approx (\sigma^2(L))^2$. While the optimal lattice quantizer in dimension d always satisfies $R(L) = \sigma^2(L) \cdot I_d$ [14], [60], we do not know whether it also has the additional required properties, e.g., resilience to noise. We must therefore resort to analyzing a random ensemble of nested lattices, and show that in addition to all other required properties, they also typically have small $\frac{1}{d}\|R(L)\|_F^2$. Our proof that a random lattice has small $\frac{1}{d}\|R(L)\|_F^2$ relies on the fact that the covering density of a random lattice is only polynomial in the dimension, which was recently proved in [16]. We prove the following result in Appendix B. Except for item 4 which required new ideas, the proofs for all other items follow the techniques developed in the papers on lattice goodness that were cited above.

Theorem 13: There are universal constants C_1, C_2 such that for any distribution P_U on $\mathbb{R}^d$, any $r_U > 0$, $D > 0$, $\alpha, \beta > 0$, $0 < \varepsilon \le \frac{1}{\sqrt{2}}$, and

$$R \ge \frac{1}{2}\log\left(\beta^2 + \alpha^2 \frac{r_U}{D}\right) + C_1\left(\varepsilon + \frac{\log d}{d}\right) \tag{153}$$

there exists a pair of nested lattices $\Lambda_c \subset \Lambda_f$ in $\mathbb{R}^d$ satisfying the following

1) $\frac{1}{2}2^{dR} \le |\Lambda_f/\Lambda_c| \le 2^{dR}$
2) $r_{\mathrm{cov}}(\Lambda_f) \le \sqrt{dD}$ and $r_{\mathrm{cov}}(\Lambda_c) \le 2^R\sqrt{dD}$;
3) $\sigma^2(\Lambda_f) \le D$;
4) $\frac{1}{d}\|R(\Lambda_f)\|_F^2 \le D^2(1 + \frac{C_2 \log^3 d}{d})$;
5) For $U \sim P_U$, $Z \sim \mathrm{Uniform}(\mathcal{V}_{\Lambda_f})$, $Z \perp\!\!\!\perp U$, we have

$$\Pr(\alpha U + \beta Z \notin \mathcal{V}_{\Lambda_c}) \le \Pr(\|U\|^2 > d \cdot r_U) + 6e^{-d\frac{\varepsilon^2}{2}}. \tag{154}$$

Remark 4: If we further require that $R = \log q$ for some integer $q \ge 2$, there exists a pair of self-similar nested lattices $\Lambda_c = q\Lambda \subset \Lambda = \Lambda_f$ satisfying the statements in Theorem 13. The proof is essentially the same.

Remark 5: While our proof for Theorem 13 does not impose any particular structure on the lattices $\Lambda_c \subset \Lambda_f$, it is possible to prove the existence of Construction A lattices $\Lambda_c \subset \Lambda_f$ satisfying Theorem 13. This follows from [57, Corollary 1.5] that shows that the covering density of a typical Construction A lattice (with judiciously chosen parameters) is also polynomial in the dimension.

B. Proof of Theorem 11

1) Dithered Nested Lattice Quantization for Inner Product: Let $d = \lfloor \kappa n \rfloor$, and denote $U_{[d]} = (U_1, \ldots, U_d)^\top$ and similarly $V_{[d]} = \sqrt{\rho}U_{[d]} + \sqrt{1-\rho^2}Z_{[d]}$. Let $\tilde{R} = \frac{n}{d}R \ge \frac{R}{\kappa}$, and let $\Lambda_c \subset \Lambda_f$ be a pair of nested lattices in

$\mathbb{R}^d$, with $|\Lambda_f/\Lambda_c| \le 2^{d\tilde{R}}$. Let $\tilde{Z}_1, \tilde{Z}_2 \sim \mathrm{Uniform}(\mathcal{V}_{\Lambda_f})$ be statistically independent dither vectors. Our encoders $f_1, f_2 : \mathbb{R}^n \to [2^{nR}]$ compute

$$\tilde{U}_{[d]} = \left[Q_{\Lambda_f}\left(U_{[d]} + \tilde{Z}_1\right)\right] \bmod \Lambda_c \tag{155}$$

$$\tilde{V}_{[d]} = \left[Q_{\Lambda_f}\left(V_{[d]} + \tilde{Z}_2\right)\right] \bmod \Lambda_c, \tag{156}$$

and each of them maps the result to $nR = d\tilde{R}$ bits (which is possible since $|\Lambda_f/\Lambda_c| \le 2^{d\tilde{R}}$).

The decoder $g(f_1(U), f_2(V))$ computes

$$\hat{U}_{[d]} = \left[\tilde{U}_{[d]} - \tilde{Z}_1\right] \bmod \Lambda_c \tag{157}$$

$$\hat{V}_{[d]} = \left[\tilde{V}_{[d]} - \tilde{Z}_2\right] \bmod \Lambda_c, \tag{158}$$

and estimates the inner product as

$$g(f_1(U), f_2(V)) = \hat{U}_{[d]}^\top \hat{V}_{[d]}. \tag{159}$$

2) Analysis: We now analyze the performance of this scheme. First, note that

$$\begin{aligned}
\hat{U}_{[d]} = \left[\tilde{U}_{[d]} - \tilde{Z}_1\right] \bmod \Lambda_c &= \left[\left[Q_{\Lambda_f}\left(U_{[d]} + \tilde{Z}_1\right)\right] \bmod \Lambda_c - \tilde{Z}_1\right] \bmod \Lambda_c \\
&= \left[Q_{\Lambda_f}\left(U_{[d]} + \tilde{Z}_1\right) - \tilde{Z}_1\right] \bmod \Lambda_c \\
&= \left[U_{[d]} + \left(Q_{\Lambda_f}\left(U_{[d]} + \tilde{Z}_1\right) - (U_{[d]} + \tilde{Z}_1)\right)\right] \bmod \Lambda_c \\
&= \left[U_{[d]} + Z_1\right] \bmod \Lambda_c,
\end{aligned} \tag{160}$$

where

$$Z_1 = Q_{\Lambda_f}\left(U_{[d]} + \tilde{Z}_1\right) - \left(U_{[d]} + \tilde{Z}_1\right) \tag{161}$$

is uniform over $-\mathcal{V}_{\Lambda_f} = \mathcal{V}_{\Lambda_f}$ and statistically independent of U (and everything else), by the Crypto Lemma [13], [14]. Similarly, we obtain

$$\hat{V}_{[d]} = \left[\tilde{V}_{[d]} - \tilde{Z}_2\right] \bmod \Lambda_c = \left[V_{[d]} + Z_2\right] \bmod \Lambda_c, \tag{162}$$

where $Z_2 \sim \mathrm{Uniform}(\mathcal{V}_f)$ is statistically independent of V (and everything else).

Let

$$\hat{U}_{[d],\mathrm{ideal}} = U_{[d]} + Z_1 \tag{163}$$

$$\hat{V}_{[d],\mathrm{ideal}} = V_{[d]} + Z_2 \tag{164}$$

$$\tag{165}$$

and define the overload events

$$\mathrm{OL}_1 = \{U_{[d]} + Z_1 \notin \mathcal{V}_{\Lambda_c}\}, \quad \mathrm{OL}_2 = \{V_{[d]} + Z_2 \notin \mathcal{V}_{\Lambda_c}\}, \quad \mathrm{OL} = \mathrm{OL}_1 \cup \mathrm{OL}_2. \tag{166}$$

In particular, if OL did not occur, the $\bmod \Lambda_c$ operation in (160) and in (162) is inactive, and $\hat{U}_{[d]} = \hat{U}_{[d],\mathrm{ideal}}$ and $\hat{V}_{[d]} = \hat{V}_{[d],\mathrm{ideal}}$. Let

$$e = \alpha g(f_1(U), f_2(V)) - U^\top V = \alpha \hat{U}_{[d]}^\top \hat{V}_{[d]} - U^\top V, \tag{167}$$

and

$$e_{\mathrm{ideal}} = \alpha \hat{U}_{[d],\mathrm{ideal}}^\top \hat{V}_{[d],\mathrm{ideal}} - U^\top V. \tag{168}$$

If $\Pr(\mathrm{OL})$ is very small, then intuitively $\mathbb{E}[e^2]$ should be close to $\mathbb{E}[e_{\mathrm{ideal}}^2]$. Indeed, we prove the following in Appendix C

Proposition 4:

$$\mathbb{E}(e^2) \le \mathbb{E}[e_{\mathrm{ideal}}^2] + 75 \Pr(\mathrm{OL}) \cdot M^4(r_{\mathrm{cov}}(\Lambda_c)) \tag{169}$$

where

$$M(r_{\rm cov}(\Lambda_c)) = \max\{\sqrt{n}, r_{\rm cov}(\Lambda_c)\}. \tag{170}$$

We therefore proceed to compute $\mathbb{E}(e_{\rm ideal}^2)$. Note that

$$e_{\rm ideal} = \alpha\left(U_{[d]}^\top V_{[d]} + U_{[d]}^\top Z_2 + V_{[d]}^\top Z_1 + Z_1^\top Z_2\right) - \rho n. \tag{171}$$

Since $P_{UVZ_1Z_2} = P_{UV}P_{Z_1}P_{Z_2}$ and all random vectors U, V, Z_1, Z_2 have zero mean, we have

$$\mathbb{E}[U_{[d]}^\top V_{[d]}] = \rho d,\ \mathbb{E}[U_{[d]}^\top Z_2] = 0,\ \mathbb{E}[V_{[d]}^\top Z_1] = 0,\ \mathbb{E}[Z_1^\top Z_2] = 0, \tag{172}$$

and therefore

$$\mathbb{E}[e_{\rm ideal}^2] = \rho^2 n^2 - 2\alpha\rho^2 nd + \alpha^2\left(\mathbb{E}[(U_{[d]}^\top V_{[d]})^2] + \mathbb{E}[(U_{[d]}^\top Z_2)^2] + \mathbb{E}[(V_{[d]}^\top Z_1)^2] + \mathbb{E}[(Z_1^\top Z_2)^2]\right). \tag{173}$$

In Appendix D we show that

$$\mathbb{E}[(U_{[d]}^\top V_{[d]})^2] \le \rho^2 n\frac{d(d+1)}{n} + \frac{d(n-d)}{n}, \tag{174}$$

Furthermore,

$$\mathbb{E}[(V_{[d]}^\top Z_1)^2] = \mathbb{E}[(U_{[d]}^\top Z_2)^2] = \operatorname{tr}\mathbb{E}[U_{[d]}U_{[d]}^\top]\mathbb{E}[Z_2Z_2^\top] = \operatorname{tr}\mathbb{E}[Z_2Z_2^\top] = \mathbb{E}\|Z_2\|^2 = d\cdot\sigma^2(\Lambda_f), \tag{175}$$

$$\mathbb{E}[(Z_1^\top Z_2)^2] = \operatorname{tr}\mathbb{E}[Z_1Z_1^\top]\mathbb{E}[Z_2Z_2^\top] = \operatorname{tr}R^2(\Lambda_f) = \|R(\Lambda_f)\|_F^2, \tag{176}$$

where (175) follows since $\mathbb{E}[U_{[d]}U_{[d]}^\top] = I_d$. Thus,

$$\mathbb{E}[e_{\rm ideal}^2] \le \rho^2 n^2 - 2\alpha\rho^2 nd + \alpha^2\left(\rho^2 n\frac{d(d+1)}{n} + \frac{d(n-d)}{n} + 2d\sigma^2(\Lambda_f) + d\frac{1}{d}\|R(\Lambda_f)\|_F^2\right) \tag{177}$$

$$= n\left[\rho^2 n\left(1 - 2\alpha\frac{d}{n} + \alpha^2\frac{d(d+1)}{n^2}\right) + \alpha^2\left(\frac{d(n-d)}{n^2} + 2\frac{d}{n}\sigma^2(\Lambda_f) + \frac{d}{n}\frac{1}{d}\|R(\Lambda_f)\|_F^2\right)\right] \tag{178}$$

$$= n\xi(\Lambda_f) \tag{179}$$

where

$$\xi(\Lambda_f) = \rho^2 n\left(\left(1 - \frac{d}{n}\alpha\right)^2 + \alpha^2\frac{d}{n^2}\right) + \alpha^2\frac{d}{n}\left(1 - \frac{d}{n} + \psi\left(\sigma^2(\Lambda_f)\right)\right) + \alpha^2\frac{d}{n}\left(\frac{1}{d}\|R(\Lambda_f)\|_F^2 - \sigma^4(\Lambda_f)\right) \tag{180}$$

and $\psi(t) = 2t + t^2$. Note that $\xi(\Lambda_f)$ is monotonically increasing in both $\sigma^2(\Lambda_f)$ and $\frac{1}{d}\|R(\Lambda_f)\|_F^2$. We have therefore obtained that

$$\frac{1}{n}\mathbb{E}(e^2) \le \xi(\Lambda_f) + \frac{75M^4(r_{\rm cov}(\Lambda_c))}{n}\Pr(\mathrm{OL}). \tag{181}$$

The expression in (181), holds for any pair of nested lattices $\Lambda_c \subset \Lambda_f$. We now evaluate it for "good" nested lattices, whose existence is guaranteed by Theorem 13. Recall that $R > 0$ is fixed. Applying this theorem with $P_{U_{[d]}}$ taken as the uniform (Haar) distribution over $\sqrt{n}\mathbb{S}^{n-1}$ projected to the first $d \le n$ coordinates, $\alpha = 1$, $\beta = 1$, $r_u = 1 + \varepsilon_0$ and some $0 < \varepsilon_0 \le \frac{1}{\sqrt{2}}$, we have that for

$$D = \frac{1}{2^{2\left(\frac{n}{d}R - \delta\right)} - 1};\quad \delta = C_1\left(2\varepsilon_0 + \frac{\log d}{d}\right) \tag{182}$$

we can find a pair of nested lattice $\Lambda_c \subset \Lambda_f$ satisfying Items 1-5. In particular, for such lattices we have that

$$\frac{1}{n}\xi(\Lambda_f) \le \rho^2 n\left(\left(1 - \frac{d}{n}\alpha\right)^2 + \alpha^2\frac{d}{n^2}\right) + \alpha^2\frac{d}{n}\left(1 - \frac{d}{n} + \psi\left(D\right)\right) + \alpha^2\frac{d}{n}D^2\frac{C_2\log^3 d}{d} \tag{183}$$

$$\frac{1}{n}M^4(\Lambda_c) \le n\max\{1, D^2 2^{4\frac{n}{d}R}\} \tag{184}$$

$$\Pr(\mathrm{OL}) \le 2\Pr(\mathrm{OL}_1) = 2\Pr\left(U_{[d]} + Z_1 \notin \mathcal{V}_{\Lambda_c}\right) = 2\left(\Pr\left(\|U_{[d]}\|^2 > (1+\varepsilon_0)d\right) + 6e^{-d\frac{\varepsilon_0^2}{2}}\right) \le 16e^{-d\frac{\varepsilon_0^2}{96}}, \tag{185}$$

where in the last inequality we have used Proposition 5, proved in Appendix E, which shows that

$$\Pr\left(\|U_{[d]}\|^2 > (1+\varepsilon_0)d\right) < 2e^{-\frac{\varepsilon_0^2}{96}d}, \tag{186}$$

for all $0 < \varepsilon_0 < 1$. Plugging these into (181) we get

$$\begin{aligned}\frac{1}{n}\mathbb{E}(e^2) \le \rho^2 n\left(\left(1-\frac{d}{n}\alpha\right)^2 + \alpha^2\frac{d}{n^2}\right) + \alpha^2\frac{d}{n}\left(1-\frac{d}{n}+\psi\left(D\right)\right)\\ + D^2\frac{C_2\log^3 d}{d} + 1200n\max\{1, D^2 2^{4\frac{n}{d}R}\}e^{-d\frac{\varepsilon_0^2}{96}}.\end{aligned} \tag{187}$$

It can be verified that

$$\psi(D) = \frac{\phi\left(\frac{D}{D+1}\right)}{1-\phi\left(\frac{D}{D+1}\right)}, \tag{188}$$

which implies that

$$\psi\left(\frac{1}{2^{2t}-1}\right) = \frac{\phi(2^{-2t})}{1-\phi(2^{-2t})}. \tag{189}$$

Thus, for any $\kappa > 0$ and $\varepsilon > 0$ we can take $\varepsilon_0 > 0$ small enough and n large enough, we have that

$$\frac{1}{n}\mathbb{E}\left(U^\top V - \alpha g(f_1(U), f_2(V))\right)^2 \le \rho^2 n\left(1-\kappa\alpha\right)^2 + \kappa\alpha^2\left(1-\kappa+\frac{\phi\left(2^{-2\frac{R}{\kappa}}\right)}{1-\phi\left(2^{-2\frac{R}{\kappa}}\right)}\right) + \varepsilon(1+\rho^2 n), \tag{190}$$

such that (112) holds. This establishes the first part of the theorem.

The second part of the theorem follows from the same nested lattice coding scheme for encoding U, setting $\hat{V_{[d]}} = V_{[d]}$, and applying the same decoder. The analysis is identical, but with $Z_2 = 0$.

VII. Practical Implementation of Nested Lattice Quantizers

In the proof of Theorem 11 we used a pair of nested lattices $\Lambda_c \subset \Lambda_f \subset \mathbb{R}^d$, with $|\Lambda_f/\Lambda_c| = 2^{dR}$. Given such a pair of lattices in $\mathbb{R}^d$, in order to implement the coding scheme described above, we need to implement the following procedures:

1) $Q_{\Lambda_f}(x) = \operatorname{argmin}_{\lambda_f\in\Lambda_f}\|x-\lambda_f\|$
2) $Q_{\Lambda_c}(x) = \operatorname{argmin}_{\lambda_c\in\Lambda_c}\|x-\lambda_c\|$
3) Mapping from Λ_f/Λ_c to dR bits
4) Mapping from dR bits to the coset representatives $\Lambda_f \cap \mathcal{V}_c$ of Λ_f/Λ_c
5) Generating a random dither $Z \sim \mathrm{Uniform}(\mathcal{V}_{\Lambda_f})$, where $\mathcal{V}_{\Lambda_f}$ is the Voronoi cell of Λ_f

Self-similar nested lattice codebooks/Voronoi codes: Let $\Lambda \subset \mathbb{R}^d$ be a lattice with generating matrix $G \in \mathbb{R}^{d\times d}$, such that $\Lambda = G\mathbb{Z}^d$. Assume that we have access to a procedure that implements the lattice quantizer $Q_\Lambda(x)$ efficiently, and that there is some $\tau > 0$ such that $\tau\mathbb{Z}^d \subset \Lambda$. The assumption that $\mathbb{Z}^d$ is nested in Λ (up to scaling) is not very important, but also not restrictive, since the majority of lattices for which efficient lattice quantizers are known do satisfy it.

Using the lattice Λ, we can construct a pair of nested lattices $\Lambda_c \subset \Lambda_f \subset \mathbb{R}^d$, with $|\Lambda_f/\Lambda_c| = 2^{dR}$, that induce an efficiently implementable coding scheme. In particular, let $\beta > 0$ and set $\Lambda_f = \beta\Lambda$, $\Lambda_c = q\Lambda_f = \beta\cdot q\Lambda$, where $q = 2^R$ is an integer. In [9], Conway and Sloane propose simple implementation of the encoders and decoder for the induced nested lattice quantizer, which they referred to as Voronoi codes. Algorithm 1 below provides the pseudo code for implementing f_1, f_2 from Subsection VI-B1 for such a nested lattice codebook. Note that the output $\mathrm{OverloadError}$ of Algorithm 1 specifies whether or not the overload event OL_i, $i = 1, 2$, defined in (166) have occurred. In order to implement the decoder g from Subsection VI-B1, one implements (157) by applying Algorithm 2 on the output of f_1, implements (158) by applying Algorithm 2 on the output of f_2, and computes the inner product of the two vectors. In order to generate the random dithers $\tilde{Z}_1, \tilde{Z}_2$, one applies Algorithm 3.

Choice of the parameter β: Using this scheme, we have that

$$D = \sigma^2(\Lambda_f) = \beta^2\sigma^2(\Lambda). \tag{191}$$

Algorithm 1 NestedLatticeEncoder

Inputs: vector to be encoded $x \in \mathbb{R}^{d'}$, lattice $\Lambda \subset \mathbb{R}^{d'}$ with generating matrix $G \in \mathbb{R}^{d' \times d'}$, nesting ratio $q \in \mathbb{N}$, dither vector $z \in \mathcal{V}_\Lambda \subset \mathbb{R}^{d'}$, scaling factor $\beta > 0$
Outputs: $\mathrm{Enc}(x) \in [q]^{d'}$ (can be represented using $\lceil d' \log q \rceil$ bits), OverloadError that indicates if a modulo error occurred

$t \leftarrow Q_\Lambda\left(\frac{x}{\beta} + z\right)$
$y \leftarrow G^{-1}t$
$\mathrm{Enc}(x) \leftarrow [y] \bmod q$ (elementwise modulo q reduction)

% check whether a modulo error occurred:

$\tilde{x} \leftarrow t - z$
$\lambda_c = q \cdot Q_\Lambda\left(\frac{\tilde{x}}{q}\right)$
$\mathrm{OverloadError} = \mathbb{1}\{\lambda_c \neq 0\}$

Algorithm 2 NestedLatticeDecoder

Inputs: The encoding $\mathrm{Enc}(x) \in [q]^{d'}$ of $x \in \mathbb{R}^{d'}$, lattice $\Lambda \subset \mathbb{R}^{d'}$ with generating matrix $G \in \mathbb{R}^{d' \times d'}$, nesting ratio $q \in \mathbb{N}$, dither vector $z \in \mathcal{V}_\Lambda \subset \mathbb{R}^{d'}$, scaling factor $\beta > 0$
Outputs: $\hat{x} \in \mathbb{R}^{d'}$

$\tilde{y} \leftarrow G \cdot \mathrm{Enc}(x) - z$
$\hat{x} \leftarrow \beta\left(\tilde{y} - q \cdot Q_\Lambda\left(\frac{\tilde{y}}{q}\right)\right)$

Algorithm 3 GenerateRandomDither

Inputs: Lattice $\Lambda \subset \mathbb{R}^{d'}$ and a number $\tau > 0$ such that $\tau\mathbb{Z}^{d'} \subset \Lambda$
Outputs: $Z \sim \mathrm{Uniform}(\mathcal{V}_\Lambda)$

$U \leftarrow \mathrm{Uniform}\left([0,\tau)^{d'}\right)$
$Z \leftarrow U - Q_\Lambda(U)$

Thus, since the base lattice Λ is given, the parameter β controls D. We also have that

$$\sigma^2(\Lambda_c) = q^2\sigma^2(\Lambda_f) = 2^{2R}D. \tag{192}$$

The "no-overload" event is equivalent to $U_{[d]} + Z_1 \in \mathcal{V}_c$ (and similarly, $V_{[d]} + Z_2 \in \mathcal{V}_c$). If Λ is a "good" high-dimensional ($d \gg 1$) lattice, that is Λ is such that $\beta q\Lambda = \Lambda_c \subset \Lambda_f = \beta\Lambda$ satisfy all items in Theorem 13, the "no-overload" event happens with high probability provided that $1 + D = \frac{1}{d}\mathbb{E}\|U + Z_1\|^2 < 2^{2R}D$, which is equivalent to $D > D^*(R) = \frac{1}{2^{2R}-1}$. In practice, we will usually work with a base lattice Λ whose second moment and coding goodness are sub-optimal. For this reason, we take $D = \gamma D^*(R) = \frac{\gamma}{2^{2R}-1}$, for some $\gamma > 0$ (where γ is not necessarily close to 1), which is done by setting

$$\beta = \left(\frac{\gamma}{2^{2R}-1} \cdot \frac{1}{\sigma^2(\Lambda)}\right)^{1/2}. \tag{193}$$

Overload avoidance mechanism: Recall that Algorithm 1 also indicates, through the variable OverloadError, whether or not a modulo error occurred, that is, whether or not $U_{[d]} + Z_1 \in \mathcal{V}_c$ (respectively, $V_{[d]} + Z_2 \in \mathcal{V}_c$). Whenever a modulo error does occur, one can increase the value of γ further to a large enough value, such that a modulo error does not occur with the new value, and inform the decoder on what value of γ was chosen. In practice, we may choose a bank of M values sorted in increasing order $\gamma \in \{\gamma_1, \ldots, \gamma_M\}$. The encoder first uses γ_1. If $\mathrm{OverloadError} = 1$ it tries again with γ_2, and keeps increasing γ to the next value until $\mathrm{OverloadError} = 0$. If γ_1 is chosen such that overload error is already not too common, and the values of γ_i increase sufficiently fast with i, say $\gamma_i = i \cdot \gamma_1$, the entropy of the first value of γ that returned $\mathrm{OverloadError} = 0$ will be small. Since we only have to report this index to the decoder once for d symbols, the effect on the quantization rate is not significant.

We note that the newest GPUs released by Nvidia use a format called *micro-block scaling* for low-precision data types such as FP4 [61]. For example, the NVFP4 data type assigns a different scale β, represented in FP8 format, to every 16 consecutive entries, where β is chosen to ensure that all 16 entries are within the dynamic range of the FP4 format, or in other words, that overload does not occur. Thus, such formats can be seen as a special case of our overload avoidance mechanism.

Next, we develop a heuristic for choosing γ_1. Recall the definition of $r_{\text{eff}}(\Lambda) = \left(\frac{\text{covol}(\Lambda)}{V_d}\right)^{1/d}$ from Section VI-A. The normalized second moment (NSM) of a lattice Λ is defined as

$$N(\Lambda) = \frac{\sigma^2(\Lambda)}{(\text{covol}(\Lambda))^{2/d}} = \frac{\sigma^2(\Lambda)}{V_d^{2/d} r_{\text{eff}}^2(\Lambda)}. \tag{194}$$

If $U_{[d]} + Z_1$ were Gaussian, the probability that it stays within $\mathcal{V}_{\Lambda_c}$ would have been upper bounded by the probability that it stays within a ball with the same volume, that is, within a ball with radius $r_{\text{eff}}(\Lambda_c)$. Thus, we need $r_{\text{eff}}^2(\Lambda_c)$ to be greater than $\mathbb{E}\|U_{[d]} + Z_1\|^2$. This corresponds to

$$1 < \frac{\frac{1}{d} r_{\text{eff}}^2(\Lambda_c)}{\frac{1}{d}\mathbb{E}\|U + Z_1\|^2} = \frac{1}{d}\frac{r_{\text{eff}}^2(\Lambda_c)}{\sigma^2(\Lambda_c)} \frac{\sigma^2(\Lambda_c)}{\frac{1}{d}\mathbb{E}\|U + Z_1\|^2} = \frac{1}{dV_d^{2/d} N(\Lambda)} \frac{2^{2R} D}{1 + D} = \frac{1}{dV_d^{2/d} N(\Lambda)} \frac{\gamma \cdot 2^{2R}}{2^{2R} + \gamma - 1} \approx \frac{\gamma}{dV_d^{2/d} N(\Lambda)}, \tag{195}$$

where the last approximation assumes that $2^{2R} + \gamma - 1 \approx 2^{2R}$. Thus, we will take

$$\gamma_1 \gtrapprox dV_d^{2/d} N(\Lambda) = \frac{d\sigma^2(\Lambda)}{r_{\text{eff}}^2(\Lambda)}. \tag{196}$$

For a measurable set $\mathcal{K} \subset \mathbb{R}^d$ let $U_{\mathcal{K}} \sim \text{Uniform}(\mathcal{K})$ and $\sigma^2(\mathcal{K}) = \frac{1}{d}\mathbb{E}\|U_{\mathcal{K}}\|^2$. For all measurable sets $\mathcal{K}$ with volume $V_d r_{\text{eff}}^d(\Lambda)$, we have that $\sigma^2(\mathcal{K}) \geq \frac{r_{\text{eff}}^2(\Lambda)}{d+2}$, and this is attained by $\mathcal{K} = r_{\text{eff}}(\Lambda)\mathcal{B}$ [14]. It therefore follows that the right hand side of (196) is at least $\frac{d}{d+2}$.

Product lattices/Product quantization: In order to use the self-similar nested lattice scheme described above, we need a base lattice Λ with an efficient nearest-neighbor decoder/lattice quantizer $Q_\Lambda(x)$ and favorable quantization and coding properties. While it is easy to find (more accurately, to randomly draw) lattices in high-dimensions that are good for coding and quantization (see Section VI-A), the task of finding such lattices that also admit an efficient nearest-neighbor decoder is notoriously difficult and is perhaps the holy grail of coding theory for the additive white Gaussian noise (AWGN) channel. A popular compromise between efficiency and "goodness", is to use a *product lattice*, with a low-dimensional base lattice that is "pretty-good" for coding and quantization [62], [14].

Let d' be an integer that divides d, and $\Lambda' \subset \mathbb{R}^{d'}$ be a lattice in $\mathbb{R}^{d'}$. We construct the lattice $\Lambda \in \mathbb{R}^d$ as the product of $K = d/d'$ copies of Λ'. Namely,

$$\Lambda = \underbrace{\Lambda' \times \cdots \times \Lambda'}_{K \text{ times}} = \Lambda'^{\otimes K} \tag{197}$$

The resulting self-similar nested lattices are also the product of K nested lattice pairs

$$\Lambda_c \subset \Lambda_f = (\beta_1 \cdot q\Lambda' \subset \beta_1\Lambda') \times \cdots \times (\beta_K \cdot q\Lambda' \subset \beta_K\Lambda'), \tag{198}$$

where we allow for different choices of β for each product to accommodate for the overload avoidance mechanism described above. Algorithm 1, Algorithm 2 and Algorithm 3 tensorize, and should be applied separately for each $k = 1, \ldots, K$ using the base lattice $\Lambda' \subset \mathbb{R}^{d'}$ with generating matrix $G' \in \mathbb{R}^{d' \times d'}$. We also have that

$$\sigma^2(\Lambda) = \sigma^2(\Lambda') \cdot \frac{1}{K} \sum_{\ell=1}^{K} \beta_k^2. \tag{199}$$

Some lattices in small dimensions have excellent quantization and coding properties, as well as efficient nearest neighbor decoding algorithms. In particular, $A_3 \cong D_3$ has the highest packing density among all lattices in $\mathbb{R}^3$ [63], A_3^* has the smallest NSM among all lattices in $\mathbb{R}^3$ [63] (only slightly smaller than that of A_3), D_4 has the highest packing density among all lattices in $\mathbb{R}^4$ and lowest known NSM among all lattices in $\mathbb{R}^4$ [63], [64], and E_8 has the highest packing density (even among non-lattice packings) [65] and the smallest known NSM among all lattices in $\mathbb{R}^8$ [63], [64]. All four lattices listed above, as well as many others from the A_n, D_n and E_n families, admit a very fast lattice decoding algorithm [66]. Similarly, among all lattices in $\mathbb{R}^{24}$, the Leech lattice Λ_{24}, is the the best

known quantizer [64], has the optimal packing density [67] (this is true even among all non-lattice packings [68]), and admits a pretty fast nearest neighbor decoding (or approximate nearest neighbor decoding) algorithms [69], [70], [71]. In addition, the second moment of all these lattices (and others) is calculated in [72] and reported also in [64, Table I]. See also [73] for a study of the NSM of a random lattice. We also note that the optimal lattice quantizer in any dimension has $R(L) = \sigma^2(L) \cdot I_d$, so that $\frac{1}{d}\|R(L)\|_F^2 = \sigma^4(L)$ for those lattices. Any one of those lattices is a good candidate for the base lattice Λ'. Another important advantage of these lattices is that they are all subsets of $\mathbb{Z}^n$ up to scaling. Thus, when these lattices are used for quantization for matrix multiplication, and dithering is not applied, we can use integer multipliers (e.g., int8 tensor core in a GPU), rather than floating point multipliers, for multiplying the quantized matrices. The lattices of higher dimensions, and in particular the Leech lattice, may yield better rate-distortion tradeoff than the lower-dimensional ones, but there are advantages to using lower-dimensional lattices in terms of efficiency. One of those is described next.

Lookup tables: Note that we decode $\hat{U}_k \in \mathbb{R}^{d'}$ and $\hat{V}_k \in \mathbb{R}^{d'}$ just to compute their inner product $\hat{U}_k^\top \hat{V}_k$. If we use the same dither vectors $\tilde{Z}_1, \tilde{Z}_2 \in \mathcal{V}_{\Lambda'}$ for all $k = 1, \ldots, K$, and the same value of β, namely, $\beta_k^U = \beta_k^V = \beta$ for all $k = 1, \ldots, K$, there are only $q^{d'}$ values of $\hat{U}_k^\top$ we can get, and only $q^{d'}$ values of $\hat{V}_k^\top$ we can get. Those do not depend on k. Thus, we can pre-compute all $q^{2d'}$ possible values of $\hat{U}_k^\top \hat{V}_k$ and store them in a lookup table (LUT). Then, instead of applying the decoder twice and computing the inner product, we simply fetch the result of the inner product from the LUT. If $\beta_k^U \neq \beta$ or $\beta_k^V \neq \beta$, we simply multiply the value fetched from the LUT by $\frac{\beta_k^U}{\beta} \cdot \frac{\beta_k^V}{\beta}$. On some processors, using LUTs significantly speed up the decoding process, as it completely bypasses all lattice decoding operations, as well as all inner products. For approximate matrix multiplication $A^\top B$ of $A \in \mathbb{R}^{n \times a}$ and $B \in \mathbb{R}^{n \times b}$ using the product nested lattice quantization scheme above, we need to perform $a \cdot b \cdot (n/d')$ such operations, whereas the encoding only involves $a(n/d') + b(n/d')$ lattice encoding operations. Thus, for $a, b \gg 1$, decoding is the computationally heavy procedure, and speeding it up will result in significant speedup of the total approximate matrix multiplication procedure. Using LUTs is therefore often highly advantageous. However, in order to have a very fast access time to the LUT, we would like it to "fit" in the highest levels of the cache, ideally in the L1 cache. This level has small capacity, which restricts the values of $q^{2d'} = 2^{2Rd'}$. Thus, we must keep Rd' small. Taking small R will typically not yield satisfactory resolution, so if LUTs are used, we are limited to using lattices Λ' of small dimensions. Fortunately, a compelling solution to this shortcoming of the LUT approach was recently developed in [38]. We note that for GPUs the LUT approach may not be attractive since the tensor core computes matrix multiplications extremely fast, while LUT probing is less efficient on this hardware. On CPUs on the other hand, the LUT approach can potentially yield significant speed-up.

Hadamard transform: Our encoders f_1, f_2 for the matrix multiplication problem, as described in the proof of Theorem 12 and in Figure 1, multiply each column vector in A as well as each column vector in B (more accurately, in their centered versions $\bar{A}, \bar{B}$), by a random projection matrix S drawn from the Haar distribution on $\mathrm{O}_n(\mathbb{R})$. In general, the matrix S drawn from this distribution will have no structure, and calculating SA (respectively SB) will require $O(an^2)$ (respectively, $O(bn^2)$) real-valued multiplication and summation operations. To significantly reduce the computational burden of this step, it was proposed in [6] (see also [7]) to restrict S to a certain class of orthogonal projection matrices: The randomized Hadamard transform. Here, we also follow this approach. In particular, we draw a vector $T \sim \mathrm{Uniform}(\{-1, 1\}^n)$, and set $K = \mathrm{diag}(T)$, that is, K is a diagonal matrix with $K_{i,i} = T_i$. We then set

$$S = \frac{1}{\sqrt{n}} HK, \tag{200}$$

where $H \in \{-1, 1\}^{n \times n}$ is the Walsh-Hadamard matrix of dimension n. Here, we assumed that n is a power of 2, such that such a matrix exists. Otherwise, we can add rows of all zeros to both A and B, resulting in larger matrices $A \in \mathbb{R}^{n' \times a}$ and $B \in \mathbb{R}^{n' \times b}$, with $n' = 2^{\lceil \log_2(n) \rceil}$. Note that in (118-119) we further scale the result by $\sqrt{n}$, so this cancels out the scaling by $\frac{1}{\sqrt{n}}$ in (200). The gain for using the randomized Hadamard transform (200), is that its special fast-Fourier transform (FFT) structure allows to compute SA (respectively, SB) using only $O(an \log n)$ (respectively, $O(bn \log n)$) additions and multiplications. Despite its simple implementation, the result of applying the randomized Hadamard transform on A (or B) is quite similar to that of applying a "pure" random rotation on A (or B) from various statistical perspectives [74], [75], [76].

Representative numeric example: To better illustrate how the building blocks above connect, we provide a numerical example. We have implemented a product nested lattice codebook, with $\Lambda' = D_3$ (such that $d' = 3$) as

the base lattice. The lattice D_3 consists of all vectors in $\mathbb{Z}^3$ whose entries sum up to an even integer. In particular, $2\mathbb{Z}^3 \subset D_3$. The simple structure of D_3 also gives rise to a very simple algorithm for computing $Q_{D_3}(x)$ [66, Algorithm 2]. The lattice D_3 has the highest packing density among all lattices in $\mathbb{R}^3$ and its packing radius satisfies [63] $r_{\text{pack}}(D_3)/r_{\text{eff}}(D_3) \approx (0.74)^{1/3} \approx 0.9045$, such that its Voronoi region is quite close to a ball. We also have that $\sigma^2(D_3) = \frac{3}{24}$, so that $N(D_3) \approx 0.0787$ (since $\text{covol}(D_3) = 2$). This NSM is only slightly greater than the smallest NSM attained by any lattice in $\mathbb{R}^3$, which is $N(A_3^*) \approx 0.0785$.

We have used this base lattice with $q = 6$ to construct a product nested lattice code as in (198). We used the same dither vectors $\tilde{Z}_1, \tilde{Z}_2 \in \mathcal{V}_{D_3}$ for all $k = 1, \ldots, K$ (these vectors were drawn once at the beginning of the experiment). For this choice of $d' = 3$ and $q = 6$, we can implement the decoder using a lookup table of size $(q^3)^2 = 2^{6 \log_2 q} < 2^{15.6}$. For constructing the LUT, we used the value $\beta = 1$. While for this choice of β all inner products between vectors in D_3 are integer valued, because of the use of dithers, the entries in our LUT are not integer-valued in general. We nevertheless rounded each of them to the nearest integer, and their range allows representing each entry in the LUT using an INT8 variable. Consequently, the total size of the LUT is less than 64Kbyte, and it can be fully stored in the L1 cache of a modern processing unit.

For the lattice D_3, we have that the right-hand side of (196) evaluates to ≈ 0.6139. For the experiment, we chose $\gamma_1 = 0.7$, and set our bank of possible values of γ as $\{i \cdot \gamma_1\}_{i=1}^9$. The corresponding value of β is given by (193).

We drew two random matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n}$, with all entries iid $\mathcal{N}(0, 1)$, where $n = 3 \cdot 2^{11}$. We used the product nested lattice codebook from (198) with $K = 2^{11}$ for encoding each column of A (using the dither vector $\tilde{Z}_1$) and for encoding each column of B (using the dither vector $\tilde{Z}_2$). Since the iid Gaussian distribution is already rotation invariant, we have not implemented a random rotation. Since the distribution we have used is zero mean, we also did not implement the "centering" mechanism. We also used $\alpha = \kappa = 1$ (no MMSE scaling and no time-sharing). For each column, we further report the K values of β_i (equivalently γ_i) used for each column. The (empirical) entropy of this random variable (that takes values in $\beta_1 \cdot \{i\}_{i=1}^9$) for the choice $\gamma_1 = 0.7$ was found to be around ≈ 1.3bits. Since this value is only reported once for every $d' = 3$ symbols (using entropy coding), its contribution to the coding rate is about 0.43 bits per symbol, such that the total rate of the coding scheme is $R_{\text{eff}} \approx \log_2(6) + 0.43 \approx 3.015$ bits/symbol.

This approximate matrix multiplication algorithm attained $\frac{1}{n^3}\|\widehat{A^\top B} - A^\top B\|_F^2 \approx 0.0593$. Let $e = \widehat{A^\top B} - A^\top B$. The empirical distribution of the normalized approximation error $e/\sqrt{n}$ (among the n^2 entries) is plotted in Figure 3. Note that for $R_{\text{eff}} = 3.015$, Theorem 2 states that no scheme can attain distortion smaller than of $\Gamma(R_{\text{eff}}) = 0.0304$ for A and B drawn as above, and Theorem 1 shows that this can be attained using high-dimensional lattices. Thus, our low-complexity implementation is not far of the optimal performance attained using optimal lattice codes. For comparison, we also evaluated the approximation error for a simple 3-bit scalar quantization scheme where each column a_i is normalized by $\|a_i\|_\infty$ such that all its entries are in $[-1, 1]$, then each entry $\tilde{a}_{i,t} = \frac{a_{i,t}}{\|a_i\|_\infty}$ is quantized to $\frac{1}{4}\text{round}(4\tilde{a}_{i,t})$, and in the end the quantized entries are rescaled again by $\|a_i\|_\infty$. The empirical error attained by the 3-bit scalar quantizer is $\frac{1}{n^3}\|\widehat{A^\top B} - A^\top B\|_F^2 \approx 0.1668$, about 3 times greater than the error attained using the D_3-based scheme with the same rate. The performance gap between the two scheme grows with n, as the random variable $\|a_i\|_\infty$ concentrates around $\sqrt{2 \ln n}$ for large n. Thus, the dynamic range for the scalar quantizer increases with n, which results in greater expected squared error.

VIII. Open problems

One can interpret our Lemma 2 as follows: Let $P = \mathcal{N}(0, 1)$ and $U^n \sim P^{\otimes n}$. Then for any random variable Y we have that

$$\sum_{i=1}^{n} R_P(\lambda_i) \le I(U^n; Y), \tag{201}$$

where $R_P(D)$ is the quadratic rate-distortion function for a source with distribution P and $(\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of $\text{Cov}(U^n|Y)$. While Lemma 2 establishes (201) for the Gaussian distribution, we were not able to prove (201) for a general distribution, and we could neither find a counterexample. If (201) turns out to hold for any P, the proof of Theorem 6 could be easily extended to show that

$$D^{\text{IP}}(R, P) = \text{convex envelope of } (\phi(D_P(R)), \tag{202}$$

where $D_P(R)$ is the quadratic distortion-rate function for a source with distribution P. Thus, proving or disproving that (201) holds for all P is an interesting problem for future research.

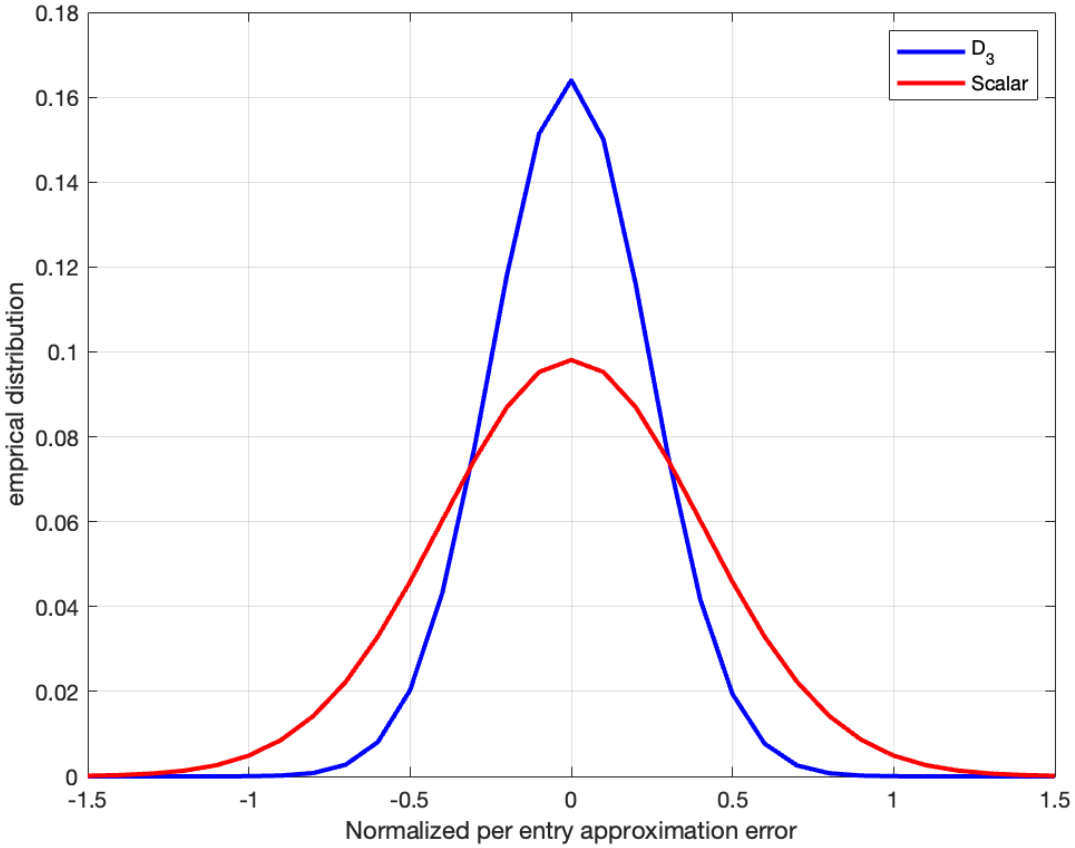

Fig. 3: The approximation error of the D_3-based product nested lattice coding scheme with $q = 6$, for random iid Gaussian matrices $A, B \in \mathbb{R}^{n\times n}$, $n = 3 \cdot 2^{11}$. We plot the histogram of the entries of $\frac{1}{\sqrt{n}}(\widehat{A^\top B} - A^\top B)$ in blue. For comparison, we also plot the histogram of the entries of $\frac{1}{\sqrt{n}}(\widehat{A^\top B} - A^\top B)$ for a 3-bit scalar quantizer in red.

In Theorem 1 we have shown the existence of encoders and decoder for quantization for matrix multiplication whose expected approximation error depends only on $\|\bar{A}\|_F^2 \cdot \|\bar{B}\|_F^2$ and $\|\bar{A}^\top \bar{B}\|_F^2$, and is optimal for A and B whose entries are iid Gaussian. For iid Gaussian matrices we have that $\frac{\mathbb{E}[\|A\|_F^2\|B\|_F^2]/n}{\mathbb{E}[\|A^\top B\|_F^2]} = 1$ so that the two error terms in (4) are well-balanced, and Theorem 1 essentially gives an upper bound of $\|A^\top B\|_F^2 \cdot \Gamma(R)$ on the MSE. Is there a scheme that attains MSE at most $\|A^\top B\|_F^2 \Gamma(R)$ universally (for all matrices A, B, not just iid ones)?

Another important question is *shared randomness*. Our construction crucially depends on encoders and decoder sharing randomness (which practically is not a big issue, since the random seed used by the encoder can be stored along the compressed matrix representation). In the single-terminal lossy compression shared randomness is not necessary. Indeed, suppose we have some compact metric space E with distance d and we proved that there exist a (shared randomness) encoder-decoder pair (f, g) compressing to L bits and achieving simultaneously for all $x \in E$ guarantee:

$$\mathbb{E}[d(x, g(f(x)))] \le \Delta\,.$$

Also suppose that there exists an ϵ-net of size M_1 in E. Fix $\delta > 0$ and average the previous inequality over all M_1 elements of the net. Then there must exist a choice ω_1 of shared randomness so that at most $M_2 = \frac{M_1}{1+\delta}$ elements of the ϵ-net have distortion exceeding $(1+\delta)\Delta$. Now repeat the argument for the subset M_2 to find choice ω_2, etc. After $k \le \frac{\log M_1}{\log(1+\delta)}$ steps we get $M_{k+1} = 0$. This shows that there must exist a $k2^L$-sized $(1+\delta)\Delta$-net that approximates each of M_1 elements. Thus, the space E can be covered to within distortion $(1+\delta)\Delta + \epsilon$ without any shared randomness by compressing down to $L + \log \frac{\log M_1}{\log(1+\delta)}$ bits. By choosing $\delta \to 0$ and $\epsilon \ll \Delta$, one can

thus get rid of shared randomness.

However, this method fails in the case of distributed compression of (A, B). Indeed, the previous argument breaks down because the choice ω of shared randomness affects both quantization grids of A and B simultaneously. Thus, it is not possible for the compressor who only knows A to decide which of the k choices of ω to use for quantizing A. It remains an open question to understand fundamental limits of deterministic quantizers.

As another extension, we may consider the question of quantization for product of k matrices $\prod_{t=1}^{k} A_t$. This paper solves the case of $k = 2$, but our methods do not seem to be immediately extendable to the $k > 2$ case. One remark we want to make, however, is regarding the critical rate. For $k = 2$ as we saw quantization below $R < 0.906$ bit/entry required additional dimensionality-reduction (or Johnson-Lindenstrauss) step. This critical point was found by convexifying the function $\Gamma_1(R) = 1 - (1 - 2^{-2R})^2$. Similarly, if one simply asks the question of optimal quantization for a product of k diagonal Gaussian matrices, one would need to convexify the function $\Gamma_1(R) = 1 - (1 - 2^{-2R})^k$. The associated critical rate grows with k from $R \approx 0.906$ for $k = 2$ matrices to $R \approx 4$ for $k = 46$ matrices etc. This suggests that quantization for deep LLMs at low rates may benefit from dimensionality reduction steps.

ACKNOWLEDGEMENTS

The authors thank Omri Weinstein (HUJI) for helping them navigate through the literature on approximate matrix multiplication and Yoon Kim (MIT) for explaining hardware and performance limitations of modern quantization algorithms in LLMs. They also thank the anonymous reviewers for their constructive feedback.

APPENDIX A
CONVEX ENVELOPE OF $\Gamma_1(R)$

Recall that $\phi(t) = 2t - t^2$ and

$$\Gamma_1(R) = \phi(2^{-2R}). \tag{203}$$

We show that the convex lower envelope of $\Gamma_1(R)$ is $\Gamma(R)$. It is easy to verify that $R \mapsto \Gamma_1(R)$ is decreasing, concave on $[0, 1/2)$ and convex on $(1/2, \infty)$. Therefore, its convex envelope consists of a linear segment between $(0, \Gamma_1(0) = 1)$ and $(R^*, \Gamma_1(R^*))$ and agrees with $\Gamma_1(R)$ for $R > R^*$. The point $R^* \geq 1/2$ is chosen such that the derivative of $\Gamma(R)$ is smooth and non-decreasing. Thus, the convex envelope of $\Gamma_1(R)$ is given by

$$\Gamma(R) = \begin{cases} \Gamma_1(R^*) + \Gamma_1'(R^*)(R - R^*) & R < R^* \\ \Gamma_1(R) & R \geq R^* \end{cases} \tag{204}$$

where R^* is chosen by requiring that $\Gamma(0) = \Gamma_1(0) = 1$, or in other words, that

$$\Gamma_1(R^*) - R^* \cdot \Gamma_1'(R^*) = 1. \tag{205}$$

Since $\Gamma_1'(R^*) = -4 \ln 2 \cdot 2^{-2R^*}(1 - 2^{-2R^*})$ and we can express $\Gamma_1(R^*)$ as $\Gamma_1(R^*) = 2 \cdot 2^{-2R^*}(1 - 2^{-2R^*}) + 2^{-4R^*}$, we have that (205) corresponds to

$$\begin{aligned} & 2^{-4R^*} + 2 \cdot 2^{-2R^*}(1 - 2^{-2R^*})(1 + 2 \ln 2R^*) = 1 \\ \iff & 2 \cdot 2^{-2R^*}(1 - 2^{-2R^*})(1 + 2 \ln 2R^*) = (1 - 2^{-2R^*})(1 + 2^{-2R^*}) \\ \iff & 2 \cdot 2^{-2R^*}(1 + 2 \ln 2R^*) = 1 + 2^{-2R^*}, \qquad (206) \\ \iff & 1 + 4 \ln 2R^* = 2^{2R^*}. \qquad (207) \end{aligned}$$

APPENDIX B
GOOD NESTED LATTICES

The proof of Theorem 13 will easily follow from Lemma 5, Lemma 6 and Lemma 7, below. We first state these Lemmas, and give the proof of Theorem 13, which uses them. The proofs of Lemma 6 and the proof of Lemma 7 are brought afterwards.

Following the notation from [16], we denote by $\mathcal{L}_d$ the space of lattices of unit covolume in $\mathbb{R}^d$, and by μ_d the natural probability distribution on $\mathcal{L}_d$, which we refer to as th Haar-Siegel measure. Let $r_{\text{eff}}(1) = V_d^{-\frac{1}{d}}$ be such that $V_d r_{\text{eff}}^d(1) = 1$ and therefore $r_{\text{eff}}(L) = r_{\text{eff}}(1)$ for all $L \in \mathcal{L}_d$. For $c_0, c_1 > 0$ define the set of lattices

$$E_{c_0,c_1} = \left\{ L \in \mathcal{L}_d \ : \ V_d r_{\text{cov}}^d(L) < c_0 d^{c_1} \right\}. \tag{208}$$

Lemma 5 (Corollary of Theorem 1.2 from [16]): For any $c_1 > 2$ there exists a universal constant $c_0 > 0$ such that

$$\mu_d(\{L \notin E_{c_0,c_1}\}) < \frac{1}{3}. \tag{209}$$

Lemma 6: For any $c_0, c_1 > 0$, there are universal constants $C, C' > 0$ such that for any $L \in E_{c_0,c_1}$, any $\kappa > 0$, $\alpha > 0$, and $\beta > 0$ the following hold

1) $r_{\mathrm{cov}}(\kappa L) \leq \left(1 + C\frac{\log d}{d}\right)\kappa r_{\mathrm{eff}}(L)$;
2) $\sigma^2(\kappa L) \leq \frac{1}{d}\left(\left(1 + C\frac{\log d}{d}\right)\kappa r_{\mathrm{eff}}(L)\right)^2$;
3) Let $Z \sim \mathrm{Uniform}(\mathcal{V}_{\kappa L})$, and let U be some random variable statistically independent of Z. Then for any $\mathcal{A} \subset \mathbb{R}^d$ we have
$$\Pr(\alpha U + \beta Z \notin \mathcal{A}) \leq d^C \Pr(\alpha U + \beta \tilde{Z} \notin \mathcal{A}) \tag{210}$$
where $\tilde{Z} \sim \mathrm{Uniform}\left(\left(1 + C\frac{\log d}{d}\right)\kappa r_{\mathrm{eff}}(L)\mathcal{B}\right)$ is statistically independent of U.
4) $\frac{1}{d}\|R(\kappa L)\|_F^2 \leq \left(\frac{1}{d}\left(\left(1 + C\frac{\log d}{d}\right)\kappa r_{\mathrm{eff}}(L)\right)^2\right)^2 \left(1 + C'\frac{\log^3 d}{d}\right)$;

Lemma 7: Let $U \sim P_U$ be a random variable in $\mathbb{R}^d$ that satisfies $\|U\| \leq \sqrt{d \cdot r_U}$ with probability 1, and let $\tilde{Z} \sim \mathrm{Uniform}(\sqrt{d \cdot r_b}\mathcal{B})$ be statistically independent of U. For $\alpha, \beta, \kappa, \varepsilon > 0$ let

$$E_{\alpha,\beta,\kappa,\varepsilon}^{P_U,r_U,r_b} = \left\{L \in \mathcal{L}_d \ : \ \Pr(\alpha U + \beta \tilde{Z} \notin \mathcal{V}_{\kappa L}) < 6e^{-d\frac{\varepsilon^2}{2}}\right\}. \tag{211}$$

Then, for any $0 < \varepsilon < \frac{1}{\sqrt{2}}$ and

$$\kappa > e^{\frac{\varepsilon^2}{2}}\sqrt{1+\varepsilon}\frac{\sqrt{d(\alpha^2 r_U + \beta^2 r_b)}}{r_{\mathrm{eff}}(1)}. \tag{212}$$

we have that

$$\mu_d\left(\left\{L \notin E_{\alpha,\beta,\kappa,\varepsilon}^{P_U,r_u,r_b}\right\}\right) < \frac{1}{3}. \tag{213}$$

Proof of Theorem 13. Let p be a prime number and k be a positive integer, such that $p^k \in [1/2, 1]2^{dR}$. Such numbers must exist. Denote by $\mathrm{Gr}_{d,k}(\mathbb{F}_p)$ the collection of subspaces of dimension k in $\mathbb{F}_p^d$. Let L be some lattice in $\mathcal{L}_d$, S be some subspace in $\mathrm{Gr}_{d,k}(\mathbb{F}_p)$, and let the lattice $L(S)$ be as defined in [57, eq. 13]. We have that $L \subset L(S)$ and $|L(S)/L| = p^k$. Fix some $c_1 > 2$ and $c_0 > 0$ for which (209) holds. Let C, C' be the universal constants from Lemma 6 and let

$$\kappa = \frac{\sqrt{dD}}{(1 + C\frac{\log d}{d})p^{-\frac{k}{d}}r_{\mathrm{eff}}(1)}. \tag{214}$$

We define

$$\Lambda_f = \kappa L(S), \quad \Lambda_c = \kappa L. \tag{215}$$

Thus, $|\Lambda_f/\Lambda_c| = |L(S)/L| = p^k$, and Item 1 holds with probability 1. Let $P_{\tilde{U}} = P_{U|\|U\|^2 \leq r_U}$ and define $\varepsilon' = \sqrt{\varepsilon^2 + 2C\frac{\log d}{d}}$ such that $\frac{\varepsilon'^2}{2} = \frac{\varepsilon^2}{2} + C\frac{\log d}{d}$. Define the events

$$E_f^{\text{quant}} = \{p^{\frac{k}{d}}L(S) \in E_{c_0,c_1}\} \ , \ E_c^{\text{quant}} = \{L \in E_{c_0,c_1}\}, \ E_c^{\text{code}} = \{L \in E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\}. \tag{216}$$

and assume they all occur (later we will show that if $L \sim \mu_d$ and $S \sim \mathrm{Uniform}(\mathrm{Gr}_{d,k}(\mathbb{F}_p)$ are statistically independent, the three events indeed occur simultaneously with positive probability).

From Lemma 6 we have that

$$r_{\rm cov}(\Lambda_f) = r_{\rm cov}\left(\kappa p^{-\frac{k}{d}} p^{\frac{k}{d}} L(S)\right) \le \left(1 + C\frac{\log d}{d}\right)\kappa p^{-\frac{k}{d}} r_{\rm eff}(1) = \sqrt{dD}, \tag{217}$$

$$r_{\rm cov}(\Lambda_c) = r_{\rm cov}(\kappa L) \le \left(1 + C\frac{\log d}{d}\right)\kappa r_{\rm eff}(1) = p^{k/d}\sqrt{dD} \le 2^R\sqrt{dD}, \tag{218}$$

$$\sigma^2(\Lambda_f) = \sigma^2\left(\kappa p^{-\frac{k}{d}} p^{\frac{k}{d}} L(S)\right) \le \frac{1}{d}\left(\left(1 + C\frac{\log d}{d}\right)\kappa p^{-\frac{k}{d}} r_{\rm eff}(1)\right)^2 = D \tag{219}$$

$$\begin{aligned}\frac{1}{d}\|R(\Lambda_f)\|_F^2 = \frac{1}{d}\|R(\kappa p^{-\frac{k}{d}} p^{\frac{k}{d}} L(S))\|_F^2 &\le \left(\frac{1}{d}\left(\left(1 + C\frac{\log d}{d}\right)\kappa p^{-\frac{k}{d}} r_{\rm eff}(1)\right)^2\right)^2 \left(1 + C'\frac{\log^3 d}{d}\right) \\ &= D^2\left(1 + C'\frac{\log^3 d}{d}\right)\end{aligned} \tag{220}$$

Thus, Λ_f and Λ_c satisfy Items 2-4, with $C_2 = C'$. To show that Item 5 holds, let $\tilde{Z} \sim \text{Uniform}(\sqrt{dD}\mathcal{B})$ be statistically independent of U and write

$$\Pr(\alpha U + \beta Z \notin \mathcal{V}_{\Lambda_c}) = \Pr(\alpha U + \beta Z \notin \mathcal{V}_{\kappa L}) \le \Pr(\|U\|^2 > dr_u) + \Pr(\alpha U + \beta Z \notin \mathcal{V}_{\kappa L} |\ \|U\|^2 \le dr_u) \tag{221}$$

$$\le \Pr(\|U\|^2 > dr_u) + d^C \Pr(\alpha U + \beta\tilde{Z} \notin \mathcal{V}_{\kappa L} |\ \|U\|^2 \le dr_u) \tag{222}$$

$$\le \Pr(\|U\|^2 > dr_u) + 6e^{-d\frac{\varepsilon^2}{2}}, \tag{223}$$

where (222) follows from Item 3 in Lemma 6 and the definition of κ in (214), and (223) follows since $L \in E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}$ and since $d^C e^{-d\frac{\varepsilon'^2}{2}} = e^{-d\frac{\varepsilon^2}{2}}$.

It therefore remains to show that there exist $L \in \mathcal{L}_d$ and $S \in \text{Gr}_{d,k}(\mathbb{F}_p)$ for which L is in E_{c_0,c_1} and $E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}$ and $p^{\frac{k}{d}}L(S)$ is in E_{c_0,c_1}. To that end, let $L \sim \mu_d$ and let $S \sim \text{Uniform}(\text{Gr}_{d,k}(\mathbb{F}_p))$ be statistically independent of L. By [16, Proposition 2.2] we have that $p^{\frac{k}{d}}L(S) \sim \mu_d$. Thus,

$$\begin{aligned}&\Pr\left(L \in E_{c_0,c_1}, p^{\frac{k}{d}}L(S) \in E_{c_0,c_1}, L \in E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\right) \\ &\ge 1 - \Pr\left(L \notin E_{c_0,c_1}\right) - \Pr\left(p^{\frac{k}{d}}L(S) \notin E_{c_0,c_1}\right) - \Pr\left(L \notin E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\right) &(224)\\ &= 1 - 2\mu_d(\{L \notin E_{c_0,c_1}\}) - \mu_d(\{L \notin E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\}) &(225)\\ &> \frac{1}{3} - \mu_d(\{L \notin E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\}), &(226)\end{aligned}$$

where we have used the union bound in the first inequality and Lemma 5 in the last inequality. We will be able to use Lemma 7 with $r_b = D$ to deduce that $\mu_d(\{L \notin E_{\alpha,\beta,\kappa,\varepsilon'}^{P_{\tilde{U}},r_U,D}\}) < \frac{1}{3}$ and complete the proof, once we show that κ in (214) is greater than the right hand side of (212). To that end, we write

$$\begin{aligned}\frac{e^{\frac{\varepsilon'^2}{2}}\sqrt{1+\varepsilon'}\frac{\sqrt{d(\alpha^2 r_U + \beta^2 D)}}{r_{\rm eff}(1)}}{\frac{\sqrt{dD}}{(1+C\frac{\log d}{d})p^{-\frac{k}{d}} r_{\rm eff}(1)}} &= p^{-\frac{k}{d}}\sqrt{\alpha^2\frac{r_U}{D} + \beta^2}e^{\frac{\varepsilon'^2}{2}}\sqrt{1+\varepsilon'}(1 + C\frac{\log d}{d}) &(227)\\ &\le 2^{-\left(R - \frac{1}{2}\log(\beta^2 + \alpha^2\frac{r_U^2}{D})\right)} \cdot 2^{\frac{1}{d}} e^{\frac{\varepsilon'^2}{2}}\sqrt{1+\varepsilon'}(1 + C\frac{\log d}{d}) &(228)\\ &\le 2^{-C_1\left(\varepsilon + \frac{\log d}{d}\right)} \cdot 2^{\frac{1}{d}} e^{\frac{\varepsilon'^2}{2}}\sqrt{1+\varepsilon'}(1 + C\frac{\log d}{d}) &(229)\\ &< 1 &(230)\end{aligned}$$

where the last inequality holds for some universal C_1 large enough. ■

Proof of Lemma 6. Let $C_0 > 0$ be a universal constant satisfying

$$e^{\frac{1}{d}\ln(c_0 d^{c_1})} < 1 + C_0\frac{\log d}{d}, \quad \forall d. \tag{231}$$

Thus, for $L \in E_{c_0,c_1}$ we have

$$\frac{V_d r_{\rm cov}(L)^d}{V_d r_{\rm eff}(L)^d} < c_0 d^{c_1} \iff \frac{r_{\rm cov}(L)}{r_{\rm eff}(L)} < e^{\frac{1}{d}\ln(c_0 d^{c_1})} \implies \frac{r_{\rm cov}(L)}{r_{\rm eff}(L)} < 1 + C_0 \frac{\log d}{d} \tag{232}$$

Since $r_{\rm cov}(\kappa L) = \kappa r_{\rm cov}(L)$ and $r_{\rm eff}(\kappa L) = \kappa r_{\rm eff}(L)$, Item 1 holds for any $C \geq C_0$. Item 2 follows since $\sigma^2(\kappa L) \leq \frac{1}{d} r_{\rm cov}^2(\kappa L)$ for any lattice $L \subset \mathbb{R}^d$.

To prove Item 3, let f_Z and $f_{\tilde{Z}}$ be the densities of the random variables Z and $\tilde{Z}$, respectively. By Item 1 we have that for any $L \in E_{c_0,c_1}$ the support $\mathcal{V}_{\kappa L}$ of Z is contained in the support $\left(1 + C_0 \frac{\log d}{d}\right) \kappa r_{\rm eff}(L)\mathcal{B}$ of $\tilde{Z}$. Thus, for any $z \in \left(1 + C_0 \frac{\log d}{d}\right) \kappa r_{\rm eff}(L)\mathcal{B}$ we have that

$$f_Z(z) \leq \frac{\mathrm{Vol}\left(\left(1 + C_0 \frac{\log d}{d}\right) \kappa r_{\rm eff}(L)\mathcal{B}\right)}{\mathrm{Vol}\left(\mathcal{V}_{\kappa L}\right)} f_{\tilde{Z}}(z) \leq \left(1 + C_0 \frac{\log d}{d}\right)^d f_{\tilde{Z}}(z) \leq d^C f_{\tilde{Z}}(z). \tag{233}$$

It therefore follows that

$$f_{\alpha U + \beta Z}(x) \leq d^C f_{\alpha U + \beta \tilde{Z}}, \quad \forall x \in \mathbb{R}^d, \tag{234}$$

and therefore for any $\mathcal{A} \subset \mathbb{R}^d$

$$\Pr(\alpha U + \beta Z \notin \mathcal{A}) \leq d^C \Pr(\alpha U + \beta \tilde{Z} \notin \mathcal{A}). \tag{235}$$

We move on to proving Item 4. Let $\mathrm{eig}(R(\kappa L)) = (\lambda_1, \ldots, \lambda_d)$ be the eigenvalues of $R(\kappa L)$, such that $\sum_{i=1}^d \lambda_i = d\sigma^2(\kappa L)$, and $\|R(\kappa L)\|_F^2 = \sum_{i=1}^d \lambda_i^2$. Let $Z \sim \mathrm{Uniform}(\mathcal{V}_{\kappa L})$. Since $\mathrm{Vol}(\mathcal{V}_{\kappa L}) = V_d r_{\rm eff}^d(\kappa L)$, we have that

$$\log V_d r_{\rm eff}^d(\kappa L) = h(Z) \leq \frac{1}{2} \log\det\left((2\pi e) R(\kappa L)\right) = \sum_{i=1}^d \frac{1}{2} \log(2\pi e \lambda_i), \tag{236}$$

where the upper bound follows since the Gaussian distribution maximizes entropy under covariance constraints. Thus,

$$\sum_{i=1}^d \log(\lambda_i) \geq 2 \log V_d r_{\rm eff}^d(\kappa L) - d \log(2\pi e) \tag{237}$$

$$= 2 \log \left(V_d^{2/d} r_{\rm eff}^2(\kappa L)\right)^{d/2} - d \log(2\pi e) \tag{238}$$

$$= d \log \left(\frac{V_d^{2/d} r_{\rm eff}^2(\kappa L)}{2\pi e}\right) \tag{239}$$

$$= d \log \left(\frac{1}{2\pi e} \frac{V_d^{2/d} r_{\rm eff}^2(\kappa L)}{\sigma^2(\kappa L)} \sigma^2(\kappa L)\right) \tag{240}$$

$$= d \log \sigma^2(\kappa L) - d \log 2\pi e \cdot N(\kappa L), \tag{241}$$

where $N(\kappa L) = \frac{\sigma^2(\kappa L)}{V_d^{2/d} r_{\rm eff}^2(\kappa L)}$ is the normalized second moment (NSM) of a lattice κL in $\mathbb{R}^d$.

Denote $\delta(\kappa L) = 2\pi e \cdot G(\kappa L) - 1$, and $\|R(\kappa L)\|_2 = \max_{i=1,\ldots,n} \lambda_i$. We now show that

$$\frac{1}{d} \sum_{i=1}^d \lambda_i^2 \leq (\sigma^2(\kappa L))^2 + 2\delta(\kappa L) \|R(\kappa L)\|_2^2 \tag{242}$$

Indeed, in the range $0 < x \leq \|R(\kappa L)\|_2$ the second derivative of $x \mapsto \ln x$ is upper-bounded by $-\frac{1}{\|R(\kappa L)\|_2^2}$. Thus, from Taylor's expansion around $x = \sigma^2(\kappa L)$ we have

$$\ln \lambda_i \leq \ln \sigma^2(\kappa L) + \frac{\lambda_i - \sigma^2(\kappa L)}{\sigma^2(\kappa L)} - \frac{1}{2\|R(\kappa L)\|_2^2} (\lambda_i - \sigma^2(\kappa L))^2 .$$

Summing over i and using the facts that a) $\frac{1}{d}\sum_i \lambda_i = \sigma^2(\kappa L)$ and b) $\frac{1}{d}\sum_{i=1}^d \ln\lambda_i \geq \ln\sigma^2(\kappa L) - \ln(1+\delta(\kappa L))$ we get after rearranging terms

$$2\|R(\kappa L)\|_2^2 \ln(1+\delta(\kappa L)) \geq \frac{1}{d}\sum_i (\lambda_i - \sigma^2(\kappa L))^2 = \frac{1}{d}\sum_i \lambda_i^2 - (\sigma^2(\kappa L))^2\,,$$

completing the proof of (242).

To complete our statement, it remains to show that $\delta(\kappa L) \leq c_2\frac{\log d}{d}$ and $\|R(\kappa L)\|_2 \leq c_3 \log d \frac{r_{\rm cov}^2(\kappa L)}{d}$, for some universal constants $c_2, c_3 > 0$. This will imply, by (242), that

$$\frac{1}{d}\|R(\kappa L)\|_F^2 \leq (\sigma^2(\kappa L))^2 + 2c_2c_3^2\frac{\log d}{d}\cdot\log^2(d)\cdot\left(\frac{r_{\rm cov}^2(\kappa L)}{d}\right)^2 \tag{243}$$

$$\leq \left(\frac{1}{d}\left(\left(1+C\frac{\log d}{d}\right)\kappa r_{\rm eff}(L)\right)^2\right)^2 \cdot \left(1+2c_2c_3^2\frac{\log^3 d}{d}\right), \tag{244}$$

where we have used Items 1 and 2 in the last inequality.

For bounding $\delta(\kappa L)$, we use $\sigma^2(\kappa L) \leq \frac{1}{d}r_{\rm cov}^2(\kappa L)$ and write

$$2\pi e\cdot N(\kappa L) \leq \frac{2\pi e}{dV_d^{2/d}}\frac{r_{\rm cov}^2(\kappa L)}{r_{\rm eff}^2(\kappa L)} \leq \frac{2\pi e}{dV_d^{2/d}}\left(c_0 d^{c_1}\right)^{2/d} \leq 1 + c_2\frac{\log d}{d}, \tag{245}$$

where in the first inequality follows since $L \in E_{c_0,c_1}$, and in the second inequality we have used the fact that $\frac{2\pi e}{dV_d^{2/d}} = 1 + O(\frac{\log d}{d})$.

We now upper bound the operator norm $\|R(\kappa L)\|_2 = \max_{v\in\mathbb{S}^{n-1}}\mathbb{E}(v^\top Z)^2$, for $Z \sim \text{Uniform}(\mathcal{V}_{\kappa L})$. To that end, let $\delta_d = \sqrt{\frac{2(C+1)\log d}{d}}$. For any $v \in \mathbb{S}^{n-1}$ we have that

$$\mathbb{E}(v^\top Z)^2 = \mathbb{E}\left[\|v\|^2\|\|Z\|^2\cos^2(\angle(v,Z))\right] \tag{246}$$

$$\leq r_{\rm cov}^2(\kappa L)\mathbb{E}\left[\cos^2(\angle(v,Z))\right] \tag{247}$$

$$\leq r_{\rm cov}^2(\kappa L)\left(\delta_d^2 + \Pr(|\cos(\angle(v,Z)| > \delta_d)\right) \tag{248}$$

Let $\mathcal{A} = \{z \in \mathbb{R}^d \;:\; |v^\top z| \leq \delta_d\}$, and $\tilde{Z} \sim \text{Uniform}\left(\left((1+C\frac{\log d}{d}\right)\kappa r_{\rm eff}(L)\mathcal{B}\right)$. From Item 3, applied with $\alpha = 0$ and $\beta = 1$, we have

$$\Pr(|\cos(\angle(v,Z)| > \delta_d) = \Pr(Z \notin \mathcal{A}) \leq d^C\Pr(\tilde{Z}\notin\mathcal{A}) = d^C\Pr(|\cos(\angle(v,\tilde{Z})| > \delta_d) \tag{249}$$

$$= d^C\Pr(|\cos(\angle(v,Z_B)| > \delta_d), \tag{250}$$

where $Z_B \sim \text{Uniform}(\mathcal{B})$. Thus,

$$\mathbb{E}(v^\top Z)^2 \leq r_{\rm cov}^2(\kappa L)\left(\delta_d^2 + d^C\Pr(|\cos(\angle(v,Z_B)| > \delta_d)\right) \tag{251}$$

$$\leq r_{\rm cov}^2(\kappa L)\left(\frac{2(C+1)\log d}{d} + d^C\Pr\left(|\cos(\angle(v,Z_B)|) > \sqrt{\frac{2(C+1)\log d}{d}}\right)\right) \tag{252}$$

$$\leq c_3\log d\cdot\frac{r_{\rm cov}^2(\kappa L)}{d}, \tag{253}$$

where (253) follows since $\Pr\left(|\cos(\angle(v,Z_B)|) > \sqrt{\frac{2(C+1)\log d}{d}}\right) \leq e^{-\frac{2(C+1)\log d}{2}} = d^{-(C+1)}$, which follows from the fact that a spherical cap of height $1-\varepsilon$ has volume (w.r.t. to $\mathbb{S}^{n-1}$) at most $e^{-d\varepsilon^2/2}$ for $0 < \varepsilon < \frac{1}{\sqrt{2}}$ [77, Section 7.2]. ■

Proof of Lemma 7. For $r > 0, \kappa > 0$ and $x \in \mathbb{R}^d$, let

$$N^*(\kappa L, r\mathcal{B}, x) = |((\kappa L \setminus \{0\}) + x) \cap r\mathcal{B})|\,. \tag{254}$$

Note that for any $r > 0$ we have the inclusion of events

$$\{x \in r\mathcal{B}, N^*(\kappa L, r\mathcal{B}, -x) = 0\} \subset \{x \in \mathcal{V}_{\kappa L}\}, \tag{255}$$

which implies that

$$\{x \notin \mathcal{V}_{\kappa L}\} \subset \{x \notin r\mathcal{B}\} \cup \{N^*(\Lambda_{\kappa L}, r\mathcal{B}, -x) > 0\}. \tag{256}$$

Let $X = \alpha U + \beta \tilde{Z}$. For any given lattice κL, we therefore have that

$$P_e(L) = \Pr(X \notin \mathcal{V}_{\kappa L}) \leq \Pr(X \notin r\mathcal{B}) + \Pr\left(N^*(\kappa L, r\mathcal{B}, -X) > 0\right) \tag{257}$$

$$\leq \Pr(X \notin r\mathcal{B}) + \mathbb{E}_X[N^*(\kappa L, r\mathcal{B}, -X)], \tag{258}$$

where the last inequality follows since $\Pr(N > 0) \leq \mathbb{E}[N]$ for a random variable N supported on the non-negative integers. Taking the expectation with respect to $L \sim \mu_n$ gives

$$\mathbb{E}[P_e(L)] \leq \Pr(X \notin r\mathcal{B}) + \mathbb{E}_L \mathbb{E}_X[N^*(\kappa L, r\mathcal{B}, -X)], \tag{259}$$

Applying Siegel's summation formula, we have

$$\mathbb{E}_{\kappa L}\mathbb{E}_X[N^*(\kappa L, r\mathcal{B}, -X)] = \mathbb{E}_X[\mathbb{E}_{\kappa L}[N^*(\kappa L, r\mathcal{B}, -x)|X = x]] = \frac{\mathrm{Vol}(r\mathcal{B})}{\mathrm{covol}(\kappa L)} = \left(\frac{r}{\kappa r_{\mathrm{eff}}(1)}\right)^d, \tag{260}$$

so that

$$\mathbb{E}[P_e(L)] \leq \Pr(X \notin r\mathcal{B}) + \left(\frac{r}{\kappa r_{\mathrm{eff}}(1)}\right)^d. \tag{261}$$

Let $r^2 = d(\alpha^2 r_U + \beta^2 r_b)(1+\varepsilon)$, for $0 < \varepsilon < \frac{1}{\sqrt{2}}$ and let us upper bound the first term. Recalling that $X = \alpha U + \beta\tilde{Z}$, we have

$$\Pr(X \notin r\mathcal{B}) = \Pr((\alpha U + \beta\tilde{Z})^2 > r^2) \tag{262}$$

$$= \Pr(\alpha^2\|U\|^2 + \beta^2\|\tilde{Z}\|^2 + 2\alpha\beta U^\top Z > r^2) \tag{263}$$

$$\leq \Pr\left(d(\alpha^2 r_U + \beta^2 r_b) + 2\alpha\beta U^\top \tilde{Z} > d(\alpha^2 r_U + \beta^2 r_b)(1+\varepsilon)\right) \tag{264}$$

$$= \Pr\left(2\alpha\beta U^\top \tilde{Z} > d(\alpha^2 r_U + \beta^2 r_b)\varepsilon\right) \tag{265}$$

$$\leq \Pr\left(2\alpha\beta d\sqrt{r_U r_b}\cos(\angle(U, \tilde{Z})) > d(\alpha^2 r_U + \beta^2 r_b)\varepsilon\right) \tag{266}$$

$$= \Pr\left(\cos(\angle(U, \tilde{Z})) > \frac{d(\alpha^2 r_U + \beta^2 r_b)}{2d\alpha\beta\sqrt{r_U r_b}}\varepsilon\right) \tag{267}$$

$$\leq \Pr\left(\cos(\angle(U, \tilde{Z})) > \varepsilon\right) \tag{268}$$

$$\leq e^{-d\varepsilon^2/2} \tag{269}$$

where (268) follows from $(\alpha\sqrt{r_u} - \beta\sqrt{r_b})^2 \geq 0$, and the last inequality follows from the fact that a spherical cap of height $1-\varepsilon$ has volume (w.r.t. to $\mathbb{S}^{n-1}$) at most $e^{-d\varepsilon^2/2}$ for $0 < \varepsilon < \frac{1}{\sqrt{2}}$ [77, Section 7.2]. We have therefore obtained that

$$\mathbb{E}[P_e(L)] \leq e^{-d\varepsilon^2/2} + e^{-d\log\frac{\kappa r_{\mathrm{eff}}(1)}{r}}. \tag{270}$$

Taking

$$\kappa > e^{\frac{\varepsilon^2}{2}}\frac{r}{r_{\mathrm{eff}}(1)} = (1+\varepsilon)e^{\frac{\varepsilon^2}{2}}\frac{\sqrt{d(\alpha^2 r_U + \beta^2 r_b)}}{r_{\mathrm{eff}}(1)}, \tag{271}$$

gives

$$\mathbb{E}[P_e(L)] \leq 2e^{-d\varepsilon^2/2}. \tag{272}$$

Thus, using Markov's inequality, we obtain the claimed result. ■

APPENDIX C
BOUNDING THE EFFECT OF OVERLOAD EVENTS

Proof of Proposition 4. Let

$$e_{\rm OL} = \hat{U}_{[d]}^\top \hat{V}_{[d]} - \hat{U}_{[d],\rm ideal}^\top \hat{V}_{[d],\rm ideal}, \tag{273}$$

such that

$$e = e_{\rm ideal} + \alpha e_{\rm OL}. \tag{274}$$

With probability 1, we have that

$$|e_{\rm OL}| < |\hat{U}_{[d]}^\top \hat{V}_{[d]}| + |\hat{U}_{[d],\rm ideal}^\top \hat{V}_{[d],\rm ideal}| \le \|\hat{U}_{[d]}\| \cdot \|\hat{V}_{[d]}\| + \|\hat{U}_{[d],\rm ideal}\| \cdot \|\hat{V}_{[d],\rm ideal}\| \tag{275}$$

$$\le r_{\rm cov}^2(\Lambda_c) + (\sqrt{n} + r_{\rm cov}(\Lambda_f))^2, \tag{276}$$

where the last inequality follows since $\hat{U}_{[d]}, \hat{V}_{[d]} \in \mathcal{V}_{\Lambda_c}$ by definition, and since

$$\|\hat{U}_{[d],\rm ideal}\| = \|U_{[d]} + Z_1\| \le \|U_{[d]}\| + \|Z_1\| \le \|U\| + \|Z_1\| \le \sqrt{n} + r_{\rm cov}(\Lambda_f), \tag{277}$$

and $\|\hat{V}_{[d],\rm ideal}\|$ is bounded similarly. With probability 1, we also have that

$$|e_{\rm ideal}| = |\alpha \hat{U}_{[d],\rm ideal}^\top \hat{V}_{[d],\rm ideal} - U^\top V| \le \alpha |\hat{U}_{[d],\rm ideal}^\top \hat{V}_{[d],\rm ideal}| + |U^\top V| \le \alpha \|\hat{U}_{[d],\rm ideal}\| \cdot \|\hat{V}_{[d],\rm ideal}\| + |\rho| n \tag{278}$$

$$\le \alpha(\sqrt{n} + r_{\rm cov}(\Lambda_f))^2 + |\rho| n \tag{279}$$

$$\le (\sqrt{n} + r_{\rm cov}(\Lambda_f))^2 + n, \tag{280}$$

where (279) follows from (277), and (280) from $0 \le \alpha, |\rho| \le 1$.

Note that $\max\{\sqrt{n}, r_{\rm cov}(\Lambda_f)\} \le M(\Lambda_c)$, since $\Lambda_c \subset \Lambda_f$, and therefore $\mathcal{V}_{\Lambda_f} \subset \mathcal{V}_{\Lambda_c}$. It therefore follows that with probability 1 we have

$$|e_{\rm OL}| \le 5M^2(\Lambda_c), \quad |e_{\rm ideal}| \le 5M^2(\Lambda_c). \tag{281}$$

Consequently,

$$\mathbb{E}(e^2) = \mathbb{E}(e_{\rm ideal}^2) + \alpha^2 \mathbb{E}(e_{\rm OL}^2) + 2\alpha \mathbb{E}[e_{OL} e_{\rm ideal}] \tag{282}$$

$$\le \mathbb{E}(e_{\rm ideal}^2) + 75 \Pr({\rm OL}) M^4(\Lambda_c), \tag{283}$$

as claimed, where we have used $0 \le \alpha \le 1$ again. ■

APPENDIX D
PROJECTIONS OF RANDOM UNIFORM ORTHOGONAL VECTORS

Recall that $S \sim \mathrm{Uniform}(\mathrm{O}_n(\mathbb{R}))$ and we denote $U = \sqrt{n} S_1$, $Z = \sqrt{n} S_2$ and $V = \rho U + \sqrt{1-\rho^2} Z$. Furthermore, $d = \lfloor \kappa n \rfloor$, and we denote $U_{[d]} = (U_1, \ldots, U_d)^\top$ and similarly $V_{[d]} = \sqrt{\rho} U_{[d]} + \sqrt{1-\rho^2} Z_{[d]}$. We have

$$\mathbb{E}[(U_{[d]}^\top V_{[d]})^2] = \mathbb{E}[(\rho \|U_{[d]}\|^2 + \sqrt{1-\rho^2} U_{[d]}^\top Z_{[d]})^2] = \rho^2 \mathbb{E}\|U_{[d]}\|^4 + (1-\rho^2)\mathbb{E}[(U_{[d]}^\top Z_{[d]})^2] \tag{284}$$

where the last equation follows since $\mathbb{E}[\|U_{[d]}\|^2 U_{[d]}^\top Z_{[d]}] = 0$ from symmetry. For $d = n$ we trivially have $\mathbb{E}\|U\|^4 = n^2$ and then $\mathbb{E}[(U^\top Z)^2] = 0$. We proceed to compute $\mathbb{E}\|U_{[d]}\|^4$ and then $\mathbb{E}[(U_{[d]}^\top Z_{[d]})^2]$ for general $d \le n$.

It holds that (using the fact that $\mathbb{E}(U_i^2) = 1$)

$$\mathbb{E}\|U_{[d]}\|^2 = d. \tag{285}$$

By symmetry, we also have

$$\mathbb{E}(U_{[d]}^\top Z_{[d]}) = 0. \tag{286}$$

We will further use the fact that $\mathbb{E}(U_i^4) = \frac{3n}{n+2}$ [78]. To compute $\mathbb{E}\|U_{[d]}\|^4$, we first note that, by symmetry

$$n^2 = \mathbb{E}\|U\|^4 = \mathbb{E}\left(\sum_{i=1}^n U_i^2\right)^2 = n\mathbb{E}(U_1^4) + n(n-1)\mathbb{E}(U_1^2 U_2^2), \tag{287}$$

which implies

$$\mathbb{E}(U_1^2 U_2^2) = \frac{n - \mathbb{E}(U_1^4)}{n-1} = \frac{n}{n+2}. \tag{288}$$

With this, we can write

$$\mathbb{E}\|U_{[d]}\|^4 = \mathbb{E}\left(\sum_{i=1}^d U_i^2\right)^2 = d\mathbb{E}(U_1^4) + d(d-1)\mathbb{E}(U_1^2 U_2^2) = \frac{n}{n+2} d(d+2). \tag{289}$$

We move on to calculate $\mathbb{E}(U_{[d]}^\top Z_{[d]})^2$. We have that

$$\mathbb{E}(U_{[d]}^\top Z_{[d]})^2 = \mathbb{E}\left(\sum_{i=1}^d U_i Z_i\right)^2 = \sum_{i=1}^d \mathbb{E}(U_i^2 Z_i^2) + \sum_{j \neq i} \mathbb{E}(U_i U_j Z_i Z_j) = d\xi + d(d-1)\nu, \tag{290}$$

where

$$\xi = \mathbb{E}(U_1^2 Z_1^2), \quad \nu = \mathbb{E}(U_1 U_2 Z_1 Z_2), \tag{291}$$

and the last equality in (290) follows by symmetry. Taking $d = n$, we get that $U_{[n]}^\top Z_{[n]} = U^\top Z = 0$ w.p. 1. Invoking (290) therefore gives

$$0 = n\xi + n(n-1)\nu \Longrightarrow \nu = -\frac{\xi}{n-1}. \tag{292}$$

Substituting this into (290), we obtain

$$\mathbb{E}(U_{[d]}^\top Z_{[d]})^2 = d\left(1 - \frac{d-1}{n-1}\right)\xi = \frac{d(n-d)}{n-1}\xi. \tag{293}$$

In order to compute ξ, define $e = U - Z$. Note that the symmetry and orthogonality of U and Z implies that $e \sim \mathrm{Uniform}(\sqrt{2n}\mathbb{S}^{n-1})$, where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. It therefore follows that

$$\mathbb{E}(e_1^4) = 4\mathbb{E}(U_1^4). \tag{294}$$

On the other hand

$$\mathbb{E}(e_1^4) = \mathbb{E}(U_1 - Z_1)^4 = \sum_{i=0}^4 \binom{4}{i} \mathbb{E}(U_1^i Z_1^{4-i}) = \mathbb{E}(U_1^4) + \mathbb{E}(Z_1^4) + 6\mathbb{E}(U_1^2 Z_1^2) + 4\mathbb{E}(U_1 Z_1^3) + 4\mathbb{E}(U_1^3 Z_1). \tag{295}$$

By symmetry, we clearly have that $\mathbb{E}(Z_1^4) = \mathbb{E}(U_1^4)$. We claim that $\mathbb{E}(U_1 Z_1^3) = 0$. To see this, note that given Z, the distribution of U is invariant to negation (in other words $p_{U|Z=z}(u) = p_{U|Z=z}(-u)$). By symmetry, this also implies that $\mathbb{E}(U_1^3 Z_1) = 0$. We therefore have that

$$4\mathbb{E}(U_1^4) = \mathbb{E}(e_1^4) = 2\mathbb{E}(U_1^4) + 6\xi \tag{296}$$

$$\Longrightarrow \xi = \frac{\mathbb{E}(U_1^4)}{3} = \frac{n}{n+2}. \tag{297}$$

Substituting this into (293), we obtain

$$\mathbb{E}(U_{[d]}^\top Z_{[d]})^2 = d \cdot \frac{n(n-d)}{(n+2)(n-1)}. \tag{298}$$

Consequently, by (284), (289) and (298), we have

$$\mathbb{E}[(U_{[d]}^\top V_{[d]})^2] = \frac{nd}{n+2}\left(\rho^2(d+2) + (1-\rho^2)\frac{n-d}{n-1}\right) \tag{299}$$

$$= \frac{nd}{n+2}\left(\rho^2 \frac{n(d+1)-2}{n-1} + \frac{n-d}{n-1}\right) \tag{300}$$

$$\leq \rho^2 n \frac{d(d+1)}{n} + \frac{d(n-d)}{n}, \tag{301}$$

where in the last inequality we have used the fact that $\frac{n}{(n-1)(n+2)} \leq \frac{1}{n}$ for all $n \geq 2$. This establishes our claim.

APPENDIX E
TAIL PROBABILITY OF $U_{[d]}$

Proposition 5: Let U be uniformly distributed on $\sqrt{n}\mathbb{S}^{n-1}$, and let $U_{[d]} = (U_1, \ldots, U_d)^\top$ be its projection on the first $1 \le d \le n$ coordinates. Then, for any $0 < \varepsilon < 1$

$$\Pr(\|U_{[d]}\|^2 > (1+\varepsilon)d) \le 2e^{-\left(\frac{\varepsilon}{1+\varepsilon}\right)^2 \frac{1}{24} d} \le 2e^{-\frac{\varepsilon^2}{96} d} \tag{302}$$

Proof. Let $Z \sim \mathcal{N}(0, I_n)$ and note that U has the same distribution as $\sqrt{n}\frac{Z}{\|Z\|}$. Let

$$X_1 = \sum_{i=1}^{d} Z_i^2, \tag{303}$$

$$X_2 = \sum_{i=d+1}^{n} Z_i^2, \tag{304}$$

and note that X_1 and X_2 are independent chi-squared random variables with d and $n-d$ degrees of freedom, respectively. We therefore have

$$\Pr\left(\|U_{[d]}\|^2 > (1+\varepsilon)d\right) = \Pr\left(n\frac{X_1}{X_1+X_2} \ge (1+\varepsilon)d\right) = \Pr\left((n-(1+\varepsilon)d)X_1 \ge (1+\varepsilon)dX_2\right) \tag{305}$$

$$= \Pr\left(\frac{1}{n-d}X_2 \le \left(\frac{1}{1+\varepsilon} - \frac{d}{n-d}\frac{\varepsilon}{1+\varepsilon}\right)\frac{1}{d}X_1\right) \tag{306}$$

$$= \Pr\left(\frac{1}{n-d}X_2 \le \left(1 - \frac{\varepsilon}{1+\varepsilon}\left(1 + \frac{d}{n-d}\right)\right)\frac{1}{d}X_1\right) \tag{307}$$

$$= \Pr\left(\frac{1}{n-d}X_2 \le \left(1 - \frac{\varepsilon}{1+\varepsilon}\frac{n}{n-d}\right)\frac{1}{d}X_1\right). \tag{308}$$

Note that for $\frac{\varepsilon}{1+\varepsilon} \cdot \frac{n}{n-d} \ge 1$ the probability above is zero. Thus, for the remainder of the proof, we assume $\frac{\varepsilon}{1+\varepsilon} \cdot \frac{n}{n-d} < 1$. For any $t > 0$ we have that

$$\Pr\left(\|U_{[d]}\|^2 > (1+\varepsilon)d\right) \le \Pr\left(\frac{1}{d}X_1 > t\right) + \Pr\left(\frac{1}{n-d}X_2 \le \left(1 - \frac{\varepsilon}{1+\varepsilon}\frac{n}{n-d}\right)t\right). \tag{309}$$

Let $t = 1+\delta$ for some $0 < \delta < 1$, and let

$$\delta' = 1 - \left(1 - \frac{\varepsilon}{1+\varepsilon}\frac{n}{n-d}\right)(1+\delta). \tag{310}$$

By standard Chenroff bounds on the tail of the chi-squared distribution

$$\Pr\left(\frac{1}{d}X_1 > t\right) = \Pr\left(\frac{1}{d}X_1 > 1+\delta\right) \le \exp\left\{\frac{d}{2}\left(\ln(1+\delta) - \delta\right)\right\} \le \exp\left\{-\frac{d\delta^2}{16}\right\}, \tag{311}$$

where we have used $\ln(1+\delta) - \delta < \frac{\delta^2}{8}$ for all $0 < \delta < 1$. Similarly, if $0 < \delta' < 1$ we have

$$\Pr\left(\frac{1}{n-d}X_2 < \left(1 - \frac{\varepsilon}{1+\varepsilon}\frac{n}{n-d}\right)t\right) = \Pr\left(\frac{1}{n-d}X_2 < 1-\delta'\right) \le \exp\left\{\frac{n-d}{2}\left(\ln(1-\delta') + \delta'\right)\right\} \tag{312}$$

$$\le \exp\left\{-\frac{(n-d)\delta'^2}{16}\right\}. \tag{313}$$

We will choose

$$\delta = \varepsilon'\frac{1+\eta^2}{1+\eta-\varepsilon'(1+\eta^2)}, \quad \varepsilon' = \frac{\varepsilon}{1+\varepsilon}, \quad \eta = \sqrt{\frac{d}{n-d}}, \tag{314}$$

such that

$$\frac{d\delta^2}{16} = \frac{(n-d)\delta'^2}{16} \tag{315}$$

and $\delta, \delta' > 0$. Note further, that

$$\delta \geq \varepsilon' \frac{1+\eta^2}{1+\eta} \geq \varepsilon' 2(\sqrt{2}-1), \tag{316}$$

and we therefore have that

$$\frac{(n-d)\delta'^2}{16} = \frac{d\delta^2}{16} \geq d\varepsilon'^2 \frac{(\sqrt{2}-1)^2}{4} \geq \frac{d\varepsilon'^2}{24}, \tag{317}$$

yielding the claimed result. ■

REFERENCES

[1] V. Strassen, "Gaussian elimination is not optimal," *Numerische mathematik*, vol. 13, no. 4, pp. 354–356, 1969.

[2] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz *et al.*, "Discovering faster matrix multiplication algorithms with reinforcement learning," *Nature*, vol. 610, no. 7930, pp. 47–53, 2022.

[3] R. Duan, H. Wu, and R. Zhou, "Faster matrix multiplication via asymmetric hashing," in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2023, pp. 2129–2138.

[4] V. V. Williams, Y. Xu, Z. Xu, and R. Zhou, *New Bounds for Matrix Multiplication: from Alpha to Omega*, 2024, pp. 3792–3835.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] J. Conway and N. Sloane, "A fast encoding method for lattice codes and quantizers," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 820–824, 1983.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[11] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*. IEEE, 2014, pp. 10–14.

[12] Y. Polyanskiy and Y. Wu, *Information theory: From coding to learning*. Cambridge university press, 2024.

[13] U. Erez and R. Zamir, "Achieving 1/2 log (1+ snr) on the AWGN channel with lattice encoding and decoding," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2293–2314, 2004.

[14] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, 2014.

[15] O. Ordentlich and U. Erez, "A simple proof for the existence of "good" pairs of nested lattices," *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4439–4453, 2016.

[16] O. Ordentlich, O. Regev, and B. Weiss, "New bounds on the density of lattice coverings," *Journal of the American Mathematical Society*, vol. 35, no. 1, pp. 295–308, 2022.

[17] 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.

[18] M. W. Mahoney *et al.*, "Randomized algorithms for matrices and data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[19] P.-G. Martinsson and J. A. Tropp, "Randomized numerical linear algebra: Foundations and algorithms," *Acta Numerica*, vol. 29, pp. 403–572, 2020.

[20] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast monte carlo algorithms for matrices i: Approximating matrix multiplication," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 132–157, 2006.

[21] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 2002, pp. 380–388.

[22] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, 2004, pp. 253–262.

[23] R. Pagh, "Compressed matrix multiplication," *ACM Transactions on Computation Theory (TOCT)*, vol. 5, no. 3, pp. 1–17, 2013.

[24] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," *Advances in neural information processing systems*, vol. 29, 2016.

[25] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *Journal of Machine Learning Research*, vol. 18, no. 187, pp. 1–30, 2018.

[26] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.

[27] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[28] 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.

[29] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 168–27 183, 2022.

[30] 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.

[31] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The era of 1-bit llms: All large language models are in 1.58 bits," *arXiv preprint arXiv:2402.17764*, 2024.

[32] 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

[33] A. Tseng, Q. Sun, D. Hou, and C. De Sa, "Qtip: Quantization with trellises and incoherence processing," *arXiv preprint arXiv:2406.11235*, 2024.

[34] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[35] D. Blalock and J. Guttag, "Multiplying matrices without multiplying," in *International Conference on Machine Learning*. PMLR, 2021, pp. 992–1004.

[36] P. Stock, A. Joulin, R. Gribonval, B. Graham, and H. Jégou, "And the bit goes down: Revisiting the quantization of neural networks," *arXiv preprint arXiv:1907.05686*, 2019.

[37] R. Guo, S. Kumar, K. Choromanski, and D. Simcha, "Quantization based fast inner product search," in *Artificial intelligence and statistics*. PMLR, 2016, pp. 482–490.

[38] 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.

[39] D. Malak, "Distributed structured matrix multiplication," *arXiv preprint arXiv:2405.02904*, 2024.

[40] D. Krithivasan and S. S. Pradhan, "Lattices for distributed source coding: Jointly gaussian sources and reconstruction of a linear function," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5628–5651, 2009.

[41] A. B. Wagner, "On distributed compression of linear functions," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 79–94, 2010.

[42] A. Ingber, T. Courtade, and T. Weissman, "Compression for quadratic similarity queries," *IEEE transactions on information theory*, vol. 61, no. 5, pp. 2729–2747, 2015.

[43] I. Ochoa, A. Ingber, and T. Weissman, "Compression schemes for similarity queries," in *2014 Data Compression Conference*, 2014, pp. 332–341.

[44] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Straggler mitigation in distributed matrix multiplication: Fundamental limits and optimal coding," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1920–1933, 2020.

[45] A. Adler, J. Tang, and Y. Polyanskiy, "Efficient representation of large-alphabet probability distributions," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 4, pp. 651–663, 2022.

[46] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[47] A. Kipnis and G. Reeves, "Gaussian approximation of quantization error for estimation from compressed data," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5562–5579, 2021.

[48] U. Erez, S. Litsyn, and R. Zamir, "Lattices which are good for (almost) everything," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3401–3416, 2005.

[49] D. Krithivasan and S. S. Pradhan, "A proof of the existence of good nested lattices," *submitted to the UM CSPL technical reports series*, 2007.

[50] B. D. Kudryashov and K. V. Yurkov, "Random coding bound for the second moment of multidimensional lattices," *Problems of Information Transmission*, vol. 43, no. 1, pp. 57–68, 2007.

[51] C. Ling and J.-C. Belfiore, "Achieving awgn channel capacity with lattice gaussian coding," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 5918–5929, 2014.

[52] C. Ling, L. Luzzi, J.-C. Belfiore, and D. Stehlé, "Semantically secure lattice codes for the gaussian wiretap channel," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6399–6416, 2014.

[53] B. Nazer, V. R. Cadambe, V. Ntranos, and G. Caire, "Expanding the compute-and-forward framework: Unequal powers, signal levels, and multiple linear combinations," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 4879–4909, 2016.

[54] A. Campello, C. Ling, and J.-C. Belfiore, "Algebraic lattices achieving the capacity of the ergodic fading channel," in *2016 IEEE Information Theory Workshop (ITW)*. IEEE, 2016, pp. 459–463.

[55] Y.-C. Huang and K. R. Narayanan, "Construction π_A and π_D lattices: Construction, goodness, and decoding algorithms," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5718–5733, 2017.

[56] N. Di Pietro, G. Zémor, and J. J. Boutros, "Lda lattices without dithering achieve capacity on the gaussian channel," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1561–1594, 2017.

[57] O. Ordentlich, O. Regev, and B. Weiss, "Bounds on the density of smooth lattice coverings," *arXiv preprint arXiv:2311.04644*, 2023.

[58] M. Sadeghi, R. Qi, C. Feng, H. Khodaiemehr, and Y.-C. Huang, "A simpler proof on the existence of good nested lattice codes over imaginary quadratic integers for awgn channel," *IEEE Open Journal of the Communications Society*, 2024.

[59] L. Liu, S. Lyu, C. Ling, and B. Bai, "On the quantization goodness of polar lattices," *arXiv preprint arXiv:2405.04051*, 2024.

[60] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1152–1159, 1996.

[61] E. Alvarez, O. Almog, E. Chung, S. Layton, D. Stosic, R. Krashinsky, and K. Aubrey, "Introducing nvfp4 for efficient and accurate low-precision inference," 2025. [Online]. Available: https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference

[62] G. Forney and G. Ungerboeck, "Modulation and coding for linear gaussian channels," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2384–2415, 1998.

[63] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag, 1988.

[64] E. Agrell and B. Allen, "On the best lattice quantizers," *IEEE Transactions on Information Theory*, vol. 69, no. 12, pp. 7650–7658, 2023.

[65] M. S. Viazovska, "The sphere packing problem in dimension 8," *Annals of mathematics*, pp. 991–1015, 2017.

[66] 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.

[67] H. Cohn and A. Kumar, "Optimality and uniqueness of the leech lattice among lattices," *Annals of mathematics*, pp. 1003–1050, 2009.

[68] H. Cohn, A. Kumar, S. Miller, D. Radchenko, and M. Viazovska, "The sphere packing problem in dimension 24," *Annals of mathematics*, vol. 185, no. 3, pp. 1017–1033, 2017.

[69] A. Vardy and Y. Be'ery, "Maximum likelihood decoding of the leech lattice," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1435–1444, 1993.

[70] A. Vardy, "Even more efficient bounded-distance decoding of the hexacode, the golay code, and the leech lattice," *IEEE Transactions on Information Theory*, vol. 41, no. 5, pp. 1495–1499, 1995.
[71] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1639–1642, 1999.
[72] J. Conway and N. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 211–226, 1982.
[73] O. Ordentlich, "The Voronoi spherical cdf for lattices and linear codes: New bounds for quantization and coding," *arXiv preprint arXiv:2506.19791*, 2025.
[74] N. Ailon and B. Chazelle, "The fast johnson–lindenstrauss transform and approximate nearest neighbors," *SIAM Journal on computing*, vol. 39, no. 1, pp. 302–322, 2009.
[75] E. Liberty, *Accelerated dense random projections*. Yale University New Haven, CT, 2009.
[76] J. A. Tropp, "Improved analysis of the subsampled randomized hadamard transform," *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.
[77] S. Boucheron, G. Lugosi, and O. Bousquet, "Concentration inequalities," in *Summer school on machine learning*. Springer, 2003, pp. 208–240.
[78] A. J. Stam, "Limit theorems for uniform distributions on spheres in high-dimensional euclidean spaces," *Journal of Applied probability*, vol. 19, no. 1, pp. 221–228, 1982.