Quantizing With Randomized Hadamard Transforms: Efficient Heuristic Now Proven

Ran Ben-Basat
University College London

William Kuszmaul
Carnegie Mellon University

Michael Mitzenmacher
Harvard University

Amit Portnoy
Independent Researcher

Shay Vargaftik
VMware Research by Broadcom

Abstract

Uniform random rotations (URRs) are a common preprocessing step in modern quantization approaches used for gradient compression, inference acceleration, KV-cache compression, model weight quantization, and approximate nearest-neighbor search in vector databases. In practice, URRs are often replaced by randomized Hadamard transforms (RHTs), which preserve orthogonality while admitting fast implementations. The remaining issue is the performance for worst-case inputs. With a URR, each coordinate is individually distributed as a shifted beta distribution, which converges to a Gaussian distribution in high dimensions. Generally, one RHT is not suitable in the worst case, as individual coordinates can be far from these distributions. We show that after composing two RHTs on any d-sized input vector, the marginal distribution of every fixed coordinate of the normalized rotated vector is within $\mathcal{O}(d^{-1/2})$ of a standard Gaussian both in Kolmogorov distance and in 1-Wasserstein distance. We then plug these bounds into the analyses of modern compression schemes, namely DRIVE and QUIC-FL, and show that two RHTs achieve performance that asymptotically matches URRs.

However, while two RHTs suffice for scalar quantization, they may be insufficient for Vector Quantization (VQ), which depends on the joint distribution across fixed-size blocks of coordinates (as opposed to only marginal distribution convergence for single coordinates). We prove that a composition of three RHTs leads to decaying coordinate covariance. This ensures that any fixed, bounded, multi-dimensional VQ codebook optimized for URRs has the same expected error when using three RHTs, up to an additive term that asymptotically vanishes with the dimension.

Finally, because practical inputs are rarely adversarial, we propose a linear-time $\mathcal{O}(d)$ check on the input's moments to dynamically adapt the number of RHTs used at runtime to improve performance.

1 Introduction

A foundational technique in quantization is applying a uniform random rotation (URR) to the input vectors beforehand, which ensures the vector's squared ℓ_2 norm is evenly distributed across coordinates in expectation. Specifically, for a URR $\mathcal{R}_U$ and input $x \in \mathbb{R}^d$, $\mathbb{E}[(\mathcal{R}_U x)_i^2] = \|x\|^2 / d$. Moreover, because individual coordinates of a rotated vector rapidly converge to a Gaussian distribution as the vector's dimension grows, rotations enable stable, nearly dimension-free scalar quantization designs. This robust statistical smoothing has recently found numerous applications, including Distributed Mean Estimation (DME) and gradient compression in federated learning (e.g., QUIC-FL [9]), parameter quantization (e.g., Quartet-2 via MS-EDEN [29]), database vector search

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(e.g., RaBitQ [21]), KV cache quantization (e.g., TurboQuant [40]), privacy (e.g., ScionFL [10]), and advanced post-training quantization frameworks (e.g., HIGGS [27]).

While URRs are conceptually and mathematically attractive, they take $O(d^2)$ operations to compute for d-dimensional vectors and thus are often too slow. This is why the literature repeatedly suggests replacing them with randomized Hadamard transforms (RHTs), which are structured random rotations that can be computed more efficiently, i.e., in $O(d \log d)$ operations. RHTs have been suggested for many quantization algorithms, e.g., [7, 8, 12, 20, 24, 26, 28, 32, 33, 36, 39, 40]. Furthermore, practitioners already use successive RHTs heuristically. For example, FAISS implements three RHTs as a fast rotation [18]. FALCONN's original cross-polytope LSH construction uses three RHTs [5], while Weaviate uses a related construction with blockwise Hadamard transforms and random permutations [15]. More broadly, RHTs also underlie fast Johnson-Lindenstrauss transforms [3, 4] and subsampled randomized Hadamard transforms (SRHTs) [31]. The algorithmic advantage is immediate: the transform is orthogonal, fast, memory efficient, and GPU-friendly.

At a theoretical level, however, the delicate issue is that RHTs are more difficult to work with, as they do not fully provide the same properties as a URR. A single RHT may not be suitable for worst-case inputs; for example, the unit vector $(1/\sqrt{2}, 1/\sqrt{2}, 0, 0, \ldots, 0)$ after an RHT will have all coordinates take on values in the set $\{0, \pm\sqrt{2/d}\}$, rather than a (near-)Gaussian coordinate law. Intuitively, even with a smoother initial vector, there can be a gap in terms of readily obtaining results for applications that utilize the coordinate marginal distribution of a random rotation.

In this paper, we show that, for scalar distributions, composing two successive RHTs eliminates such bad cases. Specifically, after one RHT, the resulting intermediate vector has a small expected third moment. Given this intermediate vector, each coordinate of the final vector becomes an ordinary Rademacher sum. We then apply the appropriate Berry–Esseen inequalities, yielding the necessary bounds for comparing these coordinate distributions to a Gaussian. We analyze this comparison using two metrics: the Kolmogorov distance, suited to thresholded and bounded-variation functionals, and the 1-Wasserstein distance, suited to Lipschitz functionals. With both estimates established, we then apply them to obtain new formal guarantees for recent scalar quantization schemes (which quantize each entry separately), namely DRIVE [37] and QUIC-FL [9]. We expect similar guarantees to be derivable for other RHT-based scalar quantizers.[1]

Additionally, we turn to the problem of Vector Quantization (VQ), where blocks of entries are jointly quantized. In VQ, having a marginal distribution that is approximately Gaussian is insufficient, as the joint block distribution determines the fixed-codebook error. We show that this can be achieved by composing three successive RHTs.

Finally, while composing RHTs restores strict worst-case guarantees, natural data distributions are rarely perfectly adversarial. A corollary of our general statements is that if an initial vector already exhibits suitably small values for specific norms, additional RHT steps are redundant. We establish that these requirements, specifically, the cubed ℓ_3 norm dictating scalar marginals, and the squared ℓ_∞ norm dictating multi-dimensional cross-correlation, can be explicitly computed in $\mathcal{O}(d)$ linear time. Thus, practical systems can dynamically check an input vector to determine whether 1, 2, or 3 RHTs suffice for the stated bounds, preserving empirical speed without sacrificing theoretical bounds.

Our presentation is modular: we state general results for composing RHTs before embarking on application-specific analysis. This keeps the core probabilistic statements separate from the downstream subtleties of each application, which we expect to provide an easier path for future applications.

Our core contributions can be summarized as follows:

- **Universal Scalar Marginals (2-RHT):** We prove that for every arbitrary input vector and every fixed coordinate, two RHTs provide a uniform $\mathcal{O}(d^{-1/2})$ normal additive approximation in both the Kolmogorov (d_K) and 1-Wasserstein (W_1) metrics.
- **Application of 2-RHT Bounds to Scalar Quantizers:**
  - **Formal Guarantees for DRIVE [37] (Table 1a):** We plug these scalar W_1 bounds into the analysis of DRIVE, recovering the URR-based guarantees for both the biased and the unbiased estimator up to a decaying $\mathcal{O}(d^{-1/2})$ term.

[1] For a scalar quantizer, expected squared reconstruction error is a sum of marginal expected errors; this equality does not require independence. DRIVE's scale and DME bias are treated separately in Section 4.

	(a) DRIVE's vNMSE (§ 4)		(b) QUIC-FL's guarantees (§5)	
	Biased Scale	Unbiased Scale	(E_1, E_2, E_3, E_4)	Sent frac.
URR	$1-\frac{2}{\pi}$	(**new**) $\frac{\pi}{2}-1$	$(1.520, 0.223, 0.044, 0.0098)$	p
1-RHT	0.5	none	$(4.831, 0.692, 0.131, 0.0272)$	$3.2{\cdot}p$
2-RHT (**new**)	$1-\frac{2}{\pi}+\mathcal{O}(\frac{1}{\sqrt{d}})$	$\frac{\pi}{2}-1+\mathcal{O}(\frac{1}{\sqrt{d}})$	$(1.520, 0.223, 0.044, 0.0098)+\mathcal{O}(\frac{1}{\sqrt{d}})$	$p+\mathcal{O}(\frac{1}{\sqrt{d}})$

Table 1: A summary of DRIVE's [37] and QUIC-FL's [9] guarantees with the known URR ($\mathcal{O}(d^2)$ time) and 1-RHT bounds ($\mathcal{O}(d\log d)$ time), and the new 2-RHT ($\mathcal{O}(d\log d)$ time) bounds. Each $\mathcal{O}(d^{-1/2})$ term is an additive bound. The implicit constants may differ between entries, including components of the QUIC-FL tuple. Also, see Appendix A.2 for a sharper URR-based unbiased DRIVE vNMSE bound with the fixed scale. In (b), E_b, for $b \in \{1,2,3,4\}$, denotes QUIC-FL's vNMSE at bit budget b.

  - **Formal Guarantees for QUIC-FL [9] (Table 1b):** We apply the d_K bound to restore the exact URR-based transmission bandwidth threshold up to an additive $\mathcal{O}(d^{-1/2})$ decaying term, eliminating the $3.2\times$ multiplicative penalty of a single RHT.
- **Decorrelation and VQ (3-RHT):** We prove that two RHTs may fail to preserve the Gaussian expected error of a fixed codebook, whereas a composition of three RHTs provides the weak dependence required for VQ. This allows any fixed, bounded URR or Gaussian codebook to yield the same expected-error limit, up to a decaying additive term.
- **Adaptive $\mathcal{O}(d)$ Linear-Time Check:** We translate our theoretical constraints into a simple linear-time method that checks sufficient conditions for using fewer RHTs based on the ℓ_3 and ℓ_∞ norms.

2 Notations and definitions

Randomized Hadamard Transform (RHT). Fix $d = 2^m$. Let H be the Sylvester-Hadamard matrix [23]. Thus, $H \in \{\pm 1\}^{d\times d}$ and $HH^\top = dI_d$. Additionally, let $\widetilde{H} = \frac{1}{\sqrt{d}}H$. Thus, $\widetilde{H}$ is orthogonal.

For $k \geq 1$, let $R_k = \widetilde{H}D_k\widetilde{H}D_{k-1}\cdots\widetilde{H}D_1$, where $D_1, D_2, \ldots, D_k$ are diagonal matrices with i.i.d. ± 1 Rademacher entries. We call R_k a composition of k RHTs. For convenience, we denote $R_0 = I_d$.

Kolmogorov distance. Given two distributions whose CDFs are P, Q, their Kolmogorov distance is defined as $d_K(P,Q) = \sup_{x\in\mathbb{R}} |P(x) - Q(x)|$.

1-Wasserstein distance (also known as the Earth Mover's Distance)**.** Utilizing the Kantorovich-Rubinstein duality [6, 13, 38], we use an alternative definition of the 1-Wasserstein distance given by $W_1(X,Y) = \sup_{f\in\text{Lip}(1)} \Big|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\Big|$.

Unit sphere. We denote by S^{d-1} the $(d-1)$-dimensional unit sphere, i.e., the set of all vectors $a \in \mathbb{R}^d$ such that $\|a\|_2 = 1$.

Coordinate. For a vector $a \in \mathbb{R}^d$, we use a_i and $(a)_i$ to denote its i'th coordinate.

Binary sign convention. We use $\text{sign}(z) = 1$ if $z \geq 0$ and $\text{sign}(z) = -1$ otherwise, applied coordinatewise to vectors.

Normalized input For an input $x \in \mathbb{R}^d \setminus \{0\}$, let $\widetilde{x} = \frac{x}{\|x\|_2} \in S^{d-1}$ be its unit direction vector.

3 Composing Two RHTs

We start by considering the setting of composing two RHTs, under the Kolmogorov and 1-Wasserstein distance metrics. Specifically, we are interested in the properties of R_2x for any $x \in \mathbb{R}^d \setminus \{0\}$.

3.1 Kolmogorov distance with two RHTs

Let $a \in S^{d-1}$ and define $\rho_3(a) = \sum_{i=0}^{d-1} |a_i|^3$, which satisfies $d^{-1/2} \leq \rho_3(a) \leq 1$. Intuitively, the ℓ_3 norm cubed $\rho_3(a)$ measures how "flat" the unit vector a is. If the coordinates of a all have size

about $d^{-1/2}$, then $\rho_3(a)$ is of order $d^{-1/2}$ while if one coordinate is large, then $\rho_3(a)$ is large; also, $\rho_3(a) \le \|a\|_\infty$. We use ρ_3 because it is the third-moment quantity in the Berry–Esseen bound. Thus, the analysis for two RHTs works by showing that the first RHT makes ρ_3 sufficiently small on average, and then that the second RHT ensures that an individual coordinate is close to a Gaussian-distributed random variable when $\rho_3(a)$ is small. We start with the following lemma.

Lemma 1 (Average ℓ_3 smoothing after one RHT)**.** *For every $\widetilde{x} \in S^{d-1}$ and a random sign diagonal D, it holds that, $\mathbb{E}\rho_3(\widetilde{H}D\widetilde{x}) \le \frac{C_3}{\sqrt{d}}$ where $C_3 = 3^{3/4} \approx 2.2795$.*

Proof. Let $a = \widetilde{H}D\widetilde{x}$. Thus, $a_1 = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} \widetilde{x}_i \varepsilon_i$, where the ε_i are i.i.d. Rademacher ± 1 signs (the signs of H are absorbed into the Rademacher signs). Since all coordinates have the same marginal distribution, $\mathbb{E}\rho_3(a) = d\,\mathbb{E}|a_1|^3$. Let $M = \sum_{i=0}^{d-1} \widetilde{x}_i \varepsilon_i$. Then $\mathbb{E}M^2 = \|\widetilde{x}\|_2^2 = 1$, and a direct fourth-moment calculation gives $\mathbb{E}M^4 = \sum_{i=0}^{d-1} \widetilde{x}_i^4 + 6\sum_{i<j} \widetilde{x}_i^2 \widetilde{x}_j^2 = 3\left(\sum_{i=0}^{d-1} \widetilde{x}_i^2\right)^2 - 2\sum_{i=0}^{d-1} \widetilde{x}_i^4 \le 3$.

By Hölder's inequality, $\mathbb{E}|M|^3 \le (\mathbb{E}M^4)^{3/4} \le 3^{3/4} = C_3$. Therefore,

$$\mathbb{E}\rho_3(a) = d\,\mathbb{E}|a_1|^3 = d \cdot \mathbb{E}|\frac{1}{\sqrt{d}} M|^3 = d \cdot d^{-3/2}\mathbb{E}|M|^3 = \frac{\mathbb{E}|M|^3}{\sqrt{d}} \le \frac{C_3}{\sqrt{d}}. \quad \square$$

We also utilize a well-known Berry–Esseen bound for non-i.i.d variables derived in [35, Theorem 7] which we restate here for clarity. See also Tyurin's published analysis [34].

Theorem 7 of [35] (Berry–Esseen for non-i.i.d variables). Let $X_0, \ldots, X_{d-1}$ be independent real-valued random variables with $\mathbb{E}X_i = 0$ and $\sum_{i=0}^{d-1} \mathbb{E}X_i^2 = 1$. If $G \sim N(0,1)$, then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sum_{i=0}^{d-1} X_i \le t \right) - \mathbb{P}(G \le t) \right| \le 0.5606 \sum_{i=0}^{d-1} \mathbb{E}|X_i|^3.$$

With these results at hand, we are now ready to prove the following theorem.

Theorem 1 (Kolmogorov distance with two RHTs)**.** *Let $x \in \mathbb{R}^d \setminus \{0\}$. Define $U(x) = \sqrt{d}\,(R_2\widetilde{x})_1$. Then $d_K\big(U(x), G\big) \le \frac{0.5606\,C_3}{\sqrt{d}} \le \frac{1.28}{\sqrt{d}}$, where $G \sim N(0,1)$. Since the coordinates of $R_2 x$ are identically distributed, the same estimate holds for every coordinate of $\sqrt{d}\,R_2\widetilde{x}$.*

Proof. Let $a = \widetilde{H}D_1\widetilde{x} = R_1\widetilde{x} \in S^{d-1}$. Therefore,

$$U(x) = \sqrt{d}\,(R_2\widetilde{x})_1 = \sqrt{d}\,(\widetilde{H}D_2R_1\widetilde{x})_1 = (HD_2R_1\widetilde{x})_1 = (HD_2a)_1 = \sum_{i=0}^{d-1} a_i\varepsilon_i \ ,$$

where $\varepsilon_i = H_{1i}(D_2)_{ii}$ are i.i.d. Rademacher signs. By Theorem 7 of [35], $d_K\big((U(x) \mid a), G\big) \le 0.5606\,\rho_3(a)$. Taking expectation and applying Lemma 1 yields

$$d_K\big(U(x), G\big) \le \mathbb{E}_a[d_K\big((U(x) \mid a), G\big)] \le 0.5606\,\mathbb{E}[\rho_3(a)] \le \frac{0.5606\,C_3}{\sqrt{d}} \ . \quad \square$$

3.2 1-Wasserstein distance with two RHTs

We utilize and restate here for clarity the following lemma.

Lemma 2.4 of [13](W_1 Berry–Esseen Method[2]). Let $X_0, \ldots, X_{d-1}$ be independent zero-mean random variables with $\sum_{i=0}^{d-1} \mathbb{E}[X_i^2] = 1$. Let $G \sim \mathcal{N}(0,1)$ be a standard Gaussian. Then:

$$W_1\left(\sum_{i=0}^{d-1} X_i,\, G \right) \le C_W \sum_{i=0}^{d-1} \mathbb{E}\big[|X_i|^3\big]$$

where $C_W \le 3$ is an absolute constant.

Now we move to bound the 1-Wasserstein distance.

[2]Also appears in [14, Section 3].

Theorem 2 (1-Wasserstein distance with two RHTs)**.** *Let $x \in \mathbb{R}^d \setminus \{0\}$. Define again $U(x) = \sqrt{d}\,(R_2\widetilde{x})_1$. Then $W_1\big(U(x), G\big) \le \frac{C_W C_3}{\sqrt{d}}$, where $G \sim N(0,1)$. Since the coordinates of $R_2 x$ are identically distributed, the same estimate holds for every coordinate of $\sqrt{d}\,R_2\widetilde{x}$.*

Proof. Again, let $a = \widetilde{H}D_1\widetilde{x} = R_1\widetilde{x} \in S^{d-1}$ and $U(x) = \sum_{i=0}^{d-1} a_i\varepsilon_i$, where $\varepsilon_i = H_{1i}(D_2)_{ii}$ are i.i.d. Rademacher signs. Notice that *conditioned on* a, the random variables $X_i = a_i\varepsilon_i$ are independent, have zero mean, and satisfy $\sum_{i=0}^{d-1}\mathbb{E}[(a_i\epsilon_i)^2] = \sum_{i=0}^{d-1} a_i^2 = 1$.

Applying Lemma 2.4 of [13] yields

$$W_1(U(x) \mid a, G) \le C_W \sum_{i=0}^{d-1} \mathbb{E}\big[|a_i\epsilon_i|^3 \mid a\big] = C_W \sum_{i=0}^{d-1} |a_i|^3 = C_W \rho_3(a)\,.$$

Taking expectation and applying Lemma 1 yields

$$\begin{aligned} W_1(U(x), G) &= \sup_{f\in \mathrm{Lip}(1)} \Big|\mathbb{E}\big[\mathbb{E}[f(U(x)) \mid a]\big] - \mathbb{E}[f(G)]\Big| \\ &\le \mathbb{E}_a\big[W_1(U(x) \mid a, G)\big] \le C_W\mathbb{E}[\rho_3(a)] \le \frac{C_W C_3}{\sqrt{d}}\ . \end{aligned}$$

□

4 Applying the 2-RHT results to DRIVE

The DRIVE algorithm [37] provides a highly efficient method for biased and unbiased 1-bit (per entry) quantization. Then, in DRIVE, the input vector $x \in \mathbb{R}^d$ is rotated to $\mathcal{R}x$ where $\mathcal{R}$ is a random rotation matrix. Then, the rotated vector is deterministically quantized to a binary vector by taking its sign. To reconstruct the vector, the estimate is $\hat{x} = S \cdot \mathcal{R}^{-1}\mathrm{sign}(\mathcal{R}x)$, for some scalar $S \in \mathbb{R}$.

In the **biased** configuration, the objective is to strictly minimize the vector's Normalized Mean Squared Error (vNMSE $= \frac{\mathbb{E}[\|x-\hat{x}\|_2^2]}{\|x\|_2^2}$). DRIVE achieves this by calculating the optimal MSE-minimizing scale factor $S = \frac{1}{d}\|\mathcal{R}x\|_1$. Because the randomized rotation forces the coordinates to behave similarly to Gaussian variables in high dimensions, this biased scaling yields a vNMSE that asymptotically converges to exactly $1 - \frac{2}{\pi} \approx 0.363$ for large d.

In the **unbiased** configuration, the scale factor is either $S = \frac{\|x\|_2^2}{\|\mathcal{R}x\|_1}$ or $S = \frac{\|x\|_2^2}{\mathbb{E}\|\mathcal{R}x\|_1}$. Since the latter is more amenable to formal analysis and, as shown in [37, Appendix A.4], yields a tighter vNMSE bound, we use it. Both these specific scalar corrections explicitly guarantee that the reconstructed vector is unbiased, meaning $\mathbb{E}[\hat{x}] = x$. Imposing this constraint slightly increases the individual vNMSE, which asymptotically converges for both to $\frac{\pi}{2} - 1 \approx 0.571$.

When quantizing low-dimensional vectors, applying URRs is computationally tractable and provides rigorously bounded theoretical guarantees. However, in practical implementations, high-dimensional data overwhelmingly favor RHT. Yet, as highlighted in the analysis of the DRIVE algorithm, this substitution is not mathematically flawless; while the RHT successfully mimics a uniform rotation for many natural inputs, it can structurally fail to uniformize certain worst-case vectors (such as sparse inputs). In particular, DRIVE [37] gives a weaker $\mathrm{vNMSE} \le 0.5$ bound for the biased case when using RHT instead of URR, and no bound for the unbiased RHT-based variant.

We now show that, after two RHTs, DRIVE recovers the URR vNMSE bounds for both the biased and unbiased variants, up to an additive term inversely polynomial in d.

4.1 The biased configuration

By [37, Lemma 1], when using $S = \frac{1}{d}\|\mathcal{R}x\|_1$, for *any* rotation matrix $\mathcal{R}$, the vNMSE is:

$$\mathrm{vNMSE}_{\mathcal{R}} = 1 - \frac{1}{d}\mathbb{E}\Big[\|\mathcal{R}\widetilde{x}\|_1^2\Big]\ . \tag{1}$$

For a URR, as mentioned, the error converges to exactly $1 - \frac{2}{\pi} \approx 0.363$ for large d. Our goal is to show that a matching asymptotic upper bound holds for $\mathcal{R} \triangleq R_2$. By Jensen's inequality:

$$\frac{1}{d}\mathbb{E}\Big[\|R_2\widetilde{x}\|_1^2\Big] \ge \frac{1}{d}\Big(\mathbb{E}\big[\|R_2\widetilde{x}\|_1\big]\Big)^2\ .$$

By linearity of expectation, $\mathbb{E}\left[\|R_2\widetilde{x}\|_1\right] = \sum_{i=0}^{d-1}\mathbb{E}[|(R_2\widetilde{x})_i|] = d\cdot\mathbb{E}[|(R_2\widetilde{x})_1|]$.
Substituting this into our squared bound yields

$$\frac{1}{d}\mathbb{E}\left[\|R_2\widetilde{x}\|_1^2\right] \geq \frac{1}{d}\Big(d\,\mathbb{E}[|(R_2\widetilde{x})_1|]\Big)^2 = d\cdot\left(\mathbb{E}[|(R_2\widetilde{x})_1|]\right)^2 .$$

Substituting this into Equation (1) yields

$$\mathrm{vNMSE}_{R_2} \leq 1 - d\cdot\left(\mathbb{E}[|(R_2\widetilde{x})_1|]\right)^2 . \tag{2}$$

We have now expressed a bound on vNMSE_{R_2} in terms of a single coordinate. We now use Lemma 2 below, which uses the fact, similar to the URR, that any single coordinate is close to a Gaussian distribution. As our lower bound can be negative for small d, we define $(z)_+ = \max(z,0)$ to obtain:

$$d\cdot\left(\mathbb{E}[|(R_2\widetilde{x})_1|]\right)^2 \geq \left(\left(\sqrt{\frac{2}{\pi}} - \frac{C_W C_3}{\sqrt{d}}\right)_+\right)^2 = \frac{2}{\pi} - \mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$$

Substituting this into Equation (2) yields the desired upper bound:

$$\mathrm{vNMSE}_{R_2} \leq 1 - d\cdot\left(\mathbb{E}[|(R_2\widetilde{x})_1|]\right)^2 \leq 1 - \left(\left(\sqrt{\frac{2}{\pi}} - \frac{C_W C_3}{\sqrt{d}}\right)_+\right)^2 = 1 - \frac{2}{\pi} + \mathcal{O}\left(\frac{1}{\sqrt{d}}\right) .$$

Lemma 2. *Let $x \in \mathbb{R}^d \setminus \{0\}$ be the input vector, $\widetilde{x} = \frac{x}{\|x\|_2}$ and define $U(x) = \sqrt{d}\cdot(R_2\widetilde{x})_1$. Then,*

$$\mathbb{E}\left[|U(x)|\right] \geq \sqrt{\frac{2}{\pi}} - \frac{C_W C_3}{\sqrt{d}} .$$

Proof. For any $u, v \in \mathbb{R}$, by the reverse triangle inequality, i.e., $\big||u| - |v|\big| \leq |u - v|$, the absolute value function $f(z) = |z|$ is exactly 1-Lipschitz. Therefore, W_1 bounds the L_1 norm difference:

$$\Big|\mathbb{E}[|X|] - \mathbb{E}[|Y|]\Big| \leq W_1(X,Y) \tag{3}$$

Let $G \sim N(0,1)$. Using Equation (3), rearranging, and applying Theorem 2 yields,

$$\mathbb{E}[|U(x)|] \geq \mathbb{E}[|G|] - W_1(U(x),G) \geq \sqrt{\frac{2}{\pi}} - \frac{C_W C_3}{\sqrt{d}} . \quad \square$$

4.2 The unbiased configuration

We now analyze the bias and variance of the unbiased 2-RHT-based DRIVE estimator for an arbitrary input vector $x \in \mathbb{R}^d \setminus \{0\}$. As defined at the beginning of Section 4, to guarantee unbiasedness under a URR, the algorithm scales the quantized vector by $S = \frac{\|x\|_2^2}{\mathbb{E}[\|\mathcal{R}_U x\|_1]}$ where $\mathcal{R}_U$ is a random rotation matrix. This simplifies to $S = \frac{\|x\|_2}{c_d\sqrt{d}}$, where $c_d \triangleq \frac{1}{\sqrt{d}}\mathbb{E}[\|\mathcal{R}_U\widetilde{x}\|_1]$ is the scaled L_1 norm expectation for a uniformly distributed unit vector $\mathcal{R}_U\widetilde{x} \in \mathcal{S}^{d-1}$. Recall the estimate:

$$\hat{x} = SR_2^{-1}\mathrm{sign}(R_2 x) = \frac{\|x\|_2}{c_d\sqrt{d}}R_2^{-1}\mathrm{sign}(R_2\widetilde{x}).$$

Let us define the expected direction vector generated by the 2-RHT as $\mu = \frac{1}{\sqrt{d}}\mathbb{E}[R_2^{-1}\mathrm{sign}(R_2\widetilde{x})]$. By factoring out the scalars, our expected estimator evaluates to $\mathbb{E}[\hat{x}] = \frac{\|x\|_2}{c_d}\mu$. The estimation bias is the vector difference between this expectation and the input vector x:

$$B(x) \triangleq \mathbb{E}[\hat{x}] - x = \frac{\|x\|_2}{c_d}(\mu - c_d\widetilde{x}). \tag{4}$$

With this definition at hand, we continue with the following theorem.

Theorem 3. *Let $x \in \mathbb{R}^d \setminus \{0\}$ be an arbitrary vector, and define the 2-RHT DRIVE estimator with the uniform-rotation unbiased scale by $\hat{x} = \frac{\|x\|_2}{c_d\sqrt{d}} R_2^{-1} sign(R_2 x)$. As $d \to \infty$, the estimator's normalized squared bias decays to 0, and its normalized variance converges to the same value as for a URR:*

$$\frac{\|B(x)\|_2^2}{\|x\|_2^2} \leq \mathcal{O}(d^{-1/2}) \qquad \text{(i)} \qquad\qquad \frac{\mathrm{Var}(\hat{x})}{\|x\|_2^2} = \frac{\pi}{2} - 1 + \mathcal{O}(d^{-1/2}) \qquad \text{(ii)}$$

Proof intuition. To prove the bias decays to zero, we show that μ converges to the target vector $c_d\widetilde{x}$. We achieve this by decomposing μ into a parallel component (its projection onto $\widetilde{x}$) and an orthogonal component ($\mu_\perp$).

Analyzing these components directly against the algorithm's constant c_d involves cumbersome algebra. Instead, our proof leverages a simpler proxy scale: the expected absolute value of a standard Gaussian, $c \triangleq \sqrt{2/\pi}$. Because 2-RHT coordinates approximate standard Gaussians (Theorem 2), the scaled expected L_1 norm of any unit vector under 2-RHT concentrates around c. We use this to bound both components:

1. **Parallel component:** The scalar projection of μ onto $\widetilde{x}$ is $\langle \mu, \widetilde{x} \rangle = d^{-1/2}\mathbb{E}\|R_2\widetilde{x}\|_1$, which is close to c.
2. **Orthogonal component:** We show that the perpendicular component $\mu_\perp$ is small, and we then bound its effect on the overall bias using the Pythagorean theorem.

Finally, because the algorithm's scale c_d converges to c with $|c - c_d| = \mathcal{O}(d^{-1})$, we substitute c for c_d at the end of the proof to establish the exact variance and bias bounds. The formal proof is deferred to Appendix A.3; the required asymptotic expansion of the scaling constant c_d is given in Appendix A.1.

Implications for Distributed Mean Estimation (DME). When N arbitrary vectors $x_{(c)}$ are averaged via their independent 2-RHT-based estimates $\hat{x}_{avg} = \frac{1}{N}\sum_{c=0}^{N-1} \hat{x}_{(c)}$ (each having a decaying bias), the Normalized Mean Squared Error (NMSE) decomposes into the aggregated variance and squared bias. Because vectors use independent 2-RHT rotations, the resulting overall variance scales down by a $1/N$ factor. Concurrently, by Jensen's inequality, the squared magnitude of the aggregated bias vector is bounded by the average of the squared individual biases. Thus, the resulting error is close to the one achieved via uniform rotations: $\mathrm{NMSE} \triangleq \frac{\mathbb{E}[\|\hat{x}_{avg} - x_{avg}\|_2^2]}{\frac{1}{N}\sum_{c=0}^{N-1}\|x_{(c)}\|_2^2} \leq \frac{\frac{\pi}{2}-1}{N} + \mathcal{O}(d^{-1/2})$.

5 Applying the 2-RHT results to QUIC-FL

State-of-the-art Distributed Mean Estimation algorithms, such as QUIC-FL [9], rely heavily on *Bounded Support Quantization (BSQ)*.[3] In BSQ, the algorithm transmits coordinates that fall outside a predefined range at baseline precision (often 32 or 16 bits) to avoid outlier errors, while stochastically quantizing the rest. Following the notation of [9], coordinates outside the range $[-t_p, t_p]$ are sent exactly (i.e., at baseline precision), where the subscript p represents that t_p is chosen so that $G \sim \mathcal{N}(0,1)$ lies outside this range with probability p. This means that after a URR, each coordinate falls outside this range with probability approximately and asymptotically p.

As discussed in [9], when QUIC-FL uses a (single) RHT instead of a URR, bounds can still be proven, but they are significantly weaker. In particular, the issue is that the expected fraction of items falling outside $[-t_p, t_p]$ can be substantially more than p. An inequality proven in [11] shows that for worst-case inputs, the tail probability of a coordinate obtained using a single RHT can be up to $3.1787p \approx 3.2p$; that is, around 3.2 times larger than that of a standard Gaussian. This multiplicative penalty in the bound on outliers cascades through the QUIC-FL analysis, leading to overestimates of both bandwidth provisioning and expected quantization error.

Our results for composing two RHTs are directly applicable and remove this penalty. Specifically, using our results for the Kolmogorov metric (Theorem 1), we show that QUIC-FL can achieve essentially the same bounds when using two composed RHTs instead of a URR, up to terms asymptotically vanishing in the dimension.

[3] See also [16, 17, 25].

5.1 Expected outlier fraction and bandwidth bounds

Recall that we let $p = \mathbb{P}(|G| > t_p)$ be the expected fraction of coordinates falling outside the range $[-t_p, t_p]$ for a standard Gaussian. With a single RHT, the expected fraction of exactly-sent coordinates (Theorem G.1 in [9]) can be bounded as follows: let $Z = \sqrt{d}(R_1\widetilde{x})_1$ be the first coordinate after one RHT; then $\mathbb{P}\big(|Z| > t_p\big) \leq 3.2 \cdot \mathbb{P}\big(|G| > t_p\big) = 3.2{\cdot}p$. The bound on the expected number of coordinates requiring full precision is thus much higher than for URRs.

When composing two RHTs, our Kolmogorov bound (Theorem 1) guarantees that the expected fraction of outliers is asymptotically the same as for the Gaussian distribution.

Theorem 4 (Expected Outlier Fraction with Two RHTs)**.** *As before, let $U(x) = \sqrt{d}(R_2\widetilde{x})_1$. Let $t_p > 0$ be a threshold such that $\mathbb{P}(|G| > t_p) = p$ for $G \sim \mathcal{N}(0,1)$. Then, the probability that a coordinate falls outside $[-t_p, t_p]$ after two RHTs is bounded by:*

$$\mathbb{P}\big(|U(x)| > t_p\big) \leq p + \frac{2 \cdot 1.28}{\sqrt{d}} = p + \frac{2.56}{\sqrt{d}} = p + \mathcal{O}\big(d^{-1/2}\big) \ .$$

Proof. Because the coordinates of $\sqrt{d}(R_2\widetilde{x})$ are identically distributed, the expected fraction of outliers equals the probability that a single coordinate exceeds the threshold t_p. We rewrite this tail probability in terms of the cumulative distribution function and apply the Kolmogorov bound from Theorem 1. Let $F_U(t) = \mathbb{P}\big(U(x) \leq t\big)$ and $F_G(t) = \mathbb{P}\big(G \leq t\big)$ denote the cumulative distribution functions of $U(x)$ and G, respectively. Then:

$$\begin{aligned}\mathbb{P}\big(|U(x)| > t_p\big) &= \mathbb{P}\big(U(x) > t_p\big) + \mathbb{P}\big(U(x) < -t_p\big) \leq 1 - F_U(t_p) + F_U(-t_p) \\ &\leq 1 - F_G(t_p) + F_G(-t_p) + 2d_K(U(x), G) = \mathbb{P}\big(|G| > t_p\big) + 2d_K(U(x), G).\end{aligned}$$

Since $d_K(U(x), G) \leq \frac{1.28}{\sqrt{d}}$ by Theorem 1, substituting this bound yields the desired result. □

As d grows, the expected fraction of outliers converges to p, saving $\approx 70\%$ of the bandwidth overhead if using the bound above for a single RHT.

5.2 Quantization error and bounded total variation

To bound the vNMSE, we define the expected squared error function conditioned on the coordinate value: $e(z) = \mathbb{E}[(\hat{z} - z)^2 \mid z]$. For $|z| \leq t_p$, $e(z)$ is defined by the stochastic quantizer. For $|z| > t_p$, the values are sent exactly, meaning $e(z) = 0$.

This creates a jump discontinuity at the thresholds $\pm t_p$. Because the error drops to zero, the function is non-Lipschitz. Consequently, the 1-Wasserstein metric used in Section 4 cannot be applied to bound this expected error. Instead, in Appendix B, we show how to use the Kolmogorov distance (d_K) to tighten the attainable bounds, getting $2.77 - 3.17\times$ vNMSE reductions for $b \in \{1, 2, 3, 4\}$.

6 Beyond scalar quantization: Three RHTs for VQ

While two-RHTs are sufficient to ensure the one-dimensional marginal distributions required for *scalar* rotation-based quantization algorithms like DRIVE and QUIC-FL, some compression schemes require multi-dimensional VQ in which coordinates are partitioned into sets of size k and treated as k-dimensional vectors. This is particularly prominent in frameworks built for inference, KV cache quantization, and vector databases, e.g., [21, 27, 29].

Vector quantizers group multi-dimensional data into localized regions based on the nearest representative point. For standard codebooks to be efficient, the coordinates within a k-sized block must have a suitable joint distribution, as Gaussian marginals alone do not suffice. For example, if coordinates are highly correlated, the data cluster in a lower-dimensional subspace, leading to significant quantization error as the distribution differs substantially from the multivariate Gaussian distribution the codebook is designed for.

In this section, we show that two RHTs may not preserve the Gaussian expected error of a fixed codebook. However, we establish that adding a third RHT yields a weak coordinate correlation sufficient for using Gaussian codebooks while maintaining their theoretical performance. Here, we consider Hadamard matrices constructed via the standard Sylvester's construction.

6.1 The correlation bottleneck of Two RHTs

To understand why two RHTs may be insufficient for VQ, consider the conditional covariance between two coordinates of the transformed vector.

Lemma 3 (2-RHT Correlation Bottleneck)**.** *There exist input vectors $\widetilde{x} \in S^{d-1}$ and distinct coordinates $i \neq j$ such that, for $U = \sqrt{d}\, R_2 \widetilde{x}$ and $a = \widetilde{H} D_1 \widetilde{x}$, for every realization of the sign matrix D_1,*

$$\mathrm{Cov}_{D_2}(U_i, U_j \mid a) = \pm 1 \, .$$

Thus, the two coordinates are perfectly conditionally correlated or anti-correlated.

Proof intuition. When evaluating the covariance between two output coordinates U_i and U_j (conditioned on the intermediate vector after the first RHT), the properties of the Hadamard matrix collapse the sum into pairwise products of the original input coordinates. If the input is highly sparse (e.g., only two non-zero coordinates), this sum reduces to a single surviving term. Because the marginal variance of each coordinate is 1, a conditional covariance of ± 1 dictates that the coordinates are perfectly conditionally correlated or anti-correlated. The derivation is deferred to Appendix C.1.

The distinction between unconditional and conditional covariance is essential. For $\widetilde{x} = (e_0 + e_1)/\sqrt{2}$, the unconditional law of (U_0, U_1) converges to (G, SG), where $G \sim \mathcal{N}(0,1)$ and S is an independent Rademacher sign. This pair has zero covariance but strongly dependent coordinates. For the (optimal) fixed codebook $\mathcal{C} = \{(\pm 2/\sqrt{\pi}, 0), (0, \pm 2/\sqrt{\pi})\}$, its limiting expected error exceeds the Gaussian error, as derived in Appendix C.2. Importantly, this example shows failure of codebook universality rather than guaranteed improvement for a given codebook.

Accordingly, we next show that a 3-RHT sequence bounds the conditional covariance itself in RMS, ruling out such bad cases for fixed coordinate pairs.

6.2 The 3-RHT conditional decorrelation guarantee

By using three RHTs, we can eliminate the conditional correlation patterns observed for two RHTs.

Theorem 5 (3-RHT Decorrelation)**.** *Let $\widetilde{x} \in S^{d-1}$, $y = R_1 \widetilde{x}$, $U = \sqrt{d} R_3 \widetilde{x}$, and for $i \neq j$ define the conditional covariance $C_{i,j}(y, D_2) = \mathrm{Cov}_{D_3}(U_i, U_j \mid y, D_2)$.*

Then, for every fixed $y = R_1 \widetilde{x}$, $\mathbb{E}_{D_2}\big[C_{i,j}(y, D_2)\big] = 0$ and $\mathbb{E}_{D_2}\big[C_{i,j}(y, D_2)^2\big] \leq 2\|y\|_\infty^2$. Moreover, the Root Mean Square (RMS) of this D_3-conditional covariance over D_1, D_2 decays:

$$\left(\mathbb{E}_{D_1, D_2}\big[C_{i,j}(R_1 \widetilde{x}, D_2)^2\big]\right)^{1/2} \leq \mathcal{O}\left(\sqrt{\frac{1}{d}}\right) \, .$$

Proof intuition. While the 2-RHT may fail, e.g., on sparse inputs, the additional RHT lets us average the conditional covariance over two independent sign matrices. For any fixed output y of the first RHT, we express the conditional covariance between two coordinates after the final RHT as a signed sum over disjoint coordinate pairs. This implies zero mean over the middle sign matrix, and also gives a deterministic second-moment bound in terms of $\|y\|_\infty^2$. To obtain the worst-case rate, we average the same pairwise expression over the first RHT. A fourth-moment calculation for two distinct coordinates gives a fixed pair contribution of order $\frac{1}{d^2}$, and there are $\frac{d}{2}$ such pairs. This yields an RMS conditional covariance bound of order $\frac{1}{\sqrt{d}}$. The complete proof is given in Appendix C.3.

6.3 Codebook universality: Matching the URR limit

In k-dimensional VQ, a codebook $\mathcal{C} = \{c_1, \ldots, c_M\} \subset \mathbb{R}^k$ maps each k-dimensional block of the vector to its nearest centroid. The expected quantization error $\mathcal{E}(\mathcal{C}, V)$ is the expected squared distance between the block V and its selected centroid from the codebook $\mathcal{C}$.

For a URR, any k-dimensional block converges to a standard multivariate Gaussian $Z \sim \mathcal{N}(0, I_k)$ as $d \to \infty$. Thus, the optimal codebook is derived via k-means clustering over this distribution. We show, with a proof provided in Appendix C.4, that a 3-RHT achieves the same quantization error up to a decaying term.

d	Improvement over one RHT	Improvement over two RHTs
128	8.59%	1.77%
512	7.68%	2.67%
2048	7.15%	2.69%
8192	6.74%	3.10%

Table 2: Relative vNMSE reductions using three RHTs with the fixed HIGGS codebook.

Theorem 6 (3-RHT Codebook Universality)**.** *Let $\widetilde{x} \in S^{d-1}$ and $U = \sqrt{d}R_3\widetilde{x}$. Fix the block size k and a finite codebook $\mathcal{C} \subset \mathbb{R}^k$ with radius $B = \max_{c\in\mathcal{C}} \|c\|_2$.*

Then, the difference in expected quantization error satisfies:

$$\left|\mathbb{E}\Big[\min_{c\in\mathcal{C}} \|U_{0:k-1} - c\|_2^2\Big] - \mathbb{E}\Big[\min_{c\in\mathcal{C}} \|Z - c\|_2^2\Big]\right| \leq \mathcal{O}_{k,B}\big(1/\sqrt{d}\big) \ .$$

7 Adaptive linear-time verification

As we explain in detail in Appendix D, scalar quantization methods such as [9, 37] can use the $\rho_3(\widetilde{x})$ check to decide whether one RHT is already sufficient. VQ methods can use the $\|\widetilde{x}\|_\infty^2$ check to decide whether two RHTs already provide the desired covariance bound with a matching vanishing-term guarantee determined by the input's ℓ_∞ norm. Both $\rho_3(\widetilde{x})$ and $\|\widetilde{x}\|_\infty^2$ are computable in $O(d)$ time, enabling an efficient and informed choice of a number of RHTs sufficient for the stated bounds.

8 Empirical validation

We test the released HIGGS [27] $p = 2$ codebook at 3 bits per value on inputs with 32 consecutive nonzeros. Three RHTs improve vNMSE over one and two RHTs (Table 2). The comparison to Uniform Random Rotation (URR) shows confidence intervals that include zero in the three dimensions with URR results, indicating that the difference is not statistically significant. See Appendix E.1.

At $d = 512$, the calibrated scalar/vector checks use a second/third RHT for 9.38%/12.1% of synthetic mixed-workload groups. Adaptive vNMSE is 0.03437/0.02949, versus 0.03434/0.02949 for always-two/three, successfully skipping unneeded RHTs. On 1968 groups from 21 Qwen2.5-0.5B tensors, including weights, activations, KV-cache blocks, and gradients, the checks skip the extra RHT for 81.7%/30.3% of groups. Appendix E.2 gives the complete details.

Appendix E.3 compares scalar Gaussian approximation after one, two, and three RHTs using Kolmogorov and 1-Wasserstein distances.

9 Approximate Nearest Neighbor Search

We evaluate RaBitQ after one or two full RHTs on six ANN datasets at 2–4 bits per transformed coordinate. Two RHTs improve 2-bit Recall@100 on MSong and Image, while gains depend on the dataset and bit budget. Appendix F gives the datasets, setup, and results.

10 Conclusion

This paper addresses a fundamental gap in practitioners' heuristic use of RHT rather than URR for quantization tasks. We prove approximation bounds for the distance between the distribution of individual coordinates and small coordinate blocks resulting from URRs and a sequence of RHTs and demonstrate how they recover optimal theoretical guarantees for recent scalar and vector quantizers.

We acknowledge parallel work: Zilca and Mendelson [41] bound the distance between coordinate marginals of two composed RHTs and those of a URR. Feng et al. [19] prove a high-rate error bound for a dithered quantization scheme using one RHT.

Declaration of LLM Usage: We have used LLMs to explore different proof routes for some of our results through iterative discussion and refinement; to provide an additional source of verification of our results; to correctly identify attribution; to simplify proofs; to help with the evaluation; and to edit. We have ensured that all content is correct, original, and adheres to ethical and academic standards.

Acknowledgments: Michael Mitzenmacher was supported in part by grants NSF CNS-2107078 and NSF DMS-2023528. Ran Ben Basat was supported in part by the Google Research Scholar Award. He was also an Affiliated Researcher with Broadcom during part of this work.

References

[1] NIST Digital Library of Mathematical Functions. `https://dlmf.nist.gov/5.6.E4`. Section 5.6, Eq. 5.6.4, Gautschi's inequality.

[2] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

[3] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, 2006.

[4] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson–Lindenstrauss transform. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 185–191, 2011.

[5] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal LSH for angular distance. *Advances in neural information processing systems*, 28, 2015. Code: `https://github.com/FALCONN-LIB/FALCONN`.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. Pmlr, 2017.

[7] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

[8] Ran Ben Basat, Michael Mitzenmacher, and Shay Vargaftik. Torque: Optimizing what (not) to quantize before and after rotation. *arXiv preprint arXiv:2609.36032*, 2026.

[9] Ran Ben-Basat, Shay Vargaftik, Amit Portnoy, Gil Einziger, Yaniv Ben-Itzhak, Michael Mitzenmacher, et al. Accelerating federated learning with quick distributed mean estimation. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Yaniv Ben-Itzhak, Helen Möllering, Benny Pinkas, Thomas Schneider, Ajith Suresh, Oleksandr Tkachenko, Shay Vargaftik, Christian Weinert, Hossein Yalame, and Avishay Yanai. Scionfl: Efficient and robust secure quantized aggregation. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 490–511. IEEE, 2024.

[11] Vidmantas Kastytis Bentkus and Dainius Dzindzalieta. A tight Gaussian bound for weighted sums of Rademacher random variables. *Bernoulli*, 21(2):1231 – 1237, 2015.

[12] Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[13] Louis HY Chen. Stein's method of normal approximation: Some recollections and reflections. *The Annals of Statistics*, 49(4):1850–1863, 2021.

[14] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

[15] Tobias Christiani. 8-bit Rotational Quantization: How to Improve the Speed-Quality Tradeoff of Vector Search. Weaviate technical blog, 2025. August 26.

[16] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

[17] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

[18] FAISS contributors. FAISS: HadamardRotation. Source code: `VectorTransform.h`, lines 129–132, 2026. Commit `bce5e24`.

[19] Ying Feng, Piotr Indyk, Michael Kapralov, Dmitry Krachun, and Boris Prokhorov. Provable Quantization with Randomized Hadamard Transform. *arXiv preprint arXiv:2605.13810*, 2026.

[20] Jianyang Gao, Yutong Gou, Yuexuan Xu, Yongyi Yang, Cheng Long, and Raymond Chi-Wing Wong. Practical and asymptotically optimal quantization of high-dimensional vectors in euclidean space for approximate nearest neighbor search. *Proceedings of the ACM on Management of Data*, 3(3):1–26, 2025.

[21] Jianyang Gao and Cheng Long. Rabitq: Quantizing high-dimensional vectors with a theoretical error bound for approximate nearest neighbor search. *Proceedings of the ACM on Management of Data*, 2(3):1–27, 2024.

[22] Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys*, 38(1):77–81, 1959.

[23] Kathy Horadam. *Hadamard matrices and their applications*. Princeton university press, 2012.

[24] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[25] Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13355–13364, 2024.

[26] Minghao Li, Ran Ben Basat, Shay Vargaftik, ChonLam Lao, Kevin Xu, Michael Mitzenmacher, and Minlan Yu. THC: Accelerating Distributed Deep Learning Using Tensor Homomorphic Compression. In *USENIX Symposium on Networked Systems Design and Implementation*, 2024.

[27] Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik, and Dan Alistarh. Higgs: Pushing the limits of large language model quantization via the linearity theorem. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10857–10886, 2025.

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

[29] Andrei Panferov, Erik Schultheis, Soroush Tabesh, and Dan Alistarh. Quartet ii: Accurate llm pre-training in nvfp4 by improved unbiased gradient estimation. *arXiv preprint arXiv:2601.22813*, 2026.

[30] Martin Raič. A multivariate central limit theorem for lipschitz and smooth test functions. *arXiv preprint arXiv:1812.08268*, 2018.

[31] Joel A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1–2):115–126, 2011.

[32] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Proceedings of machine learning research*, 235:48630–48656, 2024.

[33] Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. Qtip: Quantization with trellises and incoherence processing. *Advances in Neural Information Processing Systems*, 37:59597–59620, 2024.

[34] I. S. Tyurin. On the convergence rate in Lyapunov's theorem. *Theory of Probability and Its Applications*, 55(2):253–270, 2011.

[35] Ilya Tyurin. New estimates of the convergence rate in the lyapunov theorem. *arXiv preprint arXiv:0912.0726*, 2009.

[36] Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning*, pages 21984–22014. PMLR, 2022.

[37] Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34:362–377, 2021.

[38] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[39] Yongyi Yang, Jianyang Gao, and Wei Hu. Raana: A fast, flexible, and data-efficient post-training quantization algorithm. *arXiv preprint arXiv:2504.03717*, 2025.

[40] Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In *International Conference on Learning Representations*, 2026.

[41] Tomer Zilca and Gal Mendelson. Approximating Uniform Random Rotations by Two-Stage Structured Hadamard Rotations in High Dimensions. *arXiv preprint arXiv:2604.23418v2*, 2026.

A Deferred analysis for Section 4.2

In this appendix, we first establish the asymptotic expansion of the scaling constant c_d used by the unbiased DRIVE algorithm, and then provide the formal proof of Theorem 3.

A.1 Asymptotic expansion of the scaling constant c_d.

We now formalize the convergence gap between the scaling constant c_d and the standard Gaussian absolute expectation $c = \sqrt{\frac{2}{\pi}}$, where $g \sim \mathcal{N}(0,1)$.

For a URR matrix $\mathcal{R}_U$, define $\widetilde{x}_U \triangleq \mathcal{R}_U \widetilde{x}$. Then, $\widetilde{x}_U \in \mathcal{S}^{d-1}$ is a point on the unit sphere, distributed uniformly at random. By [37, Lemma 9], its expected L_1 norm evaluates via the Beta function to:

$$\mathbb{E}[\|\widetilde{x}_U\|_1] = \frac{2d}{(d-1)\cdot B\left(\frac{1}{2}, \frac{d-1}{2}\right)} \ .$$

Substituting this into $c_d = \frac{1}{\sqrt{d}}\mathbb{E}[\|\widetilde{x}_U\|_1]$, we rewrite the Beta function in terms of the Gamma function, $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. Evaluating $\Gamma(1/2) = \sqrt{\pi}$ yields:

$$c_d = \frac{2\sqrt{d}}{d-1}\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)} \ .$$

Applying the Gamma function recurrence relation $\Gamma(z+1) = z\Gamma(z)$ to the denominator provides $(d-1)\Gamma((d-1)/2) = 2\Gamma((d+1)/2)$. Substituting this simplifies the expression for c_d to:

$$c_d = \sqrt{\frac{d}{\pi}}\frac{\Gamma(d/2)}{\Gamma((d+1)/2)}. \tag{5}$$

To bound the resulting ratio of Gamma functions, we apply the Tricomi-Erdélyi expansion [2, Eq. 6.1.47, page 257]. For $z \to \infty$ and bounded constants a and b, the expansion is stated strictly using Big-O notation as:

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2z} + \mathcal{O}\left(\frac{1}{z^2}\right)\right).$$

Letting $z = d/2$, $a = 0$, and $b = 1/2$, we obtain:

$$\frac{\Gamma(d/2)}{\Gamma((d+1)/2)} = \left(\frac{d}{2}\right)^{-1/2}\left(1 + \frac{(-1/2)(-1/2)}{2(d/2)} + \mathcal{O}\left(\frac{1}{d^2}\right)\right)$$
$$= \sqrt{\frac{2}{d}}\left(1 + \frac{1}{4d} + \mathcal{O}\left(\frac{1}{d^2}\right)\right).$$

Applying this exact expansion back into Equation (5) gives:

$$c_d = \sqrt{\frac{d}{\pi}}\sqrt{\frac{2}{d}}\left(1 + \frac{1}{4d} + \mathcal{O}\left(\frac{1}{d^2}\right)\right) = \sqrt{\frac{2}{\pi}}\left(1 + \frac{1}{4d} + \mathcal{O}\left(\frac{1}{d^2}\right)\right).$$

Finally, using $c = \sqrt{2/\pi}$:

$$|c - c_d| = \left|\sqrt{\frac{2}{\pi}} - \sqrt{\frac{2}{\pi}}\left(1 + \frac{1}{4d} + \mathcal{O}\left(\frac{1}{d^2}\right)\right)\right| = \frac{1}{2d\sqrt{2\pi}} + \mathcal{O}\left(\frac{1}{d^2}\right).$$

Thus, the scaling discrepancy $|c - c_d|$ is bounded by $\mathcal{O}(d^{-1})$.

A.2 URR unbiased fixed-scale vNMSE for DRIVE

The following non-asymptotic calculation is used for DRIVE's unbiased URR entry of Table 1a and improves over the $\frac{\pi}{2} - 1 + \mathcal{O}(\sqrt{\frac{\log d}{d}})$ bound derived in [37]. Recall the unbiased fixed-scale DRIVE estimator, namely the variant that uses $S = \frac{\|x\|_2^2}{\mathbb{E}\|\mathcal{R}_U x\|_1} = \frac{\|x\|_2}{c_d\sqrt{d}}$, where $c_d = \frac{1}{\sqrt{d}}\mathbb{E}\|\mathcal{R}_U \widetilde{x}\|_1$.

Theorem 7 (URR DRIVE unbiased fixed-scale vNMSE)**.** *Let $x \in \mathbb{R}^d \setminus \{0\}$, let $\widetilde{x} = x/\|x\|_2$, and let $\mathcal{R}_U$ be a uniform random rotation. Consider DRIVE with the unbiased fixed scale*

$$S = \frac{\|x\|_2^2}{\mathbb{E}\|\mathcal{R}_U x\|_1} = \frac{\|x\|_2}{c_d\sqrt{d}}, \qquad \hat{x} = S\mathcal{R}_U^{-1}\operatorname{sign}(\mathcal{R}_U x).$$

Then

$$\mathrm{vNMSE} = \frac{\mathbb{E}\|\hat{x} - x\|_2^2}{\|x\|_2^2} = \frac{1}{c_d^2} - 1.$$

Moreover, for every $d \geq 1$,

$$\mathrm{vNMSE} \leq \frac{\pi}{2} - 1.$$

Proof. For clarity, we divide the proof into three steps.

Step 1: Reducing the vNMSE to c_d. By the unbiasedness result for DRIVE's fixed-scale estimator [37, Appendix A.4], we have $\mathbb{E}[\hat{x}] = x$. Therefore,

$$\mathbb{E}\|\hat{x} - x\|_2^2 = \mathbb{E}\|\hat{x}\|_2^2 - 2\langle\mathbb{E}[\hat{x}], x\rangle + \|x\|_2^2 = \mathbb{E}\|\hat{x}\|_2^2 - \|x\|_2^2.$$

It remains to compute $\|\hat{x}\|_2^2$. Since $\mathcal{R}_U$ is orthogonal,

$$\|\hat{x}\|_2^2 = \left\|\frac{\|x\|_2}{c_d\sqrt{d}}\mathcal{R}_U^{-1}\operatorname{sign}(\mathcal{R}_U x)\right\|_2^2 = \frac{\|x\|_2^2}{c_d^2 d}\left\|\operatorname{sign}(\mathcal{R}_U x)\right\|_2^2.$$

The vector $\operatorname{sign}(\mathcal{R}_U x)$ has d entries, each of magnitude 1. Hence $\|\operatorname{sign}(\mathcal{R}_U x)\|_2^2 = d$, and so $\|\hat{x}\|_2^2 = \frac{\|x\|_2^2}{c_d^2}$. Thus $\mathbb{E}\|\hat{x} - x\|_2^2 = \left(\frac{1}{c_d^2} - 1\right)\|x\|_2^2$. Dividing by $\|x\|_2^2$ gives $\mathrm{vNMSE} = \frac{1}{c_d^2} - 1$.

Step 2: Lower bounding c_d. From Appendix A.1,

$$c_d = \sqrt{\frac{d}{\pi}}\frac{\Gamma(d/2)}{\Gamma((d+1)/2)}.$$

We now show that $c_d \geq \sqrt{2/\pi}$. By Gautschi's inequality [1, Eq. 5.6.4], originally due to [22], for $x > 0$ and $s \in (0, 1)$,

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)}.$$

We apply this with $x = \frac{d}{2}$, $s = \frac{1}{2}$. Then

$$\left(\frac{d}{2}\right)^{1/2} < \frac{\Gamma(d/2+1)}{\Gamma((d+1)/2)}.$$

Using the Gamma recurrence relation $\Gamma(d/2+1) = \frac{d}{2}\,\Gamma(d/2)$, we get

$$\left(\frac{d}{2}\right)^{1/2} < \frac{d}{2}\frac{\Gamma(d/2)}{\Gamma((d+1)/2)}.$$

Dividing by $d/2$ gives

$$\frac{\Gamma(d/2)}{\Gamma((d+1)/2)} > \sqrt{\frac{2}{d}}.$$

Substituting this into the formula for c_d,

$$c_d = \sqrt{\frac{d}{\pi}}\frac{\Gamma(d/2)}{\Gamma((d+1)/2)} > \sqrt{\frac{d}{\pi}}\sqrt{\frac{2}{d}} = \sqrt{\frac{2}{\pi}}.$$

In particular, $c_d^2 \geq \frac{2}{\pi}$.

Step 3: Concluding the vNMSE upper bound. From Step 1, $\mathrm{vNMSE} = \frac{1}{c_d^2} - 1$. From Step 2, $c_d^2 \geq \frac{2}{\pi}$. Therefore, $\frac{1}{c_d^2} \leq \frac{\pi}{2}$. Substituting into the exact vNMSE identity yields

$$\mathrm{vNMSE} = \frac{1}{c_d^2} - 1 \leq \frac{\pi}{2} - 1.$$

This concludes the proof. □

A.3 Proof of Theorem 3

Proof. For clarity, we divide the proof into five steps.

Step 1: Proxy scale closeness. Let $c = \sqrt{\frac{2}{\pi}}$. By Theorem 2 and the fact that the absolute value function is 1-Lipschitz, the expected absolute value of any individual 2-RHT coordinate differs from c by at most $\mathcal{O}(d^{-1/2})$. Recall that, for any arbitrary unit vector $v \in \mathcal{S}^{d-1}$, all d coordinates of $R_2 v$, share the same marginal distribution. Thus, its scaled expected L_1 norm evaluates exactly to this single-coordinate absolute expectation. Thus, we establish a universal uniform bound δ defining the closeness of the true 2-RHT expectation to our Gaussian proxy value c:

$$\delta \triangleq \sup_{v \in \mathcal{S}^{d-1}} \left| \frac{1}{\sqrt{d}}\mathbb{E}[\|R_2 v\|_1] - c \right| = \mathcal{O}(d^{-1/2}). \tag{6}$$

This guarantees that for *any* unit direction, the expected L_1 norm after 2-RHT is in $[c - \delta, c + \delta]$.

Step 2: The parallel component. We decompose the expected direction $\mu = \frac{1}{\sqrt{d}}\mathbb{E}[R_2^{-1}\mathrm{sign}(R_2\widetilde{x})]$ into its parallel projection along $\widetilde{x}$ and an orthogonal error vector $\mu_\perp$. For the parallel component, using $R_2^{-1} = R_2^T$, the identity $\langle R_2^\top a, b\rangle = \langle a, R_2 b\rangle$ and the property $\langle \mathrm{sign}(y), y\rangle = \|y\|_1$, we have:

$$\langle \mu, \widetilde{x}\rangle = \frac{1}{\sqrt{d}}\mathbb{E}[\langle \mathrm{sign}(R_2\widetilde{x}), R_2\widetilde{x}\rangle] = \frac{1}{\sqrt{d}}\mathbb{E}[\|R_2\widetilde{x}\|_1]\ .$$

By applying Equation (6), the parallel projection is bounded, that is: $\langle \mu, \widetilde{x}\rangle \in [c - \delta, c + \delta]$.

Step 3: The orthogonal component. For the orthogonal component $\mu_\perp$, let u be the unit vector pointing in its direction, ensuring $\langle \widetilde{x}, u\rangle = 0$ and $\langle \mu, u\rangle = \|\mu_\perp\|_2$. Let $t > 0$ be a positive scalar and consider the perturbed vector $R_2(\widetilde{x} + tu)$. The absolute value inequality dictates,

$$|A + B| \geq |A| + \text{sign}(A) \cdot B \ .$$

Applying this pointwise to every coordinate of the perturbed vector and summing them yields:

$$\|R_2(\widetilde{x} + tu)\|_1 \geq \|R_2\widetilde{x}\|_1 + \langle \text{sign}(R_2\widetilde{x}), R_2 tu\rangle = \|R_2\widetilde{x}\|_1 + t\langle \text{sign}(R_2\widetilde{x}), R_2 u\rangle.$$

Taking expectations, multiplying by $\frac{1}{\sqrt{d}}$, and applying the identity,

$$\mathbb{E}[\langle \text{sign}(R_2\widetilde{x}), R_2 u\rangle] = \mathbb{E}[\langle R_2^\top \text{sign}(R_2\widetilde{x}), u\rangle] = \langle \sqrt{d}\mu, u\rangle \ ,$$

bounds the orthogonal magnitude:

$$\frac{1}{\sqrt{d}}\mathbb{E}[\|R_2(\widetilde{x} + tu)\|_1] \geq \frac{1}{\sqrt{d}}\mathbb{E}[\|R_2\widetilde{x}\|_1] + t\langle \mu, u\rangle \ . \tag{7}$$

By the Pythagorean theorem, the length of the perturbed input vector is $\|\widetilde{x} + tu\|_2 = \sqrt{1 + t^2}$. Because the L_1 norm expectation scales linearly with vector length, applying (6) upper-bounds the left side by $(c + \delta)\sqrt{1 + t^2}$. We lower-bound the right side using $c - \delta + t\|\mu_\perp\|_2$. Substituting these limits into (7) yields:

$$(c + \delta)\sqrt{1 + t^2} \geq c - \delta + t\|\mu_\perp\|_2.$$

Using the standard scalar upper bound $\sqrt{1 + t^2} \leq 1 + \frac{t^2}{2}$, expanding the left side, and subtracting $c - \delta$ results in:

$$2\delta + \frac{c + \delta}{2}t^2 \geq t\|\mu_\perp\|_2.$$

If $\|\mu_\perp\|_2 = 0$, the bound holds trivially. Otherwise, dividing by $t > 0$ and setting $t = 2\sqrt{\frac{\delta}{c+\delta}}$ yields:

$$\|\mu_\perp\|_2 \leq 2\sqrt{\delta(c + \delta)} = \mathcal{O}(\sqrt{\delta}) = \mathcal{O}(d^{-1/4}).$$

Step 4: Bridging to the algorithm's scale. By the Pythagorean theorem, the squared Euclidean distance from μ to the proxy target $c\widetilde{x}$ is bounded by:

$$\|\mu - c\widetilde{x}\|_2^2 = (\langle \mu, \widetilde{x}\rangle - c)^2 + \|\mu_\perp\|_2^2 \leq \delta^2 + \mathcal{O}(\delta) = \mathcal{O}(\delta) = \mathcal{O}(d^{-1/2}).$$

The algorithm, however, uses c_d. As derived in Appendix A.1, c_d converges to c by $|c_d - c| = \mathcal{O}(d^{-1})$. Thus, applying the triangle inequality yields:

$$\|\mu - c_d\widetilde{x}\|_2 \leq \|\mu - c\widetilde{x}\|_2 + \|(c - c_d)\widetilde{x}\|_2 \leq \mathcal{O}(d^{-1/4}) + \mathcal{O}(d^{-1}) = \mathcal{O}(d^{-1/4}).$$

Squaring this yields $\|\mu - c_d\widetilde{x}\|_2^2 \leq \mathcal{O}(d^{-1/2})$.

Recall that $\|B(x)\|_2^2 = \frac{\|x\|_2^2}{c_d^2}\|\mu - c_d\widetilde{x}\|_2^2$. Dividing by $\|x\|_2^2$ proves Equation (i):

$$\frac{\|B(x)\|_2^2}{\|x\|_2^2} = \frac{1}{c_d^2}\|\mu - c_d\widetilde{x}\|_2^2 \leq \mathcal{O}(d^{-1/2}) \ .$$

Step 5: Computing the variance. The relative variance we evaluate is $\frac{\text{Var}(\hat{x})}{\|x\|_2^2} = \frac{\mathbb{E}[\|\hat{x}\|_2^2] - \|\mathbb{E}[\hat{x}]\|_2^2}{\|x\|_2^2}$. Because the 2-RHT matrix preserves lengths due to its orthogonality, the uncentered ℓ_2 norm of any d-dimensional ± 1 sign vector is identically $\sqrt{d}$. Thus, the uncentered squared norm is deterministic:

$$\mathbb{E}[\|\hat{x}\|_2^2] = S^2 d = \left(\frac{\|x\|_2}{c_d\sqrt{d}}\right)^2 d = \frac{\|x\|_2^2}{c_d^2}.$$

Dividing by $\|x\|_2^2$, the relative uncentered squared norm is exactly $1/c_d^2$. For the squared mean, we evaluate $\frac{\|\mathbb{E}[\hat{x}]\|_2^2}{\|x\|_2^2} = \frac{1}{c_d^2}\|\mu\|_2^2$. Since $\|\mu\|_2^2 = \langle \mu, \widetilde{x}\rangle^2 + \|\mu_\perp\|_2^2 = (c \pm \delta)^2 + \mathcal{O}(d^{-1/2}) = c^2 \pm \mathcal{O}(d^{-1/2})$, we subtract this from the relative uncentered norm:

$$\frac{\text{Var}(\hat{x})}{\|x\|_2^2} = \frac{1}{c_d^2} - \frac{1}{c_d^2}(c^2 \pm \mathcal{O}(d^{-1/2})) = \frac{1 - c^2}{c_d^2} \pm \mathcal{O}(d^{-1/2}).$$

As established, $c_d = c + \mathcal{O}(d^{-1})$, which directly implies $\frac{1}{c_d^2} = \frac{1}{c^2} + \mathcal{O}(d^{-1})$. We substitute this into the leading term, absorbing the $\mathcal{O}(d^{-1})$ difference into the $\mathcal{O}(d^{-1/2})$ error. Plugging in $c = \sqrt{2/\pi}$ evaluates to $\frac{1-c^2}{c^2} = \frac{1}{c^2} - 1 = \frac{\pi}{2} - 1$, proving Equation (ii):

$$\frac{\text{Var}(\hat{x})}{\|x\|_2^2} = \frac{\pi}{2} - 1 + \mathcal{O}(d^{-1/2}) \ . \qquad \square$$

Bit Budget	**1-RHT Bound** (Thm G.3)	**2-RHT Bound**	**Improvement**
$b = 1$	$E_1 \leq 4.831$	$E_1 \approx 1.520$	$\mathbf{3.17}\times$ **tighter**
$b = 2$	$E_2 \leq 0.692$	$E_2 \approx 0.223$	$\mathbf{3.10}\times$ **tighter**
$b = 3$	$E_3 \leq 0.131$	$E_3 \approx 0.044$	$\mathbf{2.97}\times$ **tighter**
$b = 4$	$E_4 \leq 0.0272$	$E_4 \approx 0.0098$	$\mathbf{2.77}\times$ **tighter**

Table 3: Comparing the $b - bit$ vNMSE, i.e., E_b, as a function of the bit-budget b for 1-RHT and 2-RHT, adapting results from [9].

B Deferred analysis for Section 5.2

B.1 Improved vNMSE bounds for QUIC-FL

To bound the actual quantization error (vNMSE) of BSQ, we define the expected squared error function conditioned on the coordinate value: $e(z) = \mathbb{E}[(\hat{z} - z)^2 \mid z]$. For $|z| \leq t_p$, $e(z)$ is defined by the stochastic quantizer. For $|z| > t_p$, the values are sent exactly, meaning $e(z) = 0$.

This creates a jump discontinuity at the thresholds $\pm t_p$. Because the error drops to zero, the function is non-Lipschitz. Consequently, the 1-Wasserstein metric used in Section 4 cannot be applied to bound this expected error. Instead, we use the Kolmogorov distance (d_K), which can bound the expectation of any function with bounded total variation (TV). For a function with jump discontinuities, the total variation is appropriately defined by $TV(f) = \sup_{n \geq 1} \sup_{z_0 < \cdots < z_n} \sum_{j=1}^{n} |f(z_j) - f(z_{j-1})| < \infty$.

Theorem 8 (Quantization Error with Two RHTs)**.** *Let $e : \mathbb{R} \to \mathbb{R}$ be right-continuous, of bounded variation, with finite limits at $\pm\infty$ (representing the expected squared error function for BSQ). Let $TV(e)$ denote its total variation. For any input $x \in \mathbb{R}^d \setminus \{0\}$, the expected quantization error of a scaled transformed coordinate $U(x) = \sqrt{d}(R_2\widetilde{x})_1$ after two RHTs satisfies:*

$$\Big|\mathbb{E}[e(U(x))] - \mathbb{E}[e(G)]\Big| \leq TV(e) d_K(U(x), G) \leq \frac{1.28 \cdot TV(e)}{\sqrt{d}} \ ,$$

where $G \sim \mathcal{N}(0, 1)$.

Proof intuition. According to Theorem 1, a scaled transformed coordinate $U(x) = \sqrt{d}(R_2\widetilde{x})_1$ is close to Gaussian in Kolmogorov distance. Since Kolmogorov distance controls how much probability the two distributions assign below any threshold, and since $TV(e)$ measures the total amount by which the quantization-error function can change, the expected value of e can change by at most their product. The complete proof is deferred to Appendix B.2.

Refined worst-case error bounds. In [9, Theorem G.3], computing worst-case bounds for the quantization error E_b required partitioning the support into coarse intervals (e.g., $[0, 1.5]$) and penalizing the entire probability mass of that interval with the maximal error occurring at its boundary. This Riemann sum overestimation was necessary due to the 1-RHT multiplicative tail bound, leading to inflated theoretical bounds (e.g., producing 1-bit quantization error $E_1 \leq 4.831$ as $d \to \infty$).

By substituting 1-RHT with 2-RHT, the theoretical vNMSE bounds converge to the continuous values (identified numerically in [9, Section 3.5]). Transitioning from 1-RHT to 2-RHT thus recovers the desired theoretical error limits up to a decaying $\mathcal{O}(d^{-1/2})$ term.

B.2 Proof of Theorem 8

Proof. Let F_U be the CDF of $U(x) = \sqrt{d}(R_2\widetilde{x})_1$, and write $F_U(z^-) = \mathbb{P}(U(x) < z)$. Let F_G be the CDF of the standard Gaussian $G \sim \mathcal{N}(0, 1)$, so $F_G(z) = \mathbb{P}(G \leq z) = \mathbb{P}(G < z)$. Let de denote the signed Stieltjes measure induced by the right-continuous bounded-variation function e.

We formulate the difference in the expected quantization error given by $e(z) = \mathbb{E}[(\hat{z} - z)^2 \mid z]$ between $U(x)$ and G using the Lebesgue-Stieltjes integral over the real line:

$$\begin{aligned}\Big|\mathbb{E}[e(U(x))] - \mathbb{E}[e(G)]\Big| &= \left|\int_{-\infty}^{\infty} e(z)\, dF_U(z) - \int_{-\infty}^{\infty} e(z)\, dF_G(z)\right| \\ &= \left|\int_{-\infty}^{\infty} e(z)\, d\big(F_U(z) - F_G(z)\big)\right| \ .\end{aligned}$$

To isolate the CDF difference, we apply integration by parts for Lebesgue-Stieltjes integrals:

$$\int_{-\infty}^{\infty} e(z)\, d\big(F_U(z) - F_G(z)\big) = \Big[e(z)\big(F_U(z) - F_G(z)\big)\Big]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \big(F_U(z^-) - F_G(z^-)\big)\, de(z) \ .$$

We first evaluate the boundary terms. By the definition of Bounded Support Quantization, coordinates outside $[-t_p, t_p]$ are sent without quantization error, meaning $e(z) = 0$ for all $|z| > t_p$. Consequently, the boundary evaluations vanish:

$$\lim_{z\to\infty} e(z)\big(F_U(z) - F_G(z)\big) = 0 \quad \text{and} \quad \lim_{z\to-\infty} e(z)\big(F_U(z) - F_G(z)\big) = 0 \ .$$

This leaves only the integral term. Taking the absolute value yields:

$$\begin{aligned}\Big|\mathbb{E}[e(U(x))] - \mathbb{E}[e(G)]\Big| &= \left|-\int_{-\infty}^{\infty} \big(F_U(z^-) - F_G(z^-)\big)\, de(z)\right| \\ &\le \int_{-\infty}^{\infty} \big|F_U(z^-) - F_G(z^-)\big|\, |de|(z) \ .\end{aligned}$$

By definition, the Kolmogorov distance uniformly bounds the absolute difference between the CDFs at any real point: $\sup_z |F_U(z) - F_G(z)| = d_K(U(x), G)$. The same bound holds for $F_U(z^-) - F_G(z^-)$. Therefore, we pull this supremum outside the integral:

$$\int_{-\infty}^{\infty} \big|F_U(z^-) - F_G(z^-)\big|\, |de|(z) \le d_K(U(x), G) \int_{-\infty}^{\infty} |de|(z) \ .$$

The remaining integral term, $\int_{-\infty}^{\infty} |de|(z)$, corresponds to the Total Variation of the error function $e(z)$, denoted $TV(e)$. To properly account for the jump discontinuities at $\pm t_p$, the Total Variation is defined by the supremum over partitions:

$$TV(e) = \sup_{n\ge1} \sup_{z_0<\cdots<z_n} \sum_{j=1}^{n} \big|e(z_j) - e(z_{j-1})\big| \ .$$

If $e(z)$ is piecewise absolutely continuous, this equals the sum of the integrals of $|e'(z)|$ on the smooth pieces plus the magnitudes of all jump discontinuities.

Substituting $TV(e)$ and applying the 2-RHT Kolmogorov bound derived in Theorem 1 gives the desired bound:

$$\Big|\mathbb{E}[e(U(x))] - \mathbb{E}[e(G)]\Big| \le TV(e)\, d_K(U(x), G) \le \frac{1.28 \cdot TV(e)}{\sqrt{d}} \ . \qquad \square$$

C Deferred analysis for Section 6

C.1 Proof of Lemma 3

Proof. For clarity, we divide the proof into six steps.

Step 1: Definitions. Let $\widetilde{x} \in S^{d-1}$ and $U = \sqrt{d} R_2 \widetilde{x} = \sqrt{d}\widetilde{H} D_2 \widetilde{H} D_1 \widetilde{x}$. Define the intermediate vector $a = \widetilde{H} D_1 \widetilde{x}$. Recalling $\sqrt{d}\widetilde{H} = H$, we can write $U = H D_2 a$. Conditioned on a, the randomness of D_1 is fixed, and the i-th coordinate of U is given by $U_i = \sum_{r=0}^{d-1} H_{ir} \varepsilon_r a_r$.

Step 2: Formulating the covariance. We seek the conditional covariance between two distinct coordinates U_i and U_j ($i \neq j$) over the randomness of D_2. Since $\mathbb{E}[\varepsilon_r] = 0$, the expected value $\mathbb{E}[U_i \mid a] = 0$, making the covariance equal to $\mathbb{E}[U_i U_j \mid a]$:

$$\mathrm{Cov}(U_i, U_j \mid a) = \mathbb{E}[U_i U_j \mid a] = \mathbb{E}\left[\left(\sum_r H_{ir}\varepsilon_r a_r\right)\left(\sum_s H_{js}\varepsilon_s a_s\right)\right] \ .$$

Because the signs ε are independent Rademacher variables, $\mathbb{E}[\varepsilon_r \varepsilon_s] = 0$ for $r \neq s$, and $\mathbb{E}[\varepsilon_r^2] = 1$. The double sum thereby collapses to:

$$\mathrm{Cov}(U_i, U_j \mid a) = \sum_{r=0}^{d-1} H_{ir} H_{jr} a_r^2 \ .$$

Step 3: The XOR property. A fundamental property of the Sylvester-Hadamard matrix (using 0-based indexing) is that the product of two elements in the same column is given by the bitwise XOR of their row indices: $H_{ir}H_{jr} = H_{i\oplus j,r}$. Substituting this into our covariance yields:

$$\mathrm{Cov}(U_i, U_j \mid a) = \sum_{r=0}^{d-1} H_{i\oplus j,r} a_r^2 \ .$$

Step 4: Expanding the intermediate vector a. Recall $a_r = \frac{1}{\sqrt{d}} \sum_{l=0}^{d-1} H_{rl} s_l \widetilde{x}_l$. Squaring yields:

$$a_r^2 = \frac{1}{d} \sum_{l=0}^{d-1} \sum_{m=0}^{d-1} H_{rl} H_{rm} s_l s_m \widetilde{x}_l \widetilde{x}_m \ .$$

Separating the diagonal ($l = m$) from the cross-terms ($l \neq m$) and applying the XOR property $H_{rl}H_{rm} = H_{r,l\oplus m}$, we get:

$$a_r^2 = \frac{1}{d} \sum_{l=0}^{d-1} \widetilde{x}_l^2 + \frac{1}{d} \sum_{l\neq m} H_{r,l\oplus m} s_l s_m \widetilde{x}_l \widetilde{x}_m \ .$$

Step 5: Canceling the independent sum. We substitute a_r^2 back into the covariance equation from Step 3. For the first term, because $i \neq j$, we know $i \oplus j \neq 0$. The sum of any non-zero row of a Hadamard matrix is 0 ($\sum_r H_{i\oplus j,r} = 0$), perfectly eliminating the independent $\frac{1}{d}\sum \widetilde{x}_l^2$ term. For the cross-terms, swapping the order of summation and using that H is symmetric yields:

$$\mathrm{Cov}(U_i, U_j \mid a) = \frac{1}{d} \sum_{l\neq m} s_l s_m \widetilde{x}_l \widetilde{x}_m \left(\sum_{r=0}^{d-1} H_{i\oplus j,r} H_{l\oplus m,r} \right) \ .$$

The inner sum is the dot product of two Hadamard rows, which equals d if $i \oplus j = l \oplus m$, and 0 otherwise. This cancels the $\frac{1}{d}$ and acts as an indicator function $\mathbb{I}(i \oplus j = l \oplus m)$, simplifying the covariance to:

$$\mathrm{Cov}(U_i, U_j \mid a) = 2 \sum_{l<m \,:\, l\oplus m = i\oplus j} s_l s_m \widetilde{x}_l \widetilde{x}_m \ .$$

Step 6: The sparse input counterexample. Consider $i = 0$, $j = 1$, and

$$\widetilde{x} = \frac{e_0 + e_1}{\sqrt{2}} \in S^{d-1}.$$

Then $i \oplus j = 1$. In the covariance formula from Step 5, the only nonzero pair satisfying $l < m$ and $l \oplus m = i \oplus j = 1$ is $(l, m) = (0, 1)$. Therefore,

$$\mathrm{Cov}_{D_2}(U_0, U_1 \mid a) = 2 s_0 s_1 \widetilde{x}_0 \widetilde{x}_1 = s_0 s_1 \in \{\pm 1\}.$$

Moreover,

$$\mathrm{Var}_{D_2}(U_0 \mid a) = \sum_{r=0}^{d-1} H_{0r}^2 a_r^2 = \|a\|_2^2 = 1$$

and similarly $\mathrm{Var}_{D_2}(U_1 \mid a) = 1$. Hence, the conditional Pearson correlation is exactly ± 1. □

C.2 Unconditional error gap for two RHTs

In Lemma 3, we computed the covariance conditioned on the first RHT. Here, we take the expectation over both D_1 and D_2 and compare the expected quantization error with that of two independent Gaussian coordinates, using the same codebook. We show that the covariance is zero, but the coordinates remain dependent, which changes the expected quantization error.

Step 1: The first two coordinates. Consider the same normalized input $\widetilde{x} = (e_0 + e_1)/\sqrt{2}$ as in Appendix C.1, and let $U = \sqrt{d}\, R_2 \widetilde{x}$. Write $D_1 = \mathrm{diag}(s_0, \ldots, s_{d-1})$ and let $S = s_0 s_1$. Since s_0, s_1 are independent Rademacher signs, S is also a Rademacher sign. The Sylvester-Hadamard matrix satisfies $H_{r0} = 1$ and $H_{r1} = H_{1r}$. Thus, the intermediate vector $a = \widetilde{H} D_1 \widetilde{x}$ satisfies

$$a_r = (\widetilde{H} D_1 \widetilde{x})_r = \frac{s_0 + H_{r1} s_1}{\sqrt{2d}} = \begin{cases} s_0\sqrt{2/d}, & H_{1r} = S, \\ 0, & H_{1r} = -S. \end{cases}$$

Row 1 contains $d/2$ entries of each sign. Therefore, a has exactly $d/2$ nonzero coordinates, at the indices r for which $H_{1r} = S$.

Let ε_r be the independent signs in D_2. Since $U = H D_2 a$, $H_{0r} = 1$, and only the coordinates with $H_{1r} = S$ contribute to the sums, we get

$$U_0 = s_0\sqrt{\frac{2}{d}} \sum_{r : H_{1r} = S} \varepsilon_r, \qquad U_1 = s_0\sqrt{\frac{2}{d}} \sum_{r : H_{1r} = S} H_{1r}\varepsilon_r = S U_0.$$

Conditioned on D_1, the $d/2$ signs $s_0\varepsilon_r$ in the sum for U_0 are independent Rademacher variables. Their distribution is the same for every choice of D_1. Thus, U_0 is independent of S, and (U_0, U_1) has the same distribution as (T_d, ST_d), where

$$T_d = \sqrt{\frac{2}{d}} \sum_{r=1}^{d/2} \xi_r.$$

Here, $\xi_1, \ldots, \xi_{d/2}$ are independent Rademacher signs, chosen independently of S.

Step 2: The limiting distribution. By the central limit theorem, $T_d \Rightarrow G \sim \mathcal{N}(0, 1)$. Since $\mathbb{E}T_d^2 = 1$ for every d, the expected absolute values converge as well:

$$\mathbb{E}|T_d| \longrightarrow \mathbb{E}|G| = \sqrt{\frac{2}{\pi}}.$$

Since T_d and S are independent, $(U_0, U_1) \Rightarrow (G, SG)$, with G and S independent. The limiting coordinates have zero covariance, since $\mathbb{E}[SG^2] = 0$, but are dependent because $|G| = |SG|$.

Step 3: Quantization error with two RHTs. Set $\alpha = 2/\sqrt{\pi}$ and consider the same four-point codebook as in Section 6,

$$\mathcal{C} = \{(\alpha, 0), (-\alpha, 0), (0, \alpha), (0, -\alpha)\}.$$

For $v = (v_1, v_2)$, all four centroids have squared norm α^2. Expanding the squared distance gives $\|v - c\|_2^2 = \|v\|_2^2 + \alpha^2 - 2\langle v, c\rangle$. Only the inner product depends on the choice of c. It is largest when we choose the coordinate of v with the larger absolute value and use the same sign. Therefore,

$$\begin{aligned} e_{\mathcal{C}}(v) &:= \min_{c \in \mathcal{C}} \|v - c\|_2^2 \\ &= \|v\|_2^2 + \alpha^2 - 2\alpha \max(|v_1|, |v_2|). \end{aligned}$$

For (T_d, ST_d), the maximum is just $|T_d|$ and the squared norm is $2T_d^2$. Using $\mathbb{E}T_d^2 = 1$ and Step 2:

$$\mathbb{E}e_{\mathcal{C}}(U_0, U_1) = 2 + \frac{4}{\pi} - \frac{4}{\sqrt{\pi}}\mathbb{E}|T_d|,$$

$$\lim_{d \to \infty} \mathbb{E}e_{\mathcal{C}}(U_0, U_1) = 2 + \frac{4}{\pi} - \frac{4\sqrt{2}}{\pi}.$$

The squared norm has expectation 2 for every d, so we only needed to compute the limit of $\mathbb{E}|T_d|$.

Step 4: Comparison with Gaussian coordinates. Let $Z = (Z_1, Z_2) \sim \mathcal{N}(0, I_2)$. Its expected squared norm is also 2, but the expected maximum of its absolute coordinates is larger. To compute that maximum, use the identity

$$\max(|z_1|, |z_2|) = \frac{|z_1 + z_2| + |z_1 - z_2|}{2}.$$

Both $Z_1 + Z_2$ and $Z_1 - Z_2$ have distribution $\mathcal{N}(0, 2)$, so each has expected absolute value $2/\sqrt{\pi}$. Taking expectations therefore gives

$$\mathbb{E}\max(|Z_1|, |Z_2|) = \frac{2}{\sqrt{\pi}}, \qquad \mathbb{E}e_{\mathcal{C}}(Z) = 2 + \frac{4}{\pi} - \frac{8}{\pi} = 2 - \frac{4}{\pi}.$$

Subtracting the Gaussian error from the limiting two-RHT error gives a positive gap that does not decrease with d:

$$\lim_{d\to\infty} \mathbb{E}e_{\mathcal{C}}(U_0, U_1) - \mathbb{E}e_{\mathcal{C}}(Z) = \frac{4(2-\sqrt{2})}{\pi} > 0.$$

Thus, even though each coordinate converges to a standard Gaussian and the covariance is zero, the expected quantization error after two RHTs does not converge to the Gaussian error for this codebook. This example shows that two RHTs cannot guarantee the same expected error for every fixed codebook.

C.3 Proof of Theorem 5

Proof. For clarity, we divide the proof into five steps.

Step 1: Reduction to the 2-RHT case. Consider $U = \sqrt{d}R_3\widetilde{x} = \sqrt{d}\widetilde{H}D_3\widetilde{H}D_2(\widetilde{H}D_1\widetilde{x})$. Let $y = \widetilde{H}D_1\widetilde{x} = R_1\widetilde{x}$ be the output of the first RHT. If we condition on y, the final two RHTs $\sqrt{d}\widetilde{H}D_3\widetilde{H}D_2y$ match the structure of the 2-RHT transformation analyzed in Lemma 3. Therefore, we can apply the previous conditional covariance formula, replacing the original input $\widetilde{x}$ with y, and using s to denote the random signs of the intermediate matrix D_2:

$$C_{i,j}(y, D_2) = 2 \sum_{l<m\,:\,l\oplus m=i\oplus j} s_l s_m y_l y_m \ .$$

Step 2: The expected covariance is zero. We analyze the expected value of this conditional covariance over the randomness of D_2. The sum iterates over indices where $l < m$, where $l \neq m$. Because the Rademacher signs of D_2 are independent and zero-mean, $\mathbb{E}_{D_2}[s_l s_m] = \mathbb{E}[s_l]\mathbb{E}[s_m] = 0$. Due to the linearity of expectation, the expression evaluates to zero for any given y: $\mathbb{E}_{D_2}[C_{i,j}(y, D_2)] = 0$.

Step 3: Variance over disjoint pairs. To bound the magnitude of the conditional correlation, we compute its variance over D_2. Since the mean is zero, for any given y, the variance is

$$\mathrm{Var}_{D_2}(C_{i,j}(y, D_2)) = \mathbb{E}_{D_2}[C_{i,j}(y, D_2)^2] \ .$$

When expanding the squared sum, we generate squared terms and cross-terms. The condition $l \oplus m = i \oplus j$ partitions the d indices into exactly $d/2$ uniquely matched disjoint pairs. Because the pairs are entirely disjoint, the cross-terms consist of independent Rademacher products which vanish under expectation. We are left with the expected values of the squared terms. Since $(s_l s_m)^2 = 1$:

$$\mathbb{E}_{D_2}[C_{i,j}(y, D_2)^2] = 4 \sum_{l<m\,:\,l\oplus m=i\oplus j} y_l^2 y_m^2 \ .$$

Step 4: Bounding with the maximum norm. Let $\alpha = i \oplus j$. Since $i \neq j$, we have $\alpha \neq 0$. We can replace the strict $l < m$ summation by a sum over all $l \in \{0, \ldots, d-1\}$, with $m = l \oplus \alpha$ (which is equivalent to $l \oplus m = \alpha$). Since $\alpha \neq 0$, the map $l \mapsto l \oplus \alpha$ pairs each index with a distinct index. Therefore, the sum over all l counts each unordered pair twice:

$$\sum_{l=0}^{d-1} y_l^2 y_{l\oplus\alpha}^2 = 2 \sum_{l<m\,:\,l\oplus m=\alpha} y_l^2 y_m^2.$$

Using the expression from Step 3, $\alpha = i \oplus j$ and the above yields,

$$\mathbb{E}_{D_2}[C_{i,j}(y, D_2)^2] = 4 \sum_{l<m\,:\,l\oplus m=\alpha} y_l^2 y_m^2 = 2 \sum_{l=0}^{d-1} y_l^2 y_{l\oplus\alpha}^2 \ .$$

We bound this sum using the maximum squared element of y, denoted by $\|y\|_\infty^2$:

$$\mathbb{E}_{D_2}[C_{i,j}(y, D_2)^2] = 2\sum_{l=0}^{d-1} y_l^2 y_{l\oplus\alpha}^2 \leq 2\max_k(y_k^2)\sum_{l=0}^{d-1} y_l^2 = 2\|y\|_\infty^2 \sum_{l=0}^{d-1} y_l^2 \ .$$

Because the first RHT preserves the ℓ_2 norm of the input vector $\widetilde{x} \in S^{d-1}$, $\|y\|_2^2 = 1$. This yields

$$\mathbb{E}_{D_2}[C_{i,j}(y, D_2)^2] \leq 2\|y\|_\infty^2 \ .$$

Step 5: Averaging directly over the first RHT. We now use the exact expression from Step 3 and average it directly over the signs of D_1. For fixed distinct p, q, write

$$y_p = \frac{1}{\sqrt{d}}\sum_{r=0}^{d-1} H_{pr}\sigma_r\widetilde{x}_r, \qquad y_q = \frac{1}{\sqrt{d}}\sum_{r=0}^{d-1} H_{qr}\sigma_r\widetilde{x}_r,$$

where σ_r are the Rademacher signs of D_1. A direct fourth-moment expansion gives

$$\mathbb{E}_{D_1}\left[y_p^2 y_q^2\right] = \frac{1}{d^2}\left(1 + 2\theta_{pq}^2 - 2\sum_{r=0}^{d-1}\widetilde{x}_r^4\right),$$

where $\theta_{pq} = \sum_{r=0}^{d-1} H_{pr}H_{qr}\widetilde{x}_r^2$. Since $|\theta_{pq}| \leq 1$, this implies $\mathbb{E}_{D_1}\left[y_p^2 y_q^2\right] \leq \frac{3}{d^2}$. Using the $d/2$ unordered pairs satisfying $p \oplus q = i \oplus j$, Step 3 yields, $\mathbb{E}_{D_1,D_2}\left[C_{i,j}(R_1\widetilde{x}, D_2)^2\right] \leq 4 \cdot \frac{d}{2} \cdot \frac{3}{d^2} = \frac{6}{d}$. Taking square roots gives $\left(\mathbb{E}_{D_1,D_2}\left[C_{i,j}(R_1\widetilde{x}, D_2)^2\right]\right)^{1/2} \leq \sqrt{\frac{6}{d}}$, and concludes the proof. □

C.4 Proof of Theorem 6

C.4.1 Main proof.

Proof. For clarity, we divide the proof into five steps.

Step 1: Deconstructing the matrix multiplication. Conditioned on the intermediate vector $y = R_2\widetilde{x}$, the last RHT applies the transformation $U = \sqrt{d}\widetilde{H}D_3 y = HD_3 y$. It is sufficient to analyze only the first k coordinates of U, which we denote as the block $U_{0:k-1}$.

Let $H_{0:k-1}$ be the sub-matrix consisting of only the first k rows of the Hadamard matrix. We can write our output block as:

$$U_{0:k-1} = H_{0:k-1}(D_3 y) \ .$$

The vector $D_3 y$ is simply the column vector y with its coordinates multiplied by the random Rademacher signs ε_j from D_3. Therefore, the j-th element of $D_3 y$ is exactly $\varepsilon_j y_j$.

By definition, multiplying $H_{0:k-1}$ by a column vector produces a linear combination of the *columns* of $H_{0:k-1}$. Let $H^{(j)}$ denote the j-th column of $H_{0:k-1}$ (which is a deterministic vector of length k). The multiplication expands to:

$$U_{0:k-1} = \sum_{j=0}^{d-1}(\varepsilon_j y_j)H^{(j)} \ .$$

To separate the random signs ε from the deterministic values (i.e., conditioned on y), we group the deterministic parts together to define a set of d fixed "step vectors", denoted as V_j:

$$V_j = y_j H^{(j)} \ .$$

Notice that the i-th coordinate of this step vector is exactly $(V_j)_i = H_{i,j}y_j$. Substituting this definition back into our sum, we have:

$$U_{0:k-1} = \sum_{j=0}^{d-1}\varepsilon_j V_j \ .$$

Now, calculating the conditional covariance matrix Σ_y of this block becomes straightforward. By definition, $\Sigma_y = \mathbb{E}[U_{0:k-1}U_{0:k-1}^\top \mid y]$. Because the signs ε are independent and zero-mean, all

cross-terms $\mathbb{E}[\varepsilon_j \varepsilon_m]$ vanish when $j \neq m$. We are left only with the sum of the outer products of the individual step vectors:

$$\Sigma_y = \sum_{j=0}^{d-1} \mathbb{E}[\varepsilon_j^2] V_j V_j^\top = \sum_{j=0}^{d-1} V_j V_j^\top \ .$$

Looking at the specific entries of this $k \times k$ covariance matrix:

- **Diagonal entries (Variance):** For coordinate i, this is $\sum_j (V_j)_i^2 = \sum_j (H_{i,j} y_j)^2$. Since the Hadamard matrix consists of ± 1, $H_{i,j}^2 = 1$, simplifying the sum to $\sum_j y_j^2 = \|y\|_2^2 = 1$.
- **Off-diagonal entries (Covariance):** For distinct coordinates i and l, this is $\sum_j (V_j)_i (V_j)_l = \sum_j H_{i,j} H_{l,j} y_j^2$. Using the Hadamard XOR property, this becomes exactly $\sum_j H_{i \oplus l, j} y_j^2$, which matches exactly the cross-correlation equation $C_{i,l}(R_1 \widetilde{x}, D_2)$ established in Theorem 5.

Step 2: Bounding the covariance difference. Write the conditioned vector as

$$y = R_2 \widetilde{x} = \widetilde{H} D_2 y^{(1)}, \qquad y^{(1)} = R_1 \widetilde{x}.$$

The diagonal entries of Σ_y are equal to 1. For $i \neq \ell$, the off-diagonal entry satisfies

$$(\Sigma_y)_{i\ell} = \sum_{j=0}^{d-1} H_{ij} H_{\ell j} y_j^2 = C_{i,\ell}(y^{(1)}, D_2),$$

where $C_{i,\ell}$ is the conditional covariance from Theorem 5. Therefore, by Jensen's inequality and Theorem 5,

$$\begin{aligned} \mathbb{E}\big[\|\Sigma_y - I_k\|_F\big] &\leq \big(\mathbb{E}\big[\|\Sigma_y - I_k\|_F^2\big]\big)^{1/2} \\ &= \left(\sum_{i \neq \ell} \mathbb{E}\big[C_{i,\ell}(R_1 \widetilde{x}, D_2)^2\big]\right)^{1/2} \leq \sqrt{\frac{6k(k-1)}{d}}. \end{aligned}$$

Step 3: Formalizing the Wasserstein bound via Stein's method. To bound the distance between our block $U_{0:k-1}$ and a standard Gaussian $Z \sim \mathcal{N}(0, I_k)$, we use a multivariate normal approximation lemma proved in Appendix C.4.2 that is based on [30, Theorem 2.15 and Eq. (3.5)]. For a sum of independent, zero-mean random vectors $W = \sum_{j=0}^{d-1} X_j$ with covariance $\Sigma = \sum_{j=0}^{d-1} \mathbb{E}[X_j X_j^\top]$, its 1-Wasserstein distance to a standard Gaussian is bounded by:

$$W_1\Big(W, \mathcal{N}(0, I_k)\Big) \leq A_k \left(\|\Sigma - I_k\|_F + \sum_{j=0}^{d-1} \mathbb{E}\big[\|X_j\|_2^3\big]\right)$$

where A_k is a constant that depends on k.

In our setting, conditioned on the intermediate vector y, our block is the sum $U_{0:k-1} = \sum_{j=0}^{d-1} \varepsilon_j V_j$. The summands $X_j = \varepsilon_j V_j$ are independent because the Rademacher signs ε_j are independent. They are zero-mean ($\mathbb{E}[\varepsilon_j V_j \mid y] = 0$), and their conditional covariance is Σ_y as derived in Step 1.

Because $\varepsilon_j \in \{-1, 1\}$, the third absolute moment of each summand simplifies to:

$$\mathbb{E}\big[\|\varepsilon_j V_j\|_2^3 \mid y\big] = \|V_j\|_2^3 \mathbb{E}\big[|\varepsilon_j|^3\big] = \|V_j\|_2^3 \ .$$

We calculate the length of each step vector as $\|V_j\|_2 = \sqrt{\sum_{i=0}^{k-1} (H_{i,j} y_j)^2} = \sqrt{k}|y_j|$. Summing these cubes over all d steps yields:

$$\sum_{j=0}^{d-1} \|V_j\|_2^3 = \sum_{j=0}^{d-1} k^{3/2} |y_j|^3 = k^{3/2} \rho_3(y) \ ,$$

where $\rho_3(y) = \sum_{j=0}^{d-1} |y_j|^3$ is the third absolute moment of the intermediate vector y.

Substituting the conditional covariance error and the third moments into the above bound gives the conditional Wasserstein distance:

$$W_1\Big(U_{0:k-1} \mid y,\, \mathcal{N}(0, I_k)\Big) \le A_k \left(\|\Sigma_y - I_k\|_F + k^{3/2}\rho_3(y)\right) .$$

To obtain the unconditional bound, we take the expectation over the random vector y. By the joint convexity of the Wasserstein metric, the expected distance is bounded by the expectation of the conditional Wasserstein distances, $W_1(U_{0:k-1}, \mathcal{N}(0, I_k)) \le \mathbb{E}_y\big[W_1(U_{0:k-1} \mid y, \mathcal{N}(0, I_k))\big]$. Applying this yields:

$$W_1\Big(U_{0:k-1},\, \mathcal{N}(0, I_k)\Big) \le A_k \left(\mathbb{E}\big[\|\Sigma_y - I_k\|_F\big] + k^{3/2}\mathbb{E}\big[\rho_3(y)\big]\right) .$$

From Step 2, we have bounded the expected covariance difference as

$$\mathbb{E}[\|\Sigma_y - I_k\|_F] \le \sqrt{\frac{6k(k-1)}{d}}.$$

Furthermore, by the 1-RHT smoothing property from Lemma 1, the expected third moment of y is bounded by

$$\mathbb{E}[\rho_3(y)] \le \frac{C_3}{\sqrt{d}}.$$

Combining these terms gives our final multi-dimensional bound:

$$W_1\Big(U_{0:k-1},\, \mathcal{N}(0, I_k)\Big) \le A_k \left(\sqrt{\frac{6k(k-1)}{d}} + \frac{k^{3/2}C_3}{\sqrt{d}}\right) = \mathcal{O}_k\left(\frac{1}{\sqrt{d}}\right).$$

Step 4: Separating the quadratic part of the VQ error. Define the VQ error function $L_{\mathcal{C}}(v) = \min_{c\in\mathcal{C}} \|v - c\|_2^2$. We decompose this error into a quadratic term and a codebook-dependent term: $L_{\mathcal{C}}(v) = \|v\|_2^2 + g_{\mathcal{C}}(v)$, where $g_{\mathcal{C}}(v) = \min_{c\in\mathcal{C}} \left(\|c\|_2^2 - 2\langle v, c\rangle\right)$. Although $L_{\mathcal{C}}$ is not globally Lipschitz because of the term $\|v\|_2^2$, the function $g_{\mathcal{C}}$ is globally Lipschitz since for any $v, w \in \mathbb{R}^k$,

$$\begin{aligned} |g_{\mathcal{C}}(v) - g_{\mathcal{C}}(w)| &\le \sup_{c\in\mathcal{C}} \left|\left(\|c\|_2^2 - 2\langle v, c\rangle\right) - \left(\|c\|_2^2 - 2\langle w, c\rangle\right)\right| \\ &= 2\sup_{c\in\mathcal{C}} |\langle v - w, c\rangle| \le 2B\|v - w\|_2. \end{aligned}$$

Thus, $g_{\mathcal{C}}$ is $2B$-Lipschitz.

Step 5: Final gap assembly via quadratic cancellation. Let $Z \sim \mathcal{N}(0, I_k)$. We first show that the quadratic terms agree in expectation. Conditioned on $y = R_2\widetilde{x}$, each coordinate of the block $U_{0:k-1} = HD_3y$ has conditional second moment $\mathbb{E}\left[U_i^2 \mid y\right] = \sum_{j=0}^{d-1} H_{ij}^2 y_j^2 = \|y\|_2^2 = 1$. Therefore, $\mathbb{E}\|U_{0:k-1}\|_2^2 = \sum_{i=0}^{k-1} \mathbb{E}[U_i^2] = k$. On the other hand, since $Z \sim \mathcal{N}(0, I_k)$, $\mathbb{E}\|Z\|_2^2 = k$. Hence the quadratic parts cancel: $\mathbb{E}\|U_{0:k-1}\|_2^2 - \mathbb{E}\|Z\|_2^2 = 0$.

Using the decomposition from Step 4, we now obtain

$$\left|\mathbb{E}\Big[\min_{c\in\mathcal{C}} \|U_{0:k-1} - c\|_2^2\Big] - \mathbb{E}\Big[\min_{c\in\mathcal{C}} \|Z - c\|_2^2\Big]\right| = \left|\mathbb{E}[g_{\mathcal{C}}(U_{0:k-1})] - \mathbb{E}[g_{\mathcal{C}}(Z)]\right|.$$

Since $g_{\mathcal{C}}$ is $2B$-Lipschitz, the definition of the multivariate W_1 distance yields

$$|\mathbb{E}[g_{\mathcal{C}}(U_{0:k-1})] - \mathbb{E}[g_{\mathcal{C}}(Z)]| \le 2B\, W_1\Big(U_{0:k-1}, \mathcal{N}(0, I_k)\Big).$$

Applying the Wasserstein estimate from Step 3 yields

$$\left|\mathbb{E}\Big[\min_{c\in\mathcal{C}} \|U_{0:k-1} - c\|_2^2\Big] - \mathbb{E}\Big[\min_{c\in\mathcal{C}} \|Z - c\|_2^2\Big]\right| \le \mathcal{O}_{k,B}\left(\frac{1}{\sqrt{d}}\right).$$

This concludes the proof. □

C.4.2 Auxiliary lemma for multivariate normal approximation via Stein's method

We use the following estimate in the proof of Theorem 6. The point of the lemma is to allow the covariance of the summands to be close to I_k, rather than exactly equal to I_k.

Lemma 4 (Multivariate W_1 normal approximation with covariance mismatch)**.** *Fix $k \geq 1$, and let $X_1, \ldots, X_n$ be independent mean-zero random vectors in $\mathbb{R}^k$ with finite third moments. Define $S = \sum_{j=1}^n X_j$ and $\Sigma = \mathrm{Cov}(S)$, and let $Z \sim \mathcal{N}(0, I_k)$. Then*

$$W_1(S, Z) \leq A_k \left(\|\Sigma - I_k\|_F + \sum_{j=1}^n \mathbb{E}\|X_j\|_2^3 \right),$$

where $A_k < \infty$ depends only on k. Specifically, one may take $A_k = 4C_k + 4\sqrt{k} + 1$ where $C_k = 11.1 + 0.83 \log k$.

Proof. For clarity, we divide the proof into six steps.

Step 1: The identity-covariance estimate. We first present the standard covariance case that follows from [30, Theorem 2.15 and Eq. (3.5)].

Let $Y_1, \ldots, Y_n$ be independent mean-zero random vectors in $\mathbb{R}^k$, and suppose $\mathrm{Cov}\left(\sum_{j=1}^n Y_j\right) = I_k$. Let $T = \sum_{j=1}^n Y_j$. For a test function $f : \mathbb{R}^k \to \mathbb{R}$, write $M_1(f)$ for its Lipschitz constant with respect to the Euclidean norm. Then, the bound by [30, Theorem 2.15 and Eq. (3.5)] yields:

$$|\mathbb{E}f(T) - \mathbb{E}f(Z)| \leq M_1(f) \sum_{j=1}^n \mathbb{E}\left[\|Y_j\|_2^2 \min\left\{4.5, C_k\|Y_j\|_2\right\}\right],$$

where $C_k = 11.1 + 0.83 \log k$. Since $\min\left\{4.5, C_k\|Y_j\|_2\right\} \leq C_k\|Y_j\|_2$, we obtain

$$|\mathbb{E}f(T) - \mathbb{E}f(Z)| \leq C_k M_1(f) \sum_{j=1}^n \mathbb{E}\|Y_j\|_2^3.$$

Taking the supremum over all 1-Lipschitz functions f yields: $W_1(T, Z) \leq C_k \sum_{j=1}^n \mathbb{E}\|Y_j\|_2^3$.

Step 2: Notation for the general covariance case. We now return to the original summands $X_1, \ldots, X_n$. Define $r = \|\Sigma - I_k\|_F$ and $B = \sum_{j=1}^n \mathbb{E}\|X_j\|_2^3$. We prove the desired estimate by splitting into two cases according to the size of r.

Step 3: The small covariance-mismatch case. Assume that $r \leq \frac{1}{2}$. Since Σ is a covariance matrix, it is symmetric positive semidefinite. Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of Σ. Then, the eigenvalues of $\Sigma - I_k$ are $\lambda_1 - 1, \ldots, \lambda_k - 1$. Therefore, $r^2 = \|\Sigma - I_k\|_F^2 = \sum_{\ell=1}^k (\lambda_\ell - 1)^2$. Hence, for every $\ell = 1, \ldots, k$,

$$|\lambda_\ell - 1| \leq \left(\sum_{m=1}^k (\lambda_m - 1)^2 \right)^{1/2} = r \leq \frac{1}{2}.$$

Consequently, $\frac{1}{2} \leq \lambda_\ell \leq \frac{3}{2}$ for all $\ell = 1, \ldots, k$. In particular, Σ is positive definite, so $\Sigma^{-1/2}$ is well-defined.

Define the standardized summands $\widetilde{X}_j = \Sigma^{-1/2} X_j$ and $\widetilde{S} = \sum_{j=1}^n \widetilde{X}_j = \Sigma^{-1/2} S$. The random vectors $\widetilde{X}_1, \ldots, \widetilde{X}_n$ are independent and mean zero. Moreover,

$$\mathrm{Cov}(\widetilde{S}) = \Sigma^{-1/2} \mathrm{Cov}(S) \Sigma^{-1/2} = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = I_k.$$

Applying the identity-covariance estimate from Step 1 gives

$$W_1(\widetilde{S}, Z) \leq C_k \sum_{j=1}^n \mathbb{E}\|\Sigma^{-1/2} X_j\|_2^3.$$

We now bound the third moments after this linear change of variables. For a matrix A, define its Euclidean operator norm by $\|A\|_{\mathrm{op}} = \sup_{\|u\|_2=1} \|Au\|_2$. Equivalently, $\|A\|_{\mathrm{op}}$ is the smallest number L such that $\|Av\|_2 \le L\|v\|_2$ for every $v \in \mathbb{R}^k$. Indeed, if $v \neq 0$, then $u = v/\|v\|_2$ has unit norm, and therefore $\|Av\|_2 = \|v\|_2 \left\|A\frac{v}{\|v\|_2}\right\|_2 \le \|v\|_2\|A\|_{\mathrm{op}}$. The case $v = 0$ is trivial.

Applying this with $A = \Sigma^{-1/2}$ and $v = X_j$, we obtain $\|\Sigma^{-1/2}X_j\|_2 \le \|\Sigma^{-1/2}\|_{\mathrm{op}}\|X_j\|_2$. Cubing both sides and taking expectations gives $\mathbb{E}\|\Sigma^{-1/2}X_j\|_2^3 \le \|\Sigma^{-1/2}\|_{\mathrm{op}}^3\mathbb{E}\|X_j\|_2^3$. Summing over j yields $\sum_{j=1}^n \mathbb{E}\|\Sigma^{-1/2}X_j\|_2^3 \le \|\Sigma^{-1/2}\|_{\mathrm{op}}^3 B$. Thus,

$$W_1(\widetilde{S}, Z) \le C_k\|\Sigma^{-1/2}\|_{\mathrm{op}}^3 B.$$

Since the eigenvalues of Σ lie in $[1/2, 3/2]$, the eigenvalues of $\Sigma^{-1/2}$ lie in $\left[\sqrt{\frac{2}{3}}, \sqrt{2}\right]$. Because $\Sigma^{-1/2}$ is symmetric positive definite, its operator norm is its largest eigenvalue. Therefore, $\|\Sigma^{-1/2}\|_{\mathrm{op}} \le \sqrt{2}$. Substituting this into the preceding bound gives

$$W_1(\widetilde{S}, Z) \le 2\sqrt{2}\, C_k B.$$

Step 4: Returning from the standardized covariance to I_k. By the triangle inequality,

$$W_1(S, Z) \le W_1(S, \Sigma^{1/2}Z) + W_1(\Sigma^{1/2}Z, Z).$$

For the first term, use the identity $S = \Sigma^{1/2}\widetilde{S}$. We also use the following Lipschitz property of W_1: for any matrix A and any random vectors P, Q in $\mathbb{R}^k$,

$$W_1(AP, AQ) \le \|A\|_{\mathrm{op}} W_1(P, Q).$$

To see this directly from the dual definition of W_1, let f be any 1-Lipschitz function. Define $g(v) := f(Av)$. Then g is $\|A\|_{\mathrm{op}}$-Lipschitz, since

$$|g(v) - g(w)| = |f(Av) - f(Aw)| \le \|A(v - w)\|_2 \le \|A\|_{\mathrm{op}}\|v - w\|_2.$$

If $\|A\|_{\mathrm{op}} = 0$, then the desired inequality is trivial. Otherwise, set $h(v) := \frac{g(v)}{\|A\|_{\mathrm{op}}}$. Then h is 1-Lipschitz. Hence, by the dual definition of W_1,

$$\begin{aligned}|\mathbb{E}f(AP) - \mathbb{E}f(AQ)| &= |\mathbb{E}g(P) - \mathbb{E}g(Q)| \\ &= \|A\|_{\mathrm{op}}\,|\mathbb{E}h(P) - \mathbb{E}h(Q)| \le \|A\|_{\mathrm{op}} W_1(P, Q).\end{aligned}$$

Taking the supremum over all 1-Lipschitz functions f gives

$$W_1(AP, AQ) \le \|A\|_{\mathrm{op}} W_1(P, Q).$$

Applying this property with $A = \Sigma^{1/2}$, $P = \widetilde{S}$, and $Q = Z$ gives

$$W_1(S, \Sigma^{1/2}Z) = W_1(\Sigma^{1/2}\widetilde{S}, \Sigma^{1/2}Z) \le \|\Sigma^{1/2}\|_{\mathrm{op}} W_1(\widetilde{S}, Z).$$

Since the eigenvalues of Σ lie in $[1/2, 3/2]$, the eigenvalues of $\Sigma^{1/2}$ lie in $\left[\frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}}\right]$. Thus, $\|\Sigma^{1/2}\|_{\mathrm{op}} \le \sqrt{\frac{3}{2}}$. Combining this with the bound from Step 3 gives

$$W_1(S, \Sigma^{1/2}Z) \le \sqrt{\frac{3}{2}} \cdot 2\sqrt{2}\, C_k B = 2\sqrt{3}\, C_k B \le 4C_k B.$$

For the second term, we compare $\Sigma^{1/2}Z$ and Z using the same Gaussian vector Z. For any 1-Lipschitz function f, $\left|\mathbb{E}f(\Sigma^{1/2}Z) - \mathbb{E}f(Z)\right| \le \mathbb{E}\left[\left|f(\Sigma^{1/2}Z) - f(Z)\right|\right] \le \mathbb{E}\|(\Sigma^{1/2} - I_k)Z\|_2$. Taking the supremum over all 1-Lipschitz f yields,

$$W_1(\Sigma^{1/2}Z, Z) \le \mathbb{E}\|(\Sigma^{1/2} - I_k)Z\|_2.$$

By Cauchy–Schwarz,

$$\mathbb{E}\|(\Sigma^{1/2} - I_k)Z\|_2 \le \left(\mathbb{E}\|(\Sigma^{1/2} - I_k)Z\|_2^2\right)^{1/2}.$$

Since $Z \sim \mathcal{N}(0, I_k)$,

$$\mathbb{E}\|(\Sigma^{1/2} - I_k)Z\|_2^2 = \|\Sigma^{1/2} - I_k\|_F^2.$$

Indeed, if $M = \Sigma^{1/2} - I_k$, then

$$\mathbb{E}\|MZ\|_2^2 = \mathbb{E}[Z^\top M^\top MZ] = \mathrm{tr}(M^\top M\mathbb{E}[ZZ^\top]) = \mathrm{tr}(M^\top M) = \|M\|_F^2.$$

Therefore,

$$W_1(\Sigma^{1/2}Z, Z) \le \|\Sigma^{1/2} - I_k\|_F.$$

It remains to compare $\|\Sigma^{1/2} - I_k\|_F$ with $\|\Sigma - I_k\|_F$. Since Σ is symmetric positive semidefinite, it has an orthonormal eigenbasis. In that basis, the eigenvalues of $\Sigma^{1/2} - I_k$ are $\sqrt{\lambda_\ell} - 1$, while the eigenvalues of $\Sigma - I_k$ are $\lambda_\ell - 1$. For every $\lambda_\ell \ge 0$,

$$|\sqrt{\lambda_\ell} - 1| = \frac{|\lambda_\ell - 1|}{\sqrt{\lambda_\ell} + 1} \le |\lambda_\ell - 1|.$$

Hence, $\|\Sigma^{1/2} - I_k\|_F \le \|\Sigma - I_k\|_F = r$. Thus, $W_1(\Sigma^{1/2}Z, Z) \le r$.

Combining the two terms, in the case $r \le 1/2$ we obtain

$$W_1(S, Z) \le r + 4C_k B.$$

Step 5: The large covariance-mismatch case. Assume that $r > \frac{1}{2}$. We use a crude bound that depends only on the second moments. Let f be any 1-Lipschitz function. Subtracting the constant $f(0)$ does not change the difference of expectations, so

$$\begin{aligned}|\mathbb{E}f(S) - \mathbb{E}f(Z)| &= |\mathbb{E}[f(S) - f(0)] - \mathbb{E}[f(Z) - f(0)]| \\ &\le \mathbb{E}|f(S) - f(0)| + \mathbb{E}|f(Z) - f(0)| \le \mathbb{E}\|S\|_2 + \mathbb{E}\|Z\|_2.\end{aligned}$$

Taking the supremum over all 1-Lipschitz f gives

$$W_1(S, Z) \le \mathbb{E}\|S\|_2 + \mathbb{E}\|Z\|_2.$$

By Jensen's inequality, $\mathbb{E}\|S\|_2 \le \left(\mathbb{E}\|S\|_2^2\right)^{1/2}$ and $\mathbb{E}\|Z\|_2 \le \left(\mathbb{E}\|Z\|_2^2\right)^{1/2}$.

Since $\mathbb{E}S = 0$ and $\mathrm{Cov}(S) = \Sigma$, $\mathbb{E}\|S\|_2^2 = \mathrm{tr}(\Sigma)$. Also, since $Z \sim \mathcal{N}(0, I_k)$, $\mathbb{E}\|Z\|_2^2 = k$. Therefore,

$$W_1(S, Z) \le \sqrt{\mathrm{tr}(\Sigma)} + \sqrt{k}.$$

We now bound $\mathrm{tr}(\Sigma)$. Since $\mathrm{tr}(\Sigma) = k + \mathrm{tr}(\Sigma - I_k)$, and since $\mathrm{tr}(\Sigma - I_k) = \langle \Sigma - I_k, I_k\rangle_F$, Cauchy's inequality for the Frobenius inner product yields, $\mathrm{tr}(\Sigma - I_k) \le \|\Sigma - I_k\|_F \|I_k\|_F = r\sqrt{k}$. Thus, $\mathrm{tr}(\Sigma) \le k + r\sqrt{k}$. Consequently, $\sqrt{\mathrm{tr}(\Sigma)} \le \sqrt{k + r\sqrt{k}} \le \sqrt{k} + r$, because $(\sqrt{k} + r)^2 = k + 2r\sqrt{k} + r^2 \ge k + r\sqrt{k}$. Hence, $W_1(S, Z) \le 2\sqrt{k} + r$. Since $r > 1/2$, $2\sqrt{k} \le 4\sqrt{k}\, r$.

Therefore, in the case $r > 1/2$,

$$W_1(S, Z) \le (4\sqrt{k} + 1)r.$$

Step 6: Combining the two cases. If $r \le 1/2$, Step 4 yields,

$$W_1(S, Z) \le r + 4C_k B \le (4C_k + 4\sqrt{k} + 1)(r + B).$$

If $r > 1/2$, Step 5 yields,

$$W_1(S, Z) \le (4\sqrt{k} + 1)r \le (4C_k + 4\sqrt{k} + 1)(r + B).$$

Thus, in all cases,

$$W_1(S, Z) \le A_k(r + B),$$

where $A_k = 4C_k + 4\sqrt{k} + 1$. Substituting the definitions of r and B gives

$$W_1(S, Z) \le A_k \left(\|\Sigma - I_k\|_F + \sum_{j=1}^n \mathbb{E}\|X_j\|_2^3 \right).$$

This proves the lemma. □

D Adaptive linear-time verification

Because practical data distributions are often not adversarial, we can dynamically adapt the required number of RHTs. Our proofs show that the number of required RHT layers is governed by two quantities of the normalized input $\widetilde{x} = \frac{x}{\|x\|_2}$. For the scalar marginal normal approximation, the relevant quantity is the cubed ℓ_3 mass $\rho_3(\widetilde{x}) = \sum_{r=0}^{d-1} |\widetilde{x}_r|^3$.

For deciding whether the preliminary RHT in the VQ construction can be skipped, the relevant quantity is the squared infinity norm, $\|\widetilde{x}\|_\infty^2$. Thus, before applying the worst-case number of RHTs, a system can first check whether the input is already sufficiently flat.

Scalar check. If $\rho_3(\widetilde{x}) \le \eta_3$, then the one-coordinate Berry–Esseen step used in Theorems 1 and 2 already applies to a single RHT. Namely, for every fixed coordinate $U_i = \sqrt{d}\,(\widetilde{H}D\widetilde{x})_i$, we have $d_K(U_i, G) \le 0.5606\,\eta_3,\ W_1(U_i, G) \le C_W \eta_3$, where $G \sim \mathcal{N}(0,1)$. Therefore, if $\eta_3 = \mathcal{O}(d^{-1/2})$, then 1-RHT already gives the same asymptotic guarantees as the worst-case 2-RHT theorem.

Vector check. For VQ, scalar marginal normality is insufficient; the relevant quantity is the conditional covariance between coordinates in a fixed block. The worst-case 3-RHT guarantee in Theorem 5 gives an $O(\frac{1}{\sqrt{d}})$ RMS conditional covariance bound. If we want to skip the preliminary smoothing RHT and use only two RHTs, then we apply the fixed-input part of Theorem 5 directly to the current input. More precisely, let $U = \sqrt{d}\,\widetilde{H}D_b\widetilde{H}D_a\widetilde{x}$ be the 2-stage RHT of $\widetilde{x}$. For distinct coordinates $i \ne j$, define the conditional covariance $C_{ij}(\widetilde{x}, D_a) = \mathrm{Cov}_{D_b}(U_i, U_j \mid \widetilde{x}, D_a)$. The fixed-input bound in Theorem 5, applied with $y = \widetilde{x}$, gives $\mathbb{E}_{D_a}\big[C_{ij}(\widetilde{x}, D_a)^2\big] \le 2\|\widetilde{x}\|_\infty^2$. Equivalently, if $\|\widetilde{x}\|_\infty^2 \le \eta_\infty$, then $\big(\mathbb{E}_{D_a}\big[C_{ij}(\widetilde{x}, D_a)^2\big]\big)^{1/2} \le \sqrt{2\eta_\infty}$.

Thus, the adaptive check transfers into the VQ guarantee: using two RHTs on such an input gives a covariance contribution of order $\sqrt{\eta_\infty}$, and hence a VQ error gap of order $\mathcal{O}_{k,B}\Big(\sqrt{\eta_\infty} + \frac{1}{\sqrt{d}}\Big)$.

In particular, if $\eta_\infty = \mathcal{O}(\frac{1}{d})$, then two RHTs match the worst-case 3-RHT bound $\mathcal{O}_{k,B}(\frac{1}{\sqrt{d}})$. If instead $\eta_\infty = \mathcal{O}(\frac{\log d}{d})$, then two RHTs give a weaker but still vanishing guarantee $\mathcal{O}_{k,B}\left(\sqrt{\frac{\log d}{d}}\right)$.

Implementation. Both checks can be evaluated in one pass over x. Define $S_2 = \sum_{r=0}^{d-1} x_r^2$, $S_3 = \sum_{r=0}^{d-1} |x_r|^3$, $M_2 = \max_{0 \le r < d} x_r^2$. For $x \ne 0$, $\rho_3(\widetilde{x}) = \frac{S_3}{S_2^{3/2}},\ \|\widetilde{x}\|_\infty^2 = \frac{M_2}{S_2}$. Therefore, the adaptive GPU-friendly verification step costs $\mathcal{O}(d)$ time and $\mathcal{O}(1)$ space.

In summary, scalar quantization methods such as [9, 37] can use the $\rho_3(\widetilde{x})$ check to decide whether one RHT is already sufficient. VQ methods can use the $\|\widetilde{x}\|_\infty^2$ check to decide whether two RHTs already provide the desired covariance bound with vanishing-term guarantee determined by the input's ℓ_∞ norm.

E Additional experimental results

E.1 Fixed-codebook HIGGS experiment

We compare one, two, and three RHTs using the released 3-bit HIGGS $p = 2$ codebook. For each $d \in \{128, 512, 2048, 8192\}$, we place 32 consecutive independent Rademacher entries at a random position in an otherwise zero vector. We quantize pairs of rotated coordinates with their nearest codeword. We repeat this experiment over 12 random inputs, each rotated with 8 different seeds.

Table 4 shows that three RHTs reduce vNMSE by 6.74–8.59% compared with one RHT and by 1.77–3.10% compared with two. At $d = 128, 512, 2048$, the measured relative vNMSE difference from URR is at most 0.06% in magnitude.

d	3 vs. 1 RHT	3 vs. 2 RHTs	3 RHTs vs. URR
128	$8.59\ [8.23, 8.95]$	$1.77\ [1.20, 2.34]$	$-0.06\ [-0.23, +0.10]$
512	$7.68\ [7.21, 8.14]$	$2.67\ [2.25, 3.09]$	$+0.04\ [-0.32, +0.39]$
2048	$7.15\ [5.89, 8.39]$	$2.69\ [2.21, 3.16]$	$+0.05\ [-0.30, +0.40]$
8192	$6.74\ [4.60, 8.84]$	$3.10\ [2.61, 3.59]$	Timed out

Table 4: Relative HIGGS vNMSE reductions from three RHTs over one and two RHTs, and differences from URR. Values are percentages with paired 95% confidence intervals.

E.2 Adaptive scalar and vector checks

At $d = 512$, we use the scalar check in Section 7 to choose one or two RHTs for EDEN, and the vector check to choose two or three for HIGGS. We set $c_3 = 3.9$ and $c_\infty = 2.5$ using separate calibration inputs, then keep them fixed. We repeat each experiment over 12 random input batches of 64 vectors, each rotated with 8 different seeds.

Scalar check. For 3-bit EDEN, each random input batch contains 58 Gaussian vectors and six sparse vectors. For each sparse vector, we place 32 consecutive independent Rademacher entries at a random position in an otherwise zero vector. The check uses one RHT for every Gaussian vector and two for every sparse vector. In Table 5, adaptive vNMSE is 0.03437, close to 0.03434 from always using two RHTs, while only 9.375% of vectors receive a second RHT. On entirely sparse inputs, two RHTs reduce vNMSE by 14.99% compared with one.

Method	vNMSE	Change vs. one RHT (95% CI)	Second-RHT fraction
Always one	0.03505	—	0%
Adaptive one/two	0.03437	$-1.95\%\ [-2.32, -1.58]$	9.375%
Always two	0.03434	$-2.04\%\ [-2.55, -1.52]$	100%

Table 5: 3-bit EDEN on mixed Gaussian and sparse inputs. Changes and paired 95% confidence intervals are in percent.

Vector check. For 3-bit HIGGS $p = 2$, each random input batch contains 58 Gaussian vectors and six sparse vectors. For each sparse vector, we place 32 consecutive independent Rademacher entries at a random position in an otherwise zero vector. For 2-bit HIGGS, each random input batch contains 58 dense vectors with independent Rademacher entries and six sparse vectors, each formed by placing two consecutive independent Rademacher entries at a random position in an otherwise zero vector. We quantize pairs of rotated coordinates with their nearest codeword. In Table 6, adaptive vNMSE is close to always using three RHTs, while only 12.1% and 9.4% of vectors receive a third RHT. On entirely sparse inputs, three RHTs improve over both one and two RHTs.

	3-bit, 32 nonzeros	2-bit, 2 nonzeros
Three vs. one, all sparse	$-7.47\%\ [-8.03, -6.92]$	$-23.14\%\ [-23.28, -23.01]$
Three vs. two, all sparse	$-2.74\%\ [-3.16, -2.30]$	$-13.39\%\ [-14.25, -12.52]$
Third-RHT fraction, mixed	12.1%	9.4%
Adaptive vNMSE, mixed	0.02949	0.10719
Always-three vNMSE, mixed	0.02949	0.10728

Table 6: Adaptive HIGGS results on sparse and mixed inputs. Brackets give paired 95% confidence intervals in percent.

Real tensors. We test 1968 vectors of dimension 512 from 21 Qwen2.5-0.5B tensors: weights, hidden activations, KV cache, and gradients. We apply 3-bit EDEN and 3-bit HIGGS $p = 2$ using the thresholds above. We repeat each experiment with 8 different seeds for each input vector.

Data type	Vectors	Scalar: one / two	Vector: two / three
Weights	768	764 / 4	520 / 248
Activations	336	171 / 165	1 / 335
KV cache	96	96 / 0	14 / 82
Gradients	768	577 / 191	62 / 706
Total	1968	1608 / 360	597 / 1371

Table 7: RHT counts chosen for the Qwen2.5-0.5B vectors.

The scalar check skips the second RHT for 81.7% of vectors, and the vector check skips the third for 30.3% (Table 7). Adaptive HIGGS and always using three RHTs both keep vNMSE below 0.05 in every trial. The check selects a third RHT for the only two-RHT result above 0.05, reducing vNMSE from 0.05049 to 0.02887.

E.3 Checks of scalar marginal approximation

We test e_0 and $(e_0 + e_1)/\sqrt{2}$ for $d \in \{64, 128, \ldots, 4096\}$. We compare $\sqrt{d}$ times an output coordinate with a standard Gaussian using Kolmogorov and 1-Wasserstein distances. The one- and two-RHT distances are exact; for three RHTs, we repeat the experiment with 8 batches of 128 different seeds and average the resulting estimates.

Tables 8 and 9 multiply the distances by $\sqrt{d}$. For both inputs, the two-RHT values stay nearly constant, consistent with the $d^{-1/2}$ bounds. Three RHTs give smaller distances at every dimension.

	$\sqrt{d}\, d_K$			$\sqrt{d}\, W_1$		
d	One	Two	Three	One	Two	Three
64	2.73076	0.39740	0.17195	4.28302	0.50058	0.17651
128	3.86188	0.39820	0.13101	6.05710	0.50030	0.13078
256	5.46152	0.39857	0.10259	8.56603	0.50015	0.09999
512	7.72376	0.39874	0.07733	12.11420	0.50007	0.07217
1024	10.92304	0.39886	0.06589	17.13206	0.50005	0.06950
2048	15.44751	0.39892	0.05399	24.22840	0.50004	0.05171
4096	21.84608	0.39893	0.05022	34.26413	0.50003	0.05298

Table 8: Distances to a standard Gaussian for e_0, scaled by $\sqrt{d}$.

	$\sqrt{d}\, d_K$			$\sqrt{d}\, W_1$		
d	One	Two	Three	One	Two	Three
64	2.00000	0.55979	0.23849	3.00821	0.70873	0.23901
128	2.82843	0.56201	0.18322	4.25425	0.70793	0.17458
256	4.00000	0.56313	0.12687	6.01642	0.70753	0.11949
512	5.65685	0.56367	0.09689	8.50850	0.70731	0.09130
1024	8.00000	0.56390	0.08693	12.03283	0.70720	0.07932
2048	11.31371	0.56407	0.06581	17.01699	0.70718	0.06726
4096	16.00000	0.56416	0.05262	24.06566	0.70716	0.05566

Table 9: Distances to a standard Gaussian for $(e_0 + e_1)/\sqrt{2}$, scaled by $\sqrt{d}$.

F Approximate Nearest Neighbor Search

Setup. We evaluate RaBitQ [21] after one or two full RHTs on the six datasets in Table 10. For each dataset, we use a single center c, computed as the mean of the first 4096 training vectors, and zero-pad the residuals to the next power of two. We use RaBitQ's native per-residual scale selection at 2, 3, and 4 bits per transformed coordinate. The experiments use five random seeds, with the same D_1 diagonal for the 1-RHT and 2-RHT variants.

Dataset	Size	Dimension
MSong	994,185	420
Image	2,340,373	150
GIST	1,000,000	960
Word2Vec	1,000,000	300
SIFT	1,000,000	128
Deep1M	1,000,000	256

Table 10: ANN datasets. Dimensions are reported before zero-padding to the next power of 2.

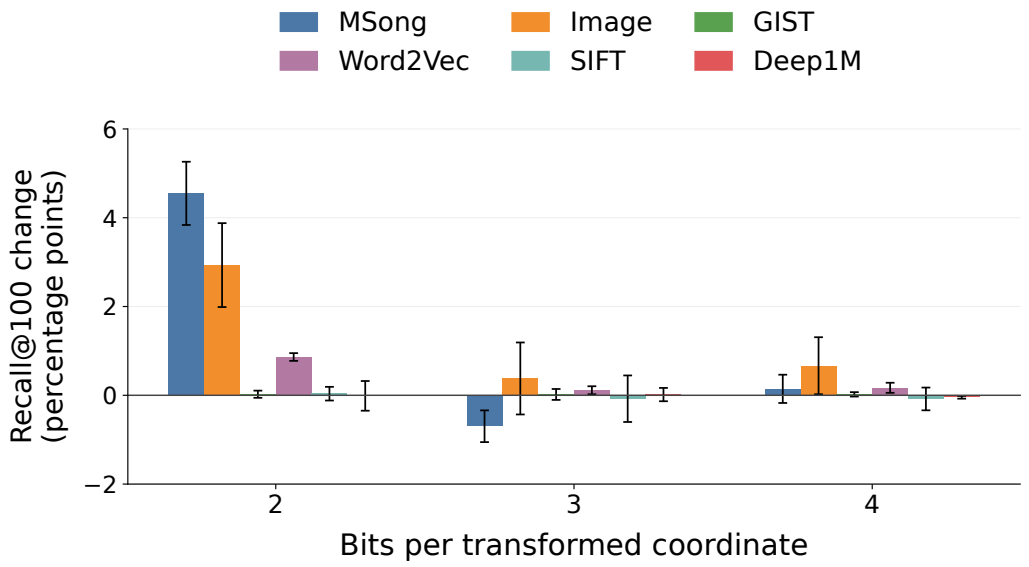

Figure 1: Change in RaBitQ Recall@100 for two full RHTs relative to one RHT, in percentage points, at 2–4 bits per transformed coordinate. Each group contains one bar per dataset. Bars show mean differences over five paired random seeds; error bars show 95% Student-t confidence intervals computed from the paired differences, with the database and queries fixed. Positive values indicate higher recall.

For each query q, we rank all database vectors using the distance estimate

$$\widehat{d}(q,x) = \|q-c\|_2^2 + \|x-c\|_2^2 - 2\langle R(q-c), \widehat{y}_x\rangle,$$

where R is the composition of the RHTs and $\widehat{y}_x$ is RaBitQ's decoded representation of $R(x-c)$ at its native reconstruction scale. Recall@100 is the fraction of the exact 100 nearest neighbors recovered among the 100 vectors with the smallest estimated distances. We average recall over queries and then over the five seeds.

Results. Figure 1 reports the change in Recall@100 from one to two RHTs. At 2 bits, two RHTs improve mean recall by 4.55 percentage points on MSong and 2.93 percentage points on Image. These gains depend on the bit budget: at 3 bits, MSong's recall decreases by approximately 0.70 percentage points with two RHTs. The changes on GIST, SIFT, and Deep1M are small. Thus, additional mixing can improve recall, but its benefit depends on the dataset and quantization regime.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's main theoretical contributions: scalar marginal convergence for two RHTs, applications to DRIVE and QUIC-FL, three-RHT decorrelation codebook universality for VQ, and the adaptive linear-time check. The formal sections and appendix provide the corresponding theorem statements and proofs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses scope limitations in the introduction, the VQ section, the adaptive verification section, and the conclusion, including the distinction between scalar quantization and block-wise VQ and the possible need for stronger results for non-fixed-size adaptive quantizers.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper is primarily theoretical. Assumptions are stated in the theorem and lemma statements, and proofs are included in the main text or deferred appendix sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendices E and F describe the input generation, quantizers, parameters, repetitions, and metrics. The ANN experiments follow the setup of RaBitQ [21], with our comparison of one and two RHTs specified in Appendix F. The remaining experiments use synthetic inputs and tensors from the publicly released Qwen2.5-0.5B model.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The RaBitQ repository provides code and data-preparation instructions for the ANN setup [21]. The Qwen model and HIGGS implementation are public, and Appendix E specifies the synthetic inputs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendices E and F give the input distributions and datasets, quantizers, bit budgets, thresholds, repetitions, and evaluation metrics. For the Qwen test, we use 32 tokens each from WikiText-2 (training split) and Penn Treebank (test split), and collect tensors from model blocks 0, 12, and 23. The model is used in its released form.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The HIGGS and EDEN comparisons in Appendix E report paired 95% confidence intervals. Figure 1 reports 95% Student-t confidence intervals over five paired random seeds, with the database and queries fixed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We report quantization accuracy and ANN recall. We do not report hardware, memory, or execution-time details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical work on randomized transforms and quantization analysis and does not involve human subjects, private data, or high-risk data/model release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

    Justification: The paper is theoretical, and it does not introduce a deployed system, dataset, or model with a direct new misuse path; thus, there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release models, datasets, or other assets that would require responsible-release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the public ANN datasets from the RaBitQ [21] setup for research and credit their original sources: SIFT/GIST, Deep1M, MSong (research publication permitted), ImageNet (noncommercial research), and Word2Vec (project: Apache 2.0). RaBitQ, HIGGS/Transformers, and Qwen2.5-0.5B use Apache 2.0; WikiText-2 lists CC BY-SA 3.0/GFDL, and Penn Treebank text states research-only use. We follow these stated terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new datasets, code packages, benchmarks, or models as assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research, so IRB or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We added the following declaration in the paper: "We have used LLMs to explore different proof routes for some of our results through iterative discussion and refinement; to provide an additional source of verification of our results; to correctly identify attribution; to simplify proofs; to help with the evaluation; and to edit. We have ensured that all content is correct, original, and adheres to ethical and academic standards."

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.