BEYOND BERRY–ESSEEN: LOSS-AWARE DISTORTION TRANSFER FOR RANDOMIZED HADAMARD QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Random rotations make scalar quantization near-optimal, and fast structured rotations are used in practice in their place. For the two-stage randomized Hadamard transform (two-RHT), which costs $O(d \log d)$ time and $2d$ sign bits, recent work shows that each rotated coordinate is only $O(d^{-1/2})$-close to its dense-Haar law, which yields an $O(d^{-1/2})$ bound on the extra distortion. We prove that the distortion itself converges much faster. For every bit-width, every symmetric nearest-centroid codebook, and every unit input, the squared-error gap between two-RHT and dense Haar is at most CB_b/d with an explicit constant, and this rate is attained at one bit. The proof reduces the loss to a few absolute-value kinks, shows by a zero-bias coupling that Rademacher sums match the Gaussian on every kink up to $\|u\|_4^4$ rather than $\|u\|_3^3$, and uses that the first Hadamard stage makes $\mathbb{E}\,\|u\|_4^4 \le 3/d$ for every input. The same argument explains when one stage suffices, and we prove that one undithered stage leaves a non-vanishing gap on sparse inputs. As corollaries, the known two-RHT guarantee for DRIVE sharpens from $O(d^{-1/2})$ to $O(1/d)$, and the gap relative to the dense-rotation distortion is $O(B_b 4^b/d)$. Structured rotations are free when $d \gg 4^b$; at $d \le 4^b$ we measure worst-case gaps of 4–41%. Exact computation on basis and sparse inputs gives $|\mathrm{gap}| \cdot d \le 0.96$ up to $d = 16384$ and a relative gap of about $0.6 \cdot 4^b/d$ at every $b \le 4$. Two-RHT with the Lloyd–Max codebook has 1.2–2× lower worst-case distortion than a provable dithered single-stage quantizer at $b \le 4$ once $d \ge 256$, mostly because of the codebook. In million-scale retrieval two-RHT matches the dense rotation within seed noise; in LLM KV caches it tracks the dense rotation, but per-channel key statistics matter more than the rotation.

1 INTRODUCTION

Random rotation followed by scalar quantization is a basic primitive for compressing embeddings, gradients, and LLM key–value (KV) caches. TurboQuant (Zandieh et al., 2026) rotates a unit vector by a dense Haar-random matrix, which makes every coordinate follow the same known law, and then applies the Lloyd–Max codebook of that law; the resulting distortion is within a factor $\sqrt{3}\pi/2 \approx 2.72$ of Shannon's lower bound. The dense rotation costs $O(d^2)$ time per vector and $O(d^3)$ to generate, so practical systems replace it with randomized Hadamard transforms (RHTs) (Ashkboos et al., 2024; Tseng et al., 2024). This raises a precise question: *how close is the distortion of a structured rotation to that of dense Haar?*

Two recent lines of work answer it at the level of coordinate *laws*. Ben-Basat et al. (2026b) prove that every coordinate of a two-stage RHT $\boldsymbol{H}\boldsymbol{D}_2\boldsymbol{H}\boldsymbol{D}_3$ is $O(d^{-1/2})$-close to Gaussian in Kolmogorov and Wasserstein distance, uniformly over inputs, and derive $O(d^{-1/2})$ guarantees for DRIVE and QUIC-FL. Feng et al. (2026) show that *one* RHT with a random dither attains $(\pi\sqrt{3}/2 + o(1))4^{-b}$, with $o(1) \to 0$ as the bit-width $b \to \infty$. Neither result says how fast the distortion of the undithered quantizer used in practice converges at a fixed bit-width.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Distortion guarantees for rotation-based scalar quantization of a unit vector at b bits per coordinate. D^{Haar} is the dense-Haar (TurboQuant) distortion and B_b the largest centroid.

rotation	cost / storage	distortion guarantee
dense Haar (Zandieh et al., 2026)	$O(d^2)$ / seed + $O(d^3)$ setup	$D^{\text{Haar}} \le \frac{\sqrt{3}\pi}{2} 4^{-b}$
RHTs in LLM quantizers (Ashkboos et al., 2024; Tseng et al., 2024)	$O(d \log d)$ / $O(d)$ bits	none for scalar distortion
two-RHT, law-level (Ben-Basat et al., 2026b)	$O(d \log d)$ / $O(d)$ bits	$D^{\text{Haar}} + O(B_b/\sqrt{d})$
one RHT + dither (Feng et al., 2026)	$O(d \log d)$ / $O(d)$ bits	$(\frac{\pi\sqrt{3}}{2} + o_{b\to\infty}(1))\, 4^{-b}$
two-RHT, loss-level (ours)	$O(d \log d)$ / $2d$ bits	$D^{\text{Haar}} + O(B_b/d)$, every b and $\boldsymbol{x}$; tight

We show that the right yardstick is the loss, not the law. Distributional distances of two-RHT coordinates are genuinely $\Theta(d^{-1/2})$, because the coordinates are lattice-supported for basis inputs, yet the squared-error distortion converges at rate $1/d$. Our contributions:

- **Sharp transfer for every bit-width and every input** (Theorem 1). For every symmetric codebook, in particular Lloyd–Max at every b, and every unit input $\boldsymbol{x}$, $|D^{\text{two-RHT}}_{\texttt{mse}}(b, \boldsymbol{x}) - D^{\text{Haar}}_{\texttt{mse}}(b)| \le 2C_K B_b\big(\frac{3-2s_4(\boldsymbol{x})}{d} + \frac{3}{d+2}\big) \le 56 B_b/d$. The rate is attained at $b = 1$ (Proposition 2), the two-RHT DRIVE guarantee sharpens to $1 - \frac{2}{\pi} + O(1/d)$ (Corollary 1), and relative to the dense-rotation distortion the gap is $O(B_b 4^b/d)$ (Corollary 2).
- **A loss-level mechanism** (Lemma 2 and Proposition 1). The loss is the common second moment plus $2^b - 1$ absolute-value kinks. On every kink, a Rademacher sum with weights u is within $C_K\|u\|_4^4$ of the Gaussian, uniformly in the kink location and with no non-lattice condition, although its CDF error is $\Theta(\|u\|_3^3)$. The first Hadamard stage makes $\mathbb{E}\,\|u\|_4^4 \le 3/d$ for every input.
- **Depth, explained** (Corollary 3 and Proposition 3). One undithered stage is $O(1/d)$-accurate for ℓ_4-delocalized inputs, and on sparse inputs its gap converges to an explicit nonzero constant. Real embeddings are delocalized, so one stage already suffices on them; the second stage buys the guarantee for every input at the cost of one more fast transform.
- **Evidence.** Exact (Monte-Carlo-free) gaps up to $d = 16384$ give $|\text{gap}| \cdot d \le 0.96$, and the relative gap collapses onto $0.6 \cdot 4^b/d$ for every $b \le 4$ (Fig. 1c); a head-to-head with the dithered single stage of Feng et al. (2026) shows 1.2–2× lower worst-case distortion for two undithered stages with Lloyd–Max at $b \le 4$ and $d \ge 256$; and with the analyzed quantizer, two-RHT matches dense Haar in million-scale retrieval and tracks it in LLM KV caches.

What is not new. Randomized Hadamard rotations are standard. The split of the nearest-centroid loss into a quadratic term and a Lipschitz part is due to Ben-Basat et al. (2026b, Sec. 6); combined with their theorem it gives the $O(d^{-1/2})$ baseline of Theorem 2. Our contribution is the $O(1/d)$ rate and its mechanism, which no Kolmogorov or W_1 argument can deliver. The result of Feng et al. (2026) is complementary: it is asymptotic in b and uses dithering, whereas ours holds at every fixed b for the undithered quantizer. Table 1 summarizes the guarantees.

2 Setting

Let d be a power of 2, $\boldsymbol{H}_d$ the normalized Walsh–Hadamard matrix (entries $\pm d^{-1/2}$), and $\boldsymbol{D}_2, \boldsymbol{D}_3$ independent diagonal matrices of i.i.d. Rademacher signs. The *two-RHT* is $\boldsymbol{\Pi}_H := \boldsymbol{H}_d\boldsymbol{D}_2\boldsymbol{H}_d\boldsymbol{D}_3$; it costs two fast Walsh–Hadamard transforms and stores $2d$ bits (an output sign layer is redundant for symmetric codebooks, Proposition 4). A unit vector $\boldsymbol{x}$ is quantized by applying a b-bit codebook $\boldsymbol{C}_b$ to each scaled coordinate $\sqrt{d}(\boldsymbol{\Pi}\boldsymbol{x})_j$ and rotating back. TurboQuant uses a dense Haar $\boldsymbol{\Pi}$ and the Lloyd–Max codebook of the scaled sphere-coordinate law $P_d := \text{law}(\sqrt{d}\,\theta_1)$, $\theta \sim \text{Unif}(\mathbb{S}^{d-1})$. Write $L_{\boldsymbol{C}}(z) := \min_{c\in\boldsymbol{C}}(z-c)^2$, $B_b := \max_{c\in\boldsymbol{C}_b}|c|$, $s_4(\boldsymbol{x}) := \|\boldsymbol{x}\|_4^4$, and $Q_d(\boldsymbol{x}) := \text{law}(\sqrt{d}(\boldsymbol{\Pi}_H\boldsymbol{x})_j)$. Because all coordinates share the same law, the total vector MSEs are $D^{\text{two-RHT}}_{\texttt{mse}}(b, \boldsymbol{x}) = \mathbb{E}_{Q_d(\boldsymbol{x})}[L_{\boldsymbol{C}_b}]$ and $D^{\text{Haar}}_{\texttt{mse}}(b) = \mathbb{E}_{P_d}[L_{\boldsymbol{C}_b}]$, and we study the *distortion gap*

$$\text{gap}(b, \boldsymbol{x}, d) \;:=\; \mathop{\mathbb{E}}_{Y\sim Q_d(\boldsymbol{x})}[L_{\boldsymbol{C}_b}(Y)] - \mathop{\mathbb{E}}_{X\sim P_d}[L_{\boldsymbol{C}_b}(X)].$$

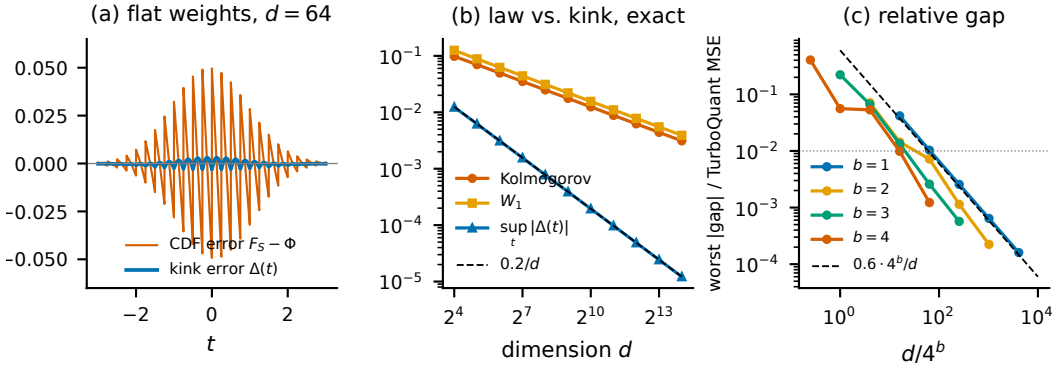


Figure 1: Why the loss converges faster than the law (all curves exact). (a) For flat weights ($S_u = d^{-1/2}\sum_k \eta_k$, the two-RHT coordinate of a basis input) the CDF error is a lattice sawtooth, while the error on a kink, $\Delta(t) = \mathbb{E}\,|S_u - t| - \mathbb{E}\,|G - t|$, is an order of magnitude smaller. (b) Kolmogorov and W_1 distances decay as $d^{-1/2}$; $\sup_t |\Delta(t)|$ decays as $0.2/d$ (Lemma 2). (c) Worst two-RHT gap over basis and sparse inputs relative to the TurboQuant distortion, against $d/4^b$: all bit-widths collapse onto $0.6 \cdot 4^b/d$ (Corollary 2).

Let $u(\delta) := \boldsymbol{H}_d \boldsymbol{D}_3 \boldsymbol{x}$ be the unit vector of *mixed weights* after the first stage, $\delta = \mathrm{diag}(\boldsymbol{D}_3)$. For each j, the signs $\sqrt{d}(\boldsymbol{H}_d)_{jk}(\boldsymbol{D}_2)_{kk}$ are i.i.d. Rademacher and independent of δ, so

$$\sqrt{d}\,(\boldsymbol{\Pi}_H \boldsymbol{x})_j \;\overset{d}{=}\; S_{u(\delta)}, \qquad S_u := \textstyle\sum_{k \le d} \eta_k u_k, \tag{1}$$

with i.i.d. Rademacher η. Thus $Q_d(\boldsymbol{x})$ is a δ-mixture of Rademacher-sum laws. All guarantees are in expectation over the signs.

3 Sharp distortion transfer

Lemma 1 (Second-moment cancellation; Ben-Basat et al., 2026b)**.** *For a symmetric codebook $\boldsymbol{C}$ with $B = \max|c_i|$, $L_{\boldsymbol{C}}(z) = z^2 + g_{\boldsymbol{C}}(z)$ with $g_{\boldsymbol{C}}(z) = \min_i(c_i^2 - 2c_i z)$ even, concave and $2B$-Lipschitz. If P, Q are centered with equal variance, $\mathbb{E}_Q[L_{\boldsymbol{C}}] - \mathbb{E}_P[L_{\boldsymbol{C}}] = \mathbb{E}_Q[g_{\boldsymbol{C}}] - \mathbb{E}_P[g_{\boldsymbol{C}}]$, which is at most $2B\,W_1(P, Q)$ in absolute value.*

With the W_1 bound of Ben-Basat et al. (2026b) this gives $|\mathrm{gap}| = O(B_b/\sqrt{d})$ (Theorem 2). No argument through W_1 or Kolmogorov distance can do better: for flat weights S_u lives on a lattice of spacing $2/\sqrt{d}$, and both distances to $N(0,1)$ are $\Theta(d^{-1/2})$ (Fig. 1b). Two ingredients give the faster rate.

Proposition 1 (ℓ_4 delocalization by the first stage)**.** *For every unit $\boldsymbol{x}$, $\mathbb{E}_\delta\,\|u(\delta)\|_4^4 = (3 - 2s_4(\boldsymbol{x}))/d \le 3/d$. Consequently $\kappa_4(Q_d(\boldsymbol{x})) = -2(3 - 2s_4(\boldsymbol{x}))/d$ and $\kappa_4(P_d) = -6/(d+2)$, where $\kappa_4(Z) = \mathbb{E}\,Z^4 - 3$.*

Lemma 2 (Kink lemma)**.** *For every unit $u \in \mathbb{S}^{d-1}$, every $t \in \mathbb{R}$ and $G \sim N(0,1)$,*

$$\big|\;\mathbb{E}\,|S_u - t| - \mathbb{E}\,|G - t|\;\big| \;\le\; C_K\,\|u\|_4^4, \qquad C_K := 3 + 2\sqrt{2}\,C_{\mathrm{BE}} \le 4.59,$$

where $C_{\mathrm{BE}} \le 0.56$ is the Berry–Esseen constant for non-identical summands (Shevtsova, 2010).

The bound is uniform in t and covers lattice-supported sums. The proof (Section A.3) writes the difference through Stein's equation for $h(w) = |w - t|$ and the zero-bias coupling (Goldstein & Reinert, 1997), which replaces each summand $\pm|u_k|$ by a uniform variable on $[-|u_k|, |u_k|]$. The resulting trapezoid-versus-mean error is second order in $|u_k|$; the kink itself enters only through an anti-concentration term, bounded by Berry–Esseen and Cauchy–Schwarz. Exact evaluation over flat, spiked and random weights gives a worst ratio of 0.205, so C_K is conservative but the $\|u\|_4^4$ scaling is attained by flat weights.

Theorem 1 (Sharp distortion transfer)**.** *Let $\boldsymbol{C}$ be any codebook symmetric about 0 with $B = \max_i |c_i|$. For every d that is a power of 2 and every unit $\boldsymbol{x}$,*

$$\Big| \mathop{\mathbb{E}}_{Q_d(\boldsymbol{x})}[L_{\boldsymbol{C}}] - \mathop{\mathbb{E}}_{P_d}[L_{\boldsymbol{C}}] \Big| \;\le\; 2B\,C_K\Big(\frac{3-2s_4(\boldsymbol{x})}{d} + \frac{3}{d+2}\Big) \;\le\; \frac{12\,C_K B}{d}.$$

In particular $|\mathrm{gap}(b, \boldsymbol{x}, d)| \le 56\,B_b/d$ for the Lloyd–Max codebook at every $b \ge 1$.

Proof outline (Section A.4). With $t_i = (c_i + c_{i+1})/2$ and $w_i = c_{i+1} - c_i$, the concave part decomposes exactly as $g_{\boldsymbol{C}}(z) = g_{\boldsymbol{C}}(0) - \sum_i w_i(|z - t_i| - |t_i|)$ with $\sum_i w_i = 2B$, so the gap is a weighted sum of kink differences. On the two-RHT side, apply Lemma 2 conditionally on δ via (1) and average with Proposition 1. On the Haar side, $P_d = \mathrm{law}(S_\theta)$ with θ uniform on the sphere and $\mathbb{E}\,\|\theta\|_4^4 = 3/(d+2)$. □

Corollary 1 (One bit and DRIVE)**.** *At $b = 1$, $|\mathrm{gap}(1, \boldsymbol{x}, d)| \le 12C_K/d$ for every unit $\boldsymbol{x}$. For DRIVE (Vargaftik et al., 2021) with rotation $\boldsymbol{\Pi}_H$, whose biased vNMSE is $1 - \frac{1}{d}\,\mathbb{E}\,\|\boldsymbol{\Pi}_H \boldsymbol{x}\|_1^2$, one has $\mathrm{vNMSE} \le 1 - \frac{2}{\pi} + \frac{22}{d}$, improving the $1 - \frac{2}{\pi} + O(d^{-1/2})$ bound of Ben-Basat et al. (2026b, Eq. (2), Sec. 4).*

Corollary 2 (Relative gap)**.** *For every $b \ge 1$ and unit $\boldsymbol{x}$, $|\mathrm{gap}(b, \boldsymbol{x}, d)|/D^{\mathrm{Haar}}_{\mathtt{mse}}(b) \le 56\,B_b\,4^b/d$.*

Proof. Every b-bit quantizer of the sphere has worst-case distortion at least 4^{-b} (Zandieh et al., 2026), and $D^{\mathrm{Haar}}_{\mathtt{mse}}(b)$ does not depend on the input, so $D^{\mathrm{Haar}}_{\mathtt{mse}}(b) \ge 4^{-b}$; divide Theorem 1 by it. □

In practice the relative gap is the number that matters. Corollary 2 says structured rotations are free once $d \gg 4^b$, and the reason is geometric: the lattice of S_u, with spacing $2/\sqrt{d}$, must be finer than the codebook cells, whose width is of order 2^{-b}. Exact gaps confirm this scaling with constant 0.6 instead of $56B_b$ (Fig. 1c): at $d = 64$ two-RHT is within 3% of dense Haar at $b = 1$ but 41% above it at $b = 4$ (Table 5), and within 1.5% at every $b \le 4$ once $d \ge 1024$.

Proposition 2 (Tightness at $b = 1$)**.** *For a basis vector $\boldsymbol{x} = \boldsymbol{e}_j$ the mixed weights are exactly flat, $Q_d(\boldsymbol{e}_j) = \mathrm{law}(d^{-1/2}\sum_{k \le d} \eta_k)$, and $\mathrm{gap}(1, \boldsymbol{e}_j, d) = (2/\pi + o(1))/d$. Hence $\sup_{\boldsymbol{x}} |\mathrm{gap}(1, \boldsymbol{x}, d)| = \Theta(1/d)$.*

Corollary 3 (When one stage suffices)**.** *For the one-RHT $\boldsymbol{HD}$, each scaled coordinate has the law of $S_{\boldsymbol{x}}$, so $|\,\mathbb{E}_{1\mathrm{RHT}}[L_{\boldsymbol{C}}] - \mathbb{E}_{P_d}[L_{\boldsymbol{C}}]| \le 2BC_K(s_4(\boldsymbol{x}) + \frac{3}{d+2})$: one stage is $O(1/d)$-accurate when $s_4(\boldsymbol{x}) = O(1/d)$, and the second stage replaces $s_4(\boldsymbol{x})$ by $\mathbb{E}_\delta\,\|u\|_4^4 \le 3/d$.*

Proposition 3 (One undithered stage fails on sparse inputs)**.** *Let $\boldsymbol{x}$ be equal-magnitude k-sparse and $R_k := k^{-1/2}\sum_{l \le k} \eta_l$. For every symmetric codebook, the one-RHT gap equals $\Gamma_k(\boldsymbol{C}) := \mathbb{E}\,L_{\boldsymbol{C}}(R_k) - \mathbb{E}\,L_{\boldsymbol{C}}(G)$ up to $6C_K B/(d+2)$. For the Gaussian Lloyd–Max codebooks, to which those of P_d converge (since $P_d \to N(0,1)$ and the Lloyd–Max quantizer of a log-concave law is unique (Kieffer, 1983)), exact evaluation gives $\Gamma_4 = +0.0764, +0.1196, +0.0206$ at $b = 1, 2, 3$, i.e. $+21\%, +102\%, +60\%$ of $D^{\mathrm{Haar}}_{\mathtt{mse}}(b)$. At $b = 4$, $|\Gamma_4| \approx 0.001$ is below the accuracy to which we resolve it at finite d.*

The sign of Γ_k varies with (b, k): a lattice law can fit a codebook better or worse than the Gaussian, but it does not match it. Tightness of the $1/d$ rate is proved only at $b = 1$; we conjecture it holds at every bit-width.

Conjecture 1. *For every fixed $b \ge 1$ there is $c_b > 0$ with $\sup_{\boldsymbol{x}} |\mathrm{gap}(b, \boldsymbol{x}, d)| \ge c_b/d$ for all large d.*

4 Experiments

Setup. Unless stated otherwise every arm applies the *same* data-independent symmetric Lloyd–Max codebook of P_d to the scaled rotated coordinates, so each comparison isolates the rotation; this is the quantizer of Section 2. Rotations are dense Haar (TurboQuant), one-RHT $\boldsymbol{HD}$, and two-RHT $\boldsymbol{HD}_2\boldsymbol{HD}_3$ (ours). The dithered single stage of Feng et al. (2026) (their Alg. 1: $\boldsymbol{HD}$, then a companding quantizer on the quantile scale of $N(0,3)$ with a uniformly shifted grid, regenerated from a seed) is the dithered baseline. Dense-Haar distortions are computed by quadrature on the density of P_d. All experiments run on one 14-core CPU; the GPU timings of Section C.1 are from a Tesla P100. Scripts and run files are listed in Section D.

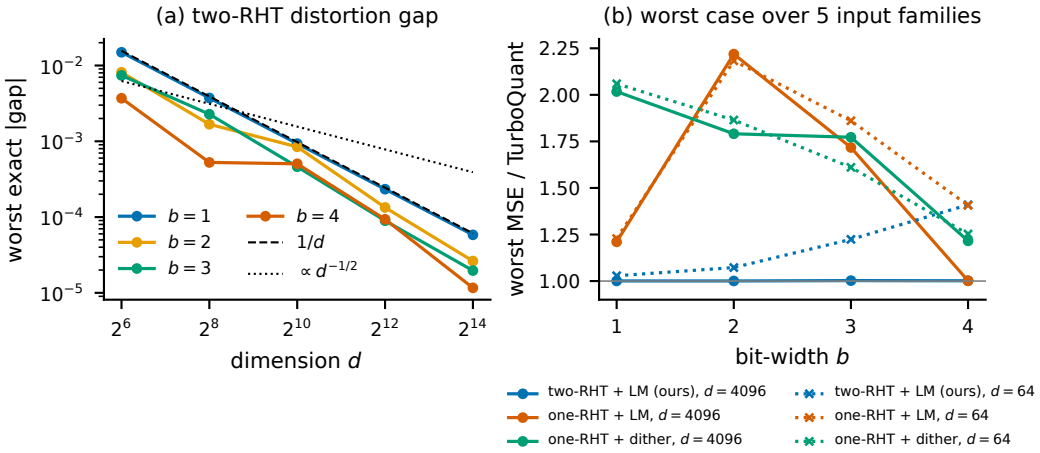


Figure 2: (a) Largest exact two-RHT gap over basis and equal-magnitude 2-, 4-, 8-sparse inputs (no Monte Carlo). Every curve follows the $1/d$ slope of Theorem 1; the $d^{-1/2}$ slope of distributional bounds is shown for reference. (b) Worst MSE over five input families relative to TurboQuant. Two undithered stages (blue) match TurboQuant at $d = 4096$; the dithered single stage (green) pays the cost of its companding quantizer at low b and wins only at small d with $b = 4$.

4.1 Exact distortion gaps

Monte-Carlo estimates of an $O(1/d)$ gap need far more than d^2 samples, so we compute the gap *exactly* on basis and sparse inputs. These produce lattice-supported mixed weights, the case that defeats distributional arguments; for delocalized inputs the fourth cumulants of $Q_d(\boldsymbol{x})$ and P_d agree up to $O(d^{-2})$ (Proposition 1), and the Monte-Carlo families of Section 4.2 show smaller gaps. For an equal-magnitude k-sparse input, only the 2^k sign patterns of $\boldsymbol{D}_3$ on the support matter, and for each pattern $\sqrt{kd}\, S_{u(\delta)}$ is an integer-valued sum of scaled Rademacher walks whose law follows by convolving binomials. Fig. 2(a) and Table 3 (appendix) report the result for $b \le 4$, $k \in \{1, 2, 4, 8\}$ and $d \le 16384$.

Every entry satisfies $|\mathrm{gap}| \cdot d \le 0.96$, about two orders of magnitude below the proven constant $12 C_K B_b \approx 44$–150. At $b = 1$, $\mathrm{gap} \cdot d$ is constant in d for every k; for basis inputs it equals 0.6366 at $d = 16384$, the value $2/\pi$ of Proposition 2. At $b \ge 2$ it oscillates with d as the lattice of S_u moves relative to the codebook knots, and it can change sign. For $b = 2, 3$ the basis-input values stay in $[0.21, 0.52]$, supporting Conjecture 1; at $b = 4$ they are too small to settle it. On the same sparse inputs one undithered stage has a d-independent gap equal to Γ_4 (Table 4), as Proposition 3 predicts.

4.2 Two undithered stages versus one dithered stage

Fig. 2(b) and Table 5 compare the worst MSE over five fixed-input families (basis, equal- and Gaussian-magnitude 4-sparse, flat sign vector, random) for $b \le 4$ and $d \in \{64, \ldots, 4096\}$, with $\ge 4 \times 10^6$ Monte-Carlo coordinates per cell. Two-RHT is within 1.5% of TurboQuant on every family at $d \ge 1024$ and within 0.3% at $d = 4096$. The dithered quantizer costs 2.03, 1.80, 1.42 and 1.22 times the TurboQuant MSE at $b = 1, \ldots, 4$ *even behind a dense rotation*, so its penalty comes from the companding quantizer, whose constant $\pi\sqrt{3}/2$ is reached only as $b \to \infty$, not from the single transform. So most of the 1.2–2× advantage belongs to the codebook, not to the second stage. Dithering does make one stage robust: its worst case stays within about $1.25\times$ of its own dense-rotation level, whereas undithered one-RHT reaches $2.2\times$. (One-stage laws depend only on the input's magnitudes, and the Gaussian-magnitude family draws fresh magnitudes at each d, so one-stage entries on that family are not monotone in d.) It wins only at small d with $b = 4$ (1.25 vs. 1.41 at $d = 64$), where the $1/d$ term is still large.

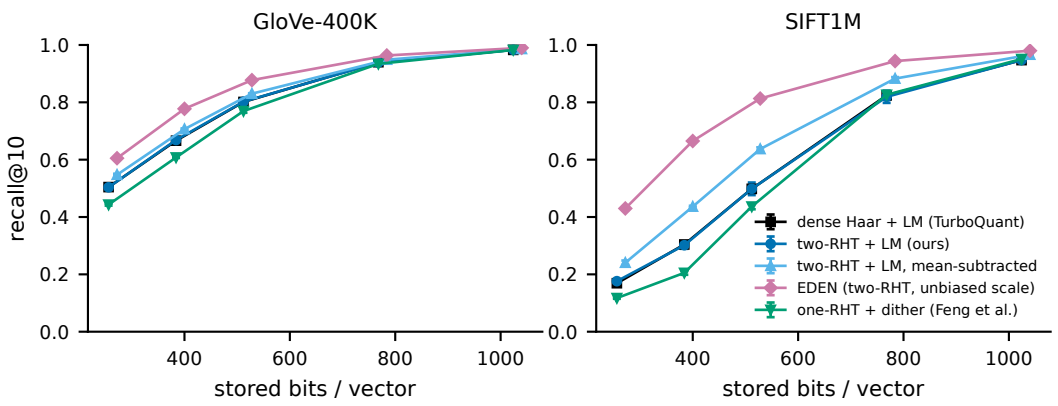


Figure 3: Recall@10 of an exact inner-product scan over the reconstructed database versus stored bits per vector (mean $\pm$ std over 5 rotation seeds). Lloyd–Max arms share one data-independent codebook; mean-subtracted and EDEN arms store 16 extra bits per vector.

4.3 Million-scale retrieval

We quantize GloVe 6B-100d (399K base vectors, 1K held-out queries, zero-padded to $d = 128$) and SIFT1M (10^6 base vectors, 1K queries, $d = 128$), all L2-normalized, at $b \in \{2, 3, 4, 6, 8\}$ bits per coordinate, and retrieve by an exact inner-product scan over the reconstructions (Fig. 3, Table 7). We add two variants that store 16 extra bits per vector: mean subtraction (the database mean is removed and the residual's norm is stored) and EDEN (Vargaftik et al., 2022), which rescales each reconstruction by the unbiased factor $\|y\|^2/\langle y, \hat{y}\rangle$.

Two-RHT matches the dense rotation. Across 5 seeds, two-RHT recall is within 0.003 of dense Haar on GloVe and within 0.007 on SIFT1M at every rate, and its database MSE is within 2%. This also holds after mean subtraction. Undithered one-RHT is equally close: normalized GloVe and SIFT vectors are ℓ_4-delocalized (median $d\, s_4(\boldsymbol{x}) = 3.7$ and 5.0, against 3 for a Gaussian vector and $d = 128$ for a basis vector), and Corollary 3 then predicts that one stage suffices. The second stage matters for sparse or spiky inputs (Proposition 3, Table 5), where one stage is up to $2.2\times$ worse; its price is one more fast transform. Real LLM gradients make the same point (Qwen2.5-0.5B, blocks of $d = 1024$, 5 seeds): with all entries kept both depths are within 3% of dense Haar, but after top-0.1% sparsification, where most blocks have one nonzero entry, one-RHT ranges from $0.44\times$ to $1.52\times$ the dense-Haar MSE over $b \le 4$, while two-RHT stays within 4%. The dithered single stage loses 6 (GloVe) and 5–10 (SIFT1M) recall points at $b \le 3$, with 1.4–$1.8\times$ the MSE, as Section 4.2 predicts, and is within one point at $b \ge 6$.

What matters more than the rotation. SIFT descriptors share a strong mean direction ($\|\bar{x}\| = 0.65$), and subtracting it adds 7–13 recall points at $b \le 4$ for 16 extra bits. EDEN's per-vector scale helps even more (0.81 vs. 0.50 on SIFT1M at $b = 4$) at nearly the same MSE. Lloyd–Max reconstructions shrink each vector by a factor that varies from vector to vector, and this biases inner-product rankings even when the average MSE is optimal. Our analysis concerns the MSE, so it is silent on this effect. At $d = 128$ a dense rotation is cheap, so these experiments test distortion transfer rather than speed.

4.4 LLM KV-cache quantization

Both models we test (Qwen2.5-0.5B-Instruct and TinyLlama-1.1B-Chat, fp32 on CPU) have head dimension 64, where a dense rotation costs almost nothing. So the only question here is whether distortion transfer survives attention. Each cached key and value vector (one per token and KV head) is quantized once on write. Its norm is stored in fp16, and the unit vector is rotated and coded with the Lloyd–Max codebook of P_{64} ($b + 0.25$ bits per coordinate). The *centered* variants first subtract a per-channel mean computed on held-out text (WikiText-2 train). The baselines are group-32 asymmetric integer codes ($b + 1$ bits per coordinate), either per token or KIVI-style with keys grouped per channel over 32 tokens (Liu et al., 2024b). We report WikiText-2 test perplexity

Table 2: KV cache at $b = 4$: needle retrieval (100 paired prompts of $\approx$ 1K tokens; Wilson 95% intervals) and greedy decode throughput (batch 1, 512-token prompt, this CPU; median of 3). Our cache stores dequantized values, so throughput measures quantization overhead, not memory savings. Perplexities for all arms and bit-widths are in Table 8.

	Qwen2.5-0.5B		TinyLlama-1.1B	
method ($b = 4$)	needle acc. [95% CI]	tok/s	needle acc. [95% CI]	tok/s
fp32 cache	100/100 [0.96, 1.00]	12.0	100/100 [0.96, 1.00]	6.6
int, per token (no rotation)	36/100 [0.27, 0.46]	11.3	99/100 [0.95, 1.00]	6.1
int, KIVI-style (keys per channel)	91/100 [0.84, 0.95]	–	–	–
dense Haar + LM (TurboQuant)	–	10.7	99/100 [0.95, 1.00]	6.2
two-RHT + LM (ours)	–	11.0	100/100 [0.96, 1.00]	6.2
one-RHT + LM	–	10.2	–	6.1
centered + dense Haar + LM	43/100 [0.34, 0.53]	–	–	–
centered + two-RHT + LM (ours)	70/100 [0.60, 0.78]	–	–	–

(16×1024 tokens, Table 8), needle retrieval (100 prompts of $\approx$ 1K tokens, the same prompts for every arm), and greedy decode throughput (Table 2).

The harness is sound, and Qwen's keys are the problem. KIVI-style integer codes are close to lossless at $b = 4$ (5 bits per coordinate with scales) on both models (perplexity 15.02 vs. 14.74 and 8.49 vs. 8.44; Table 8), matching the literature. On Qwen2.5, every quantizer that treats each key vector as a unit collapses: per-token integer codes reach 49.3, and rotated Lloyd–Max 157–255. Quantizing only the values is lossless for every method (Table 9). Qwen2.5 keys carry large per-channel offsets that a per-vector norm spreads over all coordinates. Subtracting a global per-channel mean before the rotation recovers most of the loss at $b = 4$ (25.2–25.7) but not the KIVI level, because rotary position embeddings make post-RoPE offsets position-dependent. On TinyLlama at $b = 2$ centering hurts.

Two-RHT tracks dense Haar inside attention. On TinyLlama the two rotations differ by under 1% in perplexity at $b = 3, 4$. At $b = 2$, over 3 rotation seeds, two-RHT gives 19.4 ± 1.2 and dense Haar 18.2 ± 0.3. Both reach 99–100/100 needle accuracy (exact McNemar $p = 1.0$). One-RHT is better than either at $b = 2$, on every seed (16.0 ± 0.5). We cannot explain this; head dimension 64 sits below the $d \gg 4^b$ regime of Corollary 2, so the theory does not predict it either way. On Qwen with centering, their perplexities agree (25.7 vs. 25.2). Their needle accuracies differ (70 vs. 43; McNemar $p < 10^{-3}$ over prompts), but each arm uses a single rotation draw, so we do not attribute this difference to the rotation type. Decode throughput is the same for both (11.0 vs. 10.7 tokens/s on Qwen, 6.2 vs. 6.2 on TinyLlama), 5–15% below an unquantized cache.

Scope. In these models, per-channel key statistics matter more than the rotation. KIVI at $b + 1$ bits beats rotated Lloyd–Max at $b + 0.25$ bits on both models at every b; at nearly equal rate (4.00 vs. 4.25 bits) the two are tied on TinyLlama (8.66 vs. 8.60) and KIVI still wins on Qwen2.5. The KV results support distortion transfer. They are not an argument for rotation-based scalar quantizers in KV caches.

5 Related work

Structured rotations for quantization. Randomized Hadamard transforms go back to the fast Johnson–Lindenstrauss transform (Ailon & Chazelle, 2006; Tropp, 2011); products of Hadamard and sign matrices approximate Haar rotations in kernel approximation (Yu et al., 2016; Choromanski et al., 2017), and a Hadamard rotation before scalar quantization was used for distributed mean estimation by Suresh et al. (2017). They also suppress outliers in LLM weight and activation quantization (Ashkboos et al., 2024; Tseng et al., 2024), and SpinQuant learns rotations (Liu et al., 2024a). Zilca & Mendelson (2026) prove Kolmogorov convergence of the two-block transform at rate $O(d^{-1/5})$, and Ben-Basat et al. (2026b) sharpen this to $O(d^{-1/2})$ in Kolmogorov and W_1 distance, apply it to DRIVE and QUIC-FL, and show that three RHTs decorrelate coordinate blocks for vector quantization. Feng et al. (2026) and Lin et al. (2026) analyze dithered single-stage constructions. Our work is the first to control the distortion functional of the undithered quantizer at

rate $1/d$. Fast-TurboQuant (Pereira et al., 2026) deploys a Rademacher–FWHT projection, and Ben-Basat et al. (2026a), Gao et al. (2026) and Ann et al. (2026) relate TurboQuant to EDEN, RaBitQ (Gao & Long, 2024) and block-spherical codes.

Quantization theory and KV caches. Lloyd–Max quantization (Lloyd, 1982; Max, 1960) and its high-rate asymptotics (Panter & Dite, 1951; Zador, 1964; Gersho, 1979) underlie the codebooks we use; our kink lemma builds on Stein's method with zero-bias couplings (Goldstein & Reinert, 1997; Chen et al., 2011). KV-cache quantizers such as KIVI (Liu et al., 2024b), KVQuant (Hooper et al., 2024) and RotateKV (Su et al., 2025) handle key outliers with per-channel statistics, pre-RoPE quantization, or rotations.

6 Limitations and conclusion

Limitations. (1) The constant of Theorem 1 is about $100\times$ the largest computed $|\mathrm{gap}| \cdot d$, and tightness is proved only at $b = 1$. The relative gap scales as $4^b/d$ (Corollary 2), so at high bit-width and small d, such as $b = 4$ and $d = 64$, structured rotations are measurably worse than dense Haar. (2) All guarantees are in expectation over the signs; in deployment one realization is shared by all vectors, and we do not prove high-probability bounds for a fixed realization. (3) The KV-cache study uses two small models with head dimension 64, where rotation cost is irrelevant. Per-channel key handling (KIVI) beats every rotation-based scalar quantizer there, and our analysis says nothing about anisotropic inputs such as Qwen2.5 keys (Section 4.4). (4) The GPU benchmark covers one P100, where the FWHT beats dense GEMM only for $d \gtrsim 2048$ (Section C.1). (5) Learned rotations are not evaluated.

Conclusion. For online scalar quantization the relevant question is not how fast structured-rotation coordinates approach Gaussianity but how fast the distortion approaches its dense-rotation value. For the two-RHT the answer is $O(B_b/d)$ for every bit-width and every input, attained at $b = 1$, because the loss sees the coordinate law only through kinks, on which Rademacher sums are $\|u\|_4^4$-close to Gaussian, and the first stage makes $\mathbb{E}\,\|u\|_4^4 \le 3/d$. Integrating Lemma 2 over t extends the bound to every test function whose second derivative is a finite signed measure with finite first moment. Losses outside this class, such as top-k selection, and vector quantization with block dependence, are natural next targets.

AI use statement

We used a large-language-model coding and writing assistant extensively in this work. It proposed and wrote the proofs of Lemma 2, Theorem 1, Corollary 1 and Proposition 3; wrote, debugged and ran all experiment and analysis code; drafted and edited the text; and checked every bibliographic entry against arXiv and publisher records. Every numerical ingredient of the proofs and every reported number is produced by a released script (Section D); the kink lemma and the exact gaps were additionally checked by exact numerical evaluation. We take full responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

Reproducibility statement

All theoretical claims are proved in Section A; the only numerical inputs to proofs are the constants Γ_k of Proposition 3, evaluated by `project/one_rht_limit.py`. Every table and figure is produced by a released script listed in Section D, together with its run file; `project/make_tables.py` and `project/make_figures.py` regenerate all tables and figures from the run files. Data sources, preprocessing, seeds and bit accounting are described in Section 4 and Section C.

References

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 557–563, 2006.

Heesang Ann, Joongkyu Lee, and Min hwan Oh. Block-sphere vector quantization. *arXiv preprint arXiv:2605.19972*, 2026.

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. *arXiv preprint arXiv:2404.00456*, 2024.

Ran Ben-Basat, Yaniv Ben-Itzhak, Gal Mendelson, Michael Mitzenmacher, Amit Portnoy, and Shay Vargaftik. A note on turboquant and the earlier drive/eden line of work. *arXiv preprint arXiv:2604.18555*, 2026a.

Ran Ben-Basat, William Kuszmaul, Michael Mitzenmacher, Amit Portnoy, and Shay Vargaftik. Quantizing with randomized Hadamard transforms: Efficient heuristic now proven. *arXiv preprint arXiv:2605.06014*, 2026b.

Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Probability and Its Applications. Springer, 2011. doi: 10.1007/978-3-642-15007-4.

Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ying Feng, Piotr Indyk, Michael Kapralov, Dmitry Krachun, and Boris Prokhorov. Provable quantization with randomized hadamard transform. *arXiv preprint arXiv:2605.13810*, 2026.

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.

Jianyang Gao, Yutong Gou, Yuexuan Xu, Jifan Shi, Yongyi Yang, Shuolin Li, Raymond Chi-Wing Wong, and Cheng Long. Revisiting rabitq and turboquant: A symmetric comparison of methods, theory, and experiments. *arXiv preprint arXiv:2604.19528*, 2026.

Allen Gersho. Asymptotically optimal block quantization. *IEEE Transactions on information theory*, 25(4):373–380, 1979.

Larry Goldstein. Bounds on the constant in the mean central limit theorem. *The Annals of Probability*, 38(4):1672–1689, 2010.

Larry Goldstein and Gesine Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4):935–952, 1997.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*, 2024.

John C. Kieffer. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *IEEE Transactions on Information Theory*, 29(1):42–47, 1983. doi: 10.1109/TIT.1983.1056622.

Honghao Lin, Vahab Mirrokni, and David P. Woodruff. Pairwise-independent dithering for single-stage hadamard quantization. *arXiv preprint arXiv:2608.02564*, 2026.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, et al. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024a.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024b.

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

Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960.

PF Panter and Wu Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, 1951.

Pedro M. R. Pereira, Felipe A. P. de Figueiredo, and Rausley A. A. de Souza. Fast-turboquant: A multiplier-free online vector quantization approach. *arXiv preprint arXiv:2606.21448*, 2026.

Irina G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010. doi: 10.1134/S1064562410060062.

Zunhai Su, Zhe Chen, Wang Shen, Hanyu Wei, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations, 2025. URL https://arxiv.org/abs/2501.16383.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *International Conference on Machine Learning*, pp. 3329–3337, 2017.

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

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*, 2024.

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

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*, 2022.

Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. *Advances in neural information processing systems*, 29, 2016.

Paul Laszlo Zador. *Development and evaluation of procedures for quantizing multivariate distributions*. Stanford University, 1964.

Amir Zandieh, Majid Daliri, Majid Hadian, Insu Han, Vahab Mirrokni, and Amin Karbasi. Turboquant: Random rotation for near-optimal online vector quantization. *ICLR*, 2026.

Tomer Zilca and Gal Mendelson. Approximating uniform random rotations by two-block structured hadamard rotations in high dimensions. *arXiv preprint arXiv:2604.23418*, 2026.

A PROOFS

A.1 PROOF OF LEMMA 1

Proof of Lemma 1. For each cell i of the Voronoi partition induced by $\boldsymbol{C}$, $L_{\boldsymbol{C}}(z) = (z - c_i)^2 = z^2 - 2c_i z + c_i^2$ on that cell; hence $g_{\boldsymbol{C}}(z) = \min_i(c_i^2 - 2c_i z)$ is a minimum of affine functions (so concave and continuous), even by symmetry of $\boldsymbol{C}$, piecewise linear with slopes $-2c_i \in [-2B, 2B]$, and $L_{\boldsymbol{C}}(z) = z^2 + g_{\boldsymbol{C}}(z)$ pointwise. Since both laws are centered with common second moment, $\mathbb{E}_Q[L_{\boldsymbol{C}}] - \mathbb{E}_P[L_{\boldsymbol{C}}] = \mathbb{E}_Q[g_{\boldsymbol{C}}] - \mathbb{E}_P[g_{\boldsymbol{C}}]$. By Kantorovich duality, $|\,\mathbb{E}_Q f - \mathbb{E}_P f| \le \mathrm{Lip}(f)\, W_1(P, Q)$ for every Lipschitz f, and $\mathrm{Lip}(g_{\boldsymbol{C}}) = 2B$. □

Proof of Theorem 2. By Lemma 1 and the triangle inequality,

$$\big|\, \mathop{\mathbb{E}}_{Q_d}[g] - \mathop{\mathbb{E}}_{P_d}[g]\big| \;\le\; 2B_b\big(W_1(Q_d(\boldsymbol{x}), N(0,1)) + W_1(P_d, N(0,1))\big).$$

Ben-Basat et al. (2026b) prove $W_1(Q_d(\boldsymbol{x}), N(0,1)) \le C_{\mathrm{BB}}/\sqrt{d}$ uniformly over $\boldsymbol{x} \in \mathbb{S}^{d-1}$. For the sphere side, $P_d = \mathrm{law}(S_\theta)$ with θ uniform on $\mathbb{S}^{d-1}$ (Lemma 4 below). Since $W_1(\cdot, N(0,1))$ is convex under mixtures and $W_1(S_u, N(0,1)) \le \|u\|_3^3$ for every unit u (Goldstein, 2010), we get $W_1(P_d, N(0,1)) \le \mathbb{E}\,\|\theta\|_3^3 = d\,\mathbb{E}\,|\theta_1|^3 \le d\,(\mathbb{E}\,\theta_1^4)^{3/4} = d\,\big(\frac{3}{d(d+2)}\big)^{3/4} \le 3^{3/4}/\sqrt{d}$. □

A.2 PROOF OF PROPOSITION 1

Proof of Proposition 1. Fix $\boldsymbol{x}$ and write $u_k = u_k(\delta) = \sum_\ell (\boldsymbol{H}_d)_{k\ell}\, \delta_\ell\, x_\ell$. For each k this is a Rademacher sum in δ with weights $a_\ell := (\boldsymbol{H}_d)_{k\ell} x_\ell$, where $\sum_\ell a_\ell^2 = 1/d$ and $\sum_\ell a_\ell^4 = s_4(\boldsymbol{x})/d^2$. A Rademacher sum with weights a has fourth moment $3(\sum a_\ell^2)^2 - 2\sum a_\ell^4$, hence

$$\mathop{\mathbb{E}}_{\delta}[u_k^4] = \frac{3 - 2s_4(\boldsymbol{x})}{d^2} \quad \text{for every } k, \qquad \mathop{\mathbb{E}}_{\delta}\|u(\delta)\|_4^4 = \frac{3 - 2s_4(\boldsymbol{x})}{d}.$$

For the cumulants: conditionally on δ, (1) gives $\mathbb{E}[Y^2 \mid \delta] = \|u\|_2^2 = 1$ and $\mathbb{E}[Y^4 \mid \delta] = 3 - 2\|u(\delta)\|_4^4$. Since the conditional variance is identically 1, the mixture satisfies $\mathbb{E}\,Y^2 = 1$ and $\kappa_4(Q_d(\boldsymbol{x})) = \mathbb{E}\,Y^4 - 3 = -2\,\mathbb{E}_\delta\,\|u(\delta)\|_4^4 = -2(3 - 2s_4(\boldsymbol{x}))/d$. For the sphere side, $(\sqrt{d}\,\theta_1)^2 \sim d \cdot \mathrm{Beta}(\frac{1}{2}, \frac{d-1}{2})$ gives $\mathbb{E}\,X^4 = 3d/(d+2)$ and $\kappa_4(P_d) = -6/(d+2)$. Finally $|\kappa_4(P_d) - \kappa_4(Q_d)| = \frac{|12 - 4s_4(\boldsymbol{x})(d+2)|}{d(d+2)} \le \frac{4}{d}$ using $1/d \le s_4 \le 1$, and the numerator is $O(1)$ when $s_4(\boldsymbol{x}) = O(1/d)$. □

A.3 PROOF OF THE KINK LEMMA (LEMMA 2)

Proof of Lemma 2. Fix $t \in \mathbb{R}$ and put $h(w) := |w - t|$, so h is 1-Lipschitz. Let f be the bounded solution of the Stein equation

$$f'(w) - w\, f(w) \;=\; h(w) - \mathbb{E}\,h(G). \tag{2}$$

Step 1 (regularity of f). By Chen et al. (2011, Lemma 2.4), $f \in C^1$ with $\|f\|_\infty \le 2$, $\|f'\|_\infty \le \sqrt{2/\pi}$, and f' is Lipschitz with $\|f''\|_\infty \le 2$. The right-hand side $wf(w) + h(w) - \mathbb{E}\,h(G)$ of (2) is locally Lipschitz with a.e. derivative $f + wf' + \mathrm{sgn}(\cdot - t)$, so

$$\psi(w) \;:=\; f(w) + w\, f'(w) + \mathrm{sgn}(w - t)$$

is a version of f''. Since $f + wf'$ is locally Lipschitz with a.e. derivative $2f' + wf''$, ψ has locally bounded variation and

$$d\psi \;=\; \rho(w)\, dw + 2\,\delta_t, \qquad \rho := 2f' + wf'', \qquad |\rho(w)| \le 2\sqrt{2/\pi} + 2|w|. \tag{3}$$

Step 2 (zero-bias coupling). Write $S = S_u = \sum_k \xi_k$ with $\xi_k := u_k \eta_k$: independent, mean zero, $\mathrm{Var}\,\xi_k = u_k^2$, $\sum_k u_k^2 = 1$. The zero-bias law of ξ_k (Goldstein & Reinert, 1997) is uniform on $[-|u_k|, |u_k|]$: for $U \sim \mathrm{Unif}[-1, 1]$ and every absolutely continuous φ, $\mathbb{E}[\xi_k \varphi(\xi_k)] = \frac{|u_k|}{2}\big(\varphi(|u_k|) - \varphi(-|u_k|)\big) = u_k^2\,\mathbb{E}\,\varphi'(|u_k|U)$. Let I be an index with $\Pr(I = k) = u_k^2$, and let I and U be

independent of each other and of η. Then $S^* := S - \xi_I + |u_I|U$ has the S-zero-biased law, $\mathbb{E}[Sf(S)] = \mathbb{E}\, f'(S^*)$ (Goldstein & Reinert, 1997); see also (Chen et al., 2011, Ch. 2). With $V_k := S - \xi_k$, independent of (η_k, U), (2) gives

$$\mathbb{E}\, h(S) - \mathbb{E}\, h(G) \;=\; \mathbb{E}\,\big[f'(S) - f'(S^*)\big] \;=\; \sum_k u_k^2\; \mathbb{E}\,\big[f'(V_k + \xi_k) - f'(V_k + |u_k|U)\big]. \quad (4)$$

Step 3 (trapezoid versus mean). Fix $v \in \mathbb{R}$, $a > 0$, and put $F(y) := f'(v + y)$, so F is C^1 except at $t - v$ and $F' = \psi(v + \cdot)$ has bounded variation on $[-a, a]$. Integrating by parts twice (the weight $(y^2 - a^2)/2$ vanishes at $\pm a$),

$$\begin{aligned} \tfrac{1}{2}\big(F(a) + F(-a)\big) - \tfrac{1}{2a}\int_{-a}^{a} F(y)\,dy \;&=\; \tfrac{1}{2a}\int_{-a}^{a} yF'(y)\,dy \\ &=\; \int_{[-a,a]} K_a(y)\,d\psi(v + y), \qquad K_a(y) := \frac{a^2 - y^2}{4a}. \end{aligned}$$

The kernel satisfies $K_a \geq 0$, $\sup K_a = a/4$, $\int_{-a}^{a} K_a = a^2/3$. With $a = |u_k|$, the k-th expectation in (4) is $\mathbb{E}\; T_k(V_k)$, where $T_k(v)$ denotes the left-hand side above, because $\xi_k = \pm a$ with probability $\frac{1}{2}$ each. By (3),

$$\begin{aligned} |T_k(v)| \;&\leq\; \frac{a^2}{3}\Big(2\sqrt{2/\pi} + 2|v| + 2a\Big) + 2K_a(t - v)\;\mathbb{1}\{|v - t| \leq a\} \\ &\leq\; \frac{a^2}{3}\Big(2\sqrt{2/\pi} + 2|v| + 2a\Big) + \frac{a}{2}\;\mathbb{1}\{|v - t| \leq a\}. \end{aligned}$$

Since $\mathbb{E}\,|V_k| \leq (\mathbb{E}\, V_k^2)^{1/2} \leq 1$ and $a \leq 1$, and $(2\sqrt{2/\pi} + 4)/3 < 2$,

$$\big|\,\mathbb{E}\; T_k(V_k)\big| \;\leq\; 2a^2 + \frac{a}{2}\;\Pr\big(|V_k - t| \leq a\big).$$

Step 4 (anti-concentration at the kink). If $a^2 \geq \frac{1}{2}$, bound the probability by 1: $\frac{a}{2} \leq a^2/\sqrt{2} \leq a^2$. If $a^2 < \frac{1}{2}$, V_k has variance $s^2 = 1 - a^2 > \frac{1}{2}$, and the Berry–Esseen theorem for independent non-identical summands (Shevtsova, 2010) gives $\sup_x |\Pr(V_k \leq x) - \Phi(x/s)| \leq C_{\mathrm{BE}}\|u\|_3^3/s^3 \leq 2^{3/2}C_{\mathrm{BE}}\|u\|_3^3$ (the same bound holds for left limits). Hence

$$\Pr\big(|V_k - t| \leq a\big) \;\leq\; \frac{2a}{s\sqrt{2\pi}} + 2^{5/2}C_{\mathrm{BE}}\|u\|_3^3 \;\leq\; \frac{2a}{\sqrt{\pi}} + 2^{5/2}C_{\mathrm{BE}}\|u\|_3^3.$$

In both cases $|\,\mathbb{E}\; T_k(V_k)| \leq 3a^2 + 2\sqrt{2}\,C_{\mathrm{BE}}\,a\,\|u\|_3^3$.

Step 5 (summation). Inserting into (4) with $a = |u_k|$,

$$\begin{aligned} \big|\,\mathbb{E}\, h(S) - \mathbb{E}\, h(G)\big| \;&\leq\; \sum_k u_k^2\big(3u_k^2 + 2\sqrt{2}\,C_{\mathrm{BE}}|u_k|\,\|u\|_3^3\big) \\ &=\; 3\|u\|_4^4 + 2\sqrt{2}\,C_{\mathrm{BE}}\,\|u\|_3^6 \;\leq\; \big(3 + 2\sqrt{2}\,C_{\mathrm{BE}}\big)\|u\|_4^4, \end{aligned}$$

where the last step is Cauchy–Schwarz: $\|u\|_3^6 = \big(\sum_k |u_k| \cdot u_k^2\big)^2 \leq \big(\sum_k u_k^2\big)\big(\sum_k u_k^4\big) = \|u\|_4^4$. With $C_{\mathrm{BE}} \leq 0.56$, $C_K \leq 4.59$. □

A.4 Proof of the sharp transfer theorem (Theorem 1)

Lemma 3 (Kink decomposition)**.** *Let $\boldsymbol{C} = \{c_1 < \cdots < c_n\}$ be symmetric about 0, $t_i := (c_i + c_{i+1})/2$ and $w_i := c_{i+1} - c_i > 0$ for $i < n$. Then for all $z \in \mathbb{R}$,*

$$g_{\boldsymbol{C}}(z) \;=\; g_{\boldsymbol{C}}(0) - \sum_{i<n} w_i\big(|z - t_i| - |t_i|\big), \qquad \sum_{i<n} w_i = c_n - c_1 = 2B.$$

Proof. Both sides are continuous and piecewise linear with kinks only at the t_i, and they agree at $z = 0$. On the Voronoi cell (t_{m-1}, t_m) of c_m (with $t_0 = -\infty$, $t_n = +\infty$), the right-hand side has slope $-\sum_{i<m} w_i + \sum_{i\geq m} w_i = -(c_m - c_1) + (c_n - c_m) = -2c_m$, using $c_1 = -c_n$. This is the slope of $g_{\boldsymbol{C}}$ on that cell. □

Lemma 4 (The sphere coordinate is a Rademacher mixture)**.** *Let θ be uniform on $\mathbb{S}^{d-1}$, independent of the Rademacher vector η. Then $S_\theta = \sum_k \eta_k \theta_k$ has law P_d, and $\mathbb{E}\,\|\theta\|_4^4 = 3/(d+2)$.*

Proof. Conditionally on η, $v := \eta/\sqrt{d}$ is a fixed unit vector, so by rotation invariance $\langle v, \theta\rangle \overset{d}{=} \theta_1$; hence $S_\theta = \sqrt{d}\,\langle v, \theta\rangle \overset{d}{=} \sqrt{d}\,\theta_1 \sim P_d$. Moreover $\mathbb{E}\,\|\theta\|_4^4 = d\,\mathbb{E}\,\theta_1^4 = d \cdot \frac{3}{d(d+2)}$. □

Proof of Theorem 1. For a symmetric law R write $\Delta_R(t) := \mathbb{E}_{Z\sim R}\,|Z-t| - \mathbb{E}\,|G-t|$. Both $Q_d(\boldsymbol{x})$ and P_d are centered with unit variance, so by Lemmas 1 and 3,

$$\mathop{\mathbb{E}}_{Q_d(\boldsymbol{x})}[L_{\boldsymbol{C}}] - \mathop{\mathbb{E}}_{P_d}[L_{\boldsymbol{C}}] \;=\; \mathop{\mathbb{E}}_{Q_d(\boldsymbol{x})}[g_{\boldsymbol{C}}] - \mathop{\mathbb{E}}_{P_d}[g_{\boldsymbol{C}}] \;=\; -\sum_{i<n} w_i\big(\Delta_{Q_d(\boldsymbol{x})}(t_i) - \Delta_{P_d}(t_i)\big),$$

and therefore $|\mathrm{gap}| \le 2B\,\sup_t\big(|\Delta_{Q_d(\boldsymbol{x})}(t)| + |\Delta_{P_d}(t)|\big)$. *Two-RHT side.* By (1), $\mathbb{E}_{Q_d(\boldsymbol{x})}\,|Y - t| = \mathbb{E}_\delta\,\mathbb{E}_\eta\,|S_{u(\delta)} - t|$. Applying Lemma 2 for each δ and then Proposition 1, $|\Delta_{Q_d(\boldsymbol{x})}(t)| \le C_K\,\mathbb{E}_\delta\,\|u(\delta)\|_4^4 = C_K(3-2s_4(\boldsymbol{x}))/d$. *Haar side.* By Lemma 4 and Lemma 2 applied conditionally on θ, $|\Delta_{P_d}(t)| \le C_K\,\mathbb{E}\,\|\theta\|_4^4 = 3C_K/(d+2)$. Combining the two sides gives the first inequality. The second follows from $s_4 \ge 0$ and $3/(d+2) \le 3/d$. For the Lloyd–Max codebook, $B = B_b$ and $12C_K < 56$. □

Proof of Corollary 1. Apply Theorem 1 with $\boldsymbol{C} = \{\pm m_P(d)\}$ and $B = m_P(d) = \mathbb{E}_{P_d}\,|X| \le (\mathbb{E}_{P_d}\,X^2)^{1/2} = 1$. □

Proof of Corollary 1. By Jensen, $\mathbb{E}\,\|\boldsymbol{\Pi}_H\boldsymbol{x}\|_1^2 \ge (\mathbb{E}\,\|\boldsymbol{\Pi}_H\boldsymbol{x}\|_1)^2 = d^2(\mathbb{E}\,|(\boldsymbol{\Pi}_H\boldsymbol{x})_1|)^2 = d\,m_Q(\boldsymbol{x})^2$, so vNMSE $\le 1 - m_Q(\boldsymbol{x})^2$; this is Ben-Basat et al. (2026b, Eq. (2)). By (1), Lemma 2 at $t = 0$ and Proposition 1, $m_Q(\boldsymbol{x}) = \mathbb{E}_\delta\,|S_{u(\delta)}| \ge \sqrt{2/\pi} - C_K\,\mathbb{E}_\delta\,\|u(\delta)\|_4^4 \ge \sqrt{2/\pi} - 3C_K/d$. Hence $m_Q^2 \ge \frac{2}{\pi} - 2\sqrt{2/\pi}\cdot 3C_K/d$ whenever the right-hand side of the previous bound is nonnegative, and the claim is trivial otherwise (vNMSE ≤ 1). Finally $6\sqrt{2/\pi}\,C_K \le 21.95$. □

Proof of Corollary 3. For $\boldsymbol{\Pi} = \boldsymbol{H}_d\boldsymbol{D}$, $\sqrt{d}\,(\boldsymbol{\Pi}\boldsymbol{x})_j = \sum_k \sqrt{d}\,(\boldsymbol{H}_d)_{jk}\boldsymbol{D}_{kk}x_k \overset{d}{=} S_{\boldsymbol{x}}$, because $\sqrt{d}(\boldsymbol{H}_d)_{jk}\boldsymbol{D}_{kk}$ are i.i.d. Rademacher. The proof of Theorem 1 then applies with $\mathbb{E}_\delta\,\|u(\delta)\|_4^4$ replaced by $\|\boldsymbol{x}\|_4^4 = s_4(\boldsymbol{x})$. □

A.5 Proof of Proposition 3

Proof of Proposition 3. Let $\boldsymbol{x} = k^{-1/2}\sum_{l\in S}\sigma_l\boldsymbol{e}_l$ with $|S| = k$. As in the proof of Corollary 3, every scaled one-RHT coordinate has the law of $S_{\boldsymbol{x}} = k^{-1/2}\sum_{l\in S}\sigma_l\eta_l \overset{d}{=} R_k$, which does not depend on d or S. Hence

$$\mathop{\mathbb{E}}_{1\mathrm{RHT}}[L_{\boldsymbol{C}}] - \mathop{\mathbb{E}}_{P_d}[L_{\boldsymbol{C}}] - \Gamma_k(\boldsymbol{C}) \;=\; \mathbb{E}\;L_{\boldsymbol{C}}(G) - \mathop{\mathbb{E}}_{P_d}[L_{\boldsymbol{C}}] \;=\; \mathbb{E}\;g_{\boldsymbol{C}}(G) - \mathop{\mathbb{E}}_{P_d}[g_{\boldsymbol{C}}],$$

using $\mathbb{E}\,G^2 = \mathbb{E}_{P_d}\,X^2 = 1$. By Lemma 3, this equals $-\sum_i w_i\Delta_{P_d}(t_i)$ with $\sum_i w_i = 2B$, and the Haar-side bound in the proof of Theorem 1 gives $|\Delta_{P_d}(t)| \le 3C_K/(d+2)$ for all t. The constants $\Gamma_k(\boldsymbol{C}_b^\infty)$ are a finite binomial sum minus a Gaussian integral; we evaluate them by exact summation and adaptive quadrature (`project/one_rht_limit.py`). For $k = 4$ they are $+0.07641, +0.11964, +0.02062, -0.00107$ at $b = 1,\dots,4$, and the same evaluation at the P_d-codebooks for $d \in \{64,\dots,16384\}$ agrees to within 0.005; this fixes the sign for $b \le 3$ but not at $b = 4$. Their values for other k also change sign with (b,k), e.g. $\Gamma_1(\boldsymbol{C}_1^\infty) = -0.3225$ and $\Gamma_2(\boldsymbol{C}_2^\infty) = -0.0103$. □

A.6 The lower bound at $b = 1$

At $b = 1$ the symmetric codebook is $\{-c, +c\}$ with $c = m_P(d) := \mathbb{E}_{P_d}\,|X|$, and

$$L(z) = (|z| - c)^2 = z^2 - 2c|z| + c^2 \qquad\Longrightarrow\qquad \mathop{\mathbb{E}}_{S}[L] = 1 - 2c\mathop{\mathbb{E}}_{S}|Z| + c^2,$$

so the gap is exactly $2c\,\big(m_P - m_Q(\boldsymbol{x})\big)$ where $m_Q(\boldsymbol{x}) := \mathbb{E}_{Q_d(\boldsymbol{x})}\,|Y|$.

Lemma 5 (Sphere side; exact). *$m_P(d) = \sqrt{2/\pi}\,\big(1+\frac{1}{4d}+O(d^{-2})\big)$, with the exact value $m_P(d) = \sqrt{d}\,\frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma(\frac{d+1}{2})}$, obtained from $\int_0^1 t(1-t^2)^{(d-3)/2}dt = (d-1)^{-1}$ and the asymptotic expansion of the Gamma ratio.*

Lemma 6 (Flat Rademacher side; exact). *For even d and $W_d = \sum_{k\le d}\eta_k$, $\mathbb{E}\,\frac{|W_d|}{\sqrt{d}} = \frac{\sqrt{d}}{2^d}\binom{d}{d/2} = \sqrt{\frac{2}{\pi}}\big(1-\frac{1}{4d}+O(d^{-2})\big)$, by Stirling's formula.*

Proof of Proposition 2. Let $\boldsymbol{x} = \boldsymbol{e}_j$. For every realization of $\boldsymbol{D}_3$, $\boldsymbol{D}_3\boldsymbol{x} = \pm\boldsymbol{e}_j$ and $u(\delta) = \boldsymbol{H}\boldsymbol{D}_3\boldsymbol{x} = \pm\boldsymbol{H}\boldsymbol{e}_j$ is a signed Walsh column, so $|u_k| = d^{-1/2}$ for all k and, by (1), $Q_d(\boldsymbol{x})$ is exactly the law of $d^{-1/2}W_d$. By Lemmas 5 and 6,

$$\mathrm{gap}(1,\boldsymbol{e}_j,d) = 2\,m_P\,\big(m_P - m_Q\big) = 2\cdot\frac{2}{\pi}\cdot\frac{1}{2d}\,\big(1+O(d^{-1})\big) = \frac{(2/\pi)+o(1)}{d},$$

with the exact finite-d value available in closed form; numerically $\mathrm{gap}\cdot d = 0.6366$ already at $d = 256$ (Section 4.1), matching $2/\pi = 0.63662$. □

Where the constant is lost. For flat weights and $t = 0$, Lemmas 6–5 give $|\Delta(0)|\,d \to \sqrt{2/\pi}/4 \approx 0.1995$ on the Rademacher side, which is the largest ratio $|\Delta(t)|/\|u\|_4^4$ found by exact evaluation over flat, spiked-plus-flat, and random weight vectors (worst 0.205; Section D). The kink-lemma constant $C_K \le 4.59$ is therefore loose by about $22\times$. The remaining slack in Theorem 1 comes from bounding the 2^b-1 kink terms by the triangle inequality, which ignores the cancellation between Δ_Q and Δ_P (both are negative-kurtosis perturbations of $N(0,1)$) and between neighbouring knots.

A.7 Proof of Theorem 3

Proof of Theorem 3. Unbiasedness follows exactly as in TurboQuant: the QJL residual sketch uses an independent random matrix and is unbiased for any fixed residual; the rotation choice does not enter. For the variance, the TurboQuant analysis gives $D^{(H)}_{\texttt{prod}} \le \frac{\pi}{2d}\,\|\boldsymbol{y}\|_2^2\; D^{\text{two-RHT}}_{\texttt{mse}}(b-1,\boldsymbol{x})$, and Theorem 1 at bit-width $b-1 \ge 1$ gives $D^{\text{two-RHT}}_{\texttt{mse}}(b-1,\boldsymbol{x}) \le D^{\text{Haar}}_{\texttt{mse}}(b-1) + \frac{12C_K B_{b-1}}{d} \le \frac{\sqrt{3}\pi}{2}4^{-(b-1)}\Big(1+\frac{24\,C_K B_{b-1}4^{b-1}}{\sqrt{3}\,\pi\,d}\Big)$. Multiplying through gives the claim with $C''_b = 24\,C_K B_{b-1}4^{b-1}/(\sqrt{3}\,\pi)$. □

B Additional statements

Theorem 2 (Generic transfer, a corollary of Ben-Basat et al., 2026b). *For every $b \ge 1$, unit $\boldsymbol{x}$ and d a power of 2, $|D^{\text{two-RHT}}_{\texttt{mse}}(b,\boldsymbol{x}) - D^{\text{Haar}}_{\texttt{mse}}(b)| \le 2B_b(C_{\mathrm{BB}} + 3^{3/4})/\sqrt{d}$, where C_{BB} is the uniform two-RHT Wasserstein constant of Ben-Basat et al. (2026b).*

Proposition 4 (Output-sign redundancy). *Let q be an odd scalar quantizer ($q(-z) = -q(z)$, e.g. any symmetric codebook). For any orthogonal $\mathbf{R}$ and sign matrix $\boldsymbol{D}$, $\mathbf{R}^\top\boldsymbol{D}\,q(\boldsymbol{D}\mathbf{R}\boldsymbol{x}) = \mathbf{R}^\top q(\mathbf{R}\boldsymbol{x})$, because $q(\boldsymbol{D}z) = \boldsymbol{D}q(z)$ and $\boldsymbol{D}^2 = I$. Hence $\boldsymbol{D}_1\boldsymbol{H}\boldsymbol{D}_2\boldsymbol{H}\boldsymbol{D}_3$ and $\boldsymbol{H}\boldsymbol{D}_2\boldsymbol{H}\boldsymbol{D}_3$ have the same distortion at every realization, and the two-RHT stores $2d$ sign bits.*

Theorem 3 (Inner-product estimator). *For $b \ge 2$, unit $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^d$, the TurboQuant$_{\texttt{prod}}$ estimator (a $(b-1)$-bit MSE stage followed by a 1-bit QJL sketch of the residual) with rotation $\boldsymbol{\Pi}_H$ is unbiased and satisfies $D^{(H)}_{\texttt{prod}} \le \frac{\sqrt{3}\pi^2\|\boldsymbol{y}\|_2^2}{d}4^{-b}\big(1+\frac{C''_b}{d}\big)$ with $C''_b = 24C_K B_{b-1}4^{b-1}/(\sqrt{3}\pi)$, by Theorem 1 at bit-width $b-1$.*

Remark 1 (Covariance). *For $i \ne j$, $\mathrm{cov}(y_i,y_j) = 0$ unconditionally under the two-RHT, but this is not independence: Ben-Basat et al. (2026b) exhibit sparse inputs whose conditional covariance is ±1 once the signs are drawn. Scalar quantization needs only the marginals, so our analysis never uses joint decorrelation.*

Non-power-of-two dimensions. For d not a power of 2, zero-pad $\boldsymbol{x}/\|\boldsymbol{x}\|$ to $d' = 2^{\lceil\log_2 d\rceil}$, apply the two-RHT in dimension d', and discard the padded coordinates on reconstruction; storage grows by $d'/d \le 2$, and Theorems 1 and 2 apply with d' in place of d. The rotation itself needs $2d'$ bits, compared with $32d^2$ bits for a float32 dense matrix (or a stored seed and $O(d^3)$ regeneration).

C ADDITIONAL EXPERIMENTAL RESULTS

Table 3: Exact two-RHT gap, reported as $\mathrm{gap}(b, \boldsymbol{x}, d) \cdot d$, for basis and equal-magnitude k-sparse inputs (`project/exact_gap.py`). Theorem 1 bounds every entry by $12 C_K B_b \approx 44$–150. At $b \geq 2$ and large d the entries depend on the knot positions relative to a lattice of spacing $2/\sqrt{d}$, so a 10^{-3} change in the numerically computed codebook can change their sign.

b	input	$d = 64$	$d = 256$	$d = 1024$	$d = 4096$	$d = 16384$
1	basis $\boldsymbol{e}_j$	$+0.639$	$+0.637$	$+0.637$	$+0.637$	$+0.637$
1	2-sparse	$+0.957$	$+0.955$	$+0.955$	$+0.955$	$+0.955$
1	4-sparse	$+0.472$	$+0.476$	$+0.477$	$+0.477$	$+0.477$
1	8-sparse	$+0.230$	$+0.237$	$+0.238$	$+0.239$	$+0.239$
2	basis $\boldsymbol{e}_j$	$+0.521$	$+0.429$	$+0.275$	$+0.211$	$+0.431$
2	2-sparse	$+0.431$	$+0.199$	$+0.865$	$+0.548$	$+0.218$
2	4-sparse	$+0.422$	$+0.333$	$+0.179$	$+0.115$	$+0.335$
2	8-sparse	$+0.042$	$+0.225$	$+0.141$	$+0.055$	$+0.189$
3	basis $\boldsymbol{e}_j$	$+0.474$	$+0.400$	$+0.472$	$+0.366$	$+0.322$
3	2-sparse	$+0.342$	$+0.583$	$+0.216$	$+0.010$	$+0.022$
3	4-sparse	$+0.433$	$+0.361$	$+0.434$	$+0.329$	$+0.285$
3	8-sparse	$+0.097$	$+0.034$	-0.002	$+0.006$	$+0.020$
4	basis $\boldsymbol{e}_j$	$+0.237$	$+0.101$	$+0.077$	$+0.066$	-0.101
4	2-sparse	$+0.233$	$+0.107$	$+0.517$	$+0.383$	$+0.191$
4	4-sparse	$+0.221$	$+0.088$	$+0.065$	$+0.054$	-0.113
4	8-sparse	$+0.061$	$+0.135$	$+0.080$	$+0.007$	$+0.003$

Table 4: Exact gaps on equal-magnitude 4-sparse inputs (total-MSE units). The one-RHT gap converges to Γ_4 (Proposition 3); the two-RHT gap decays as $1/d$.

	$d = 64$	$d = 256$	$d = 1024$	$d = 4096$	$d = 16384$
one-RHT, b=2	$+1.18\mathrm{e}{-1}$	$+1.19\mathrm{e}{-1}$	$+1.20\mathrm{e}{-1}$	$+1.20\mathrm{e}{-1}$	$+1.20\mathrm{e}{-1}$
two-RHT, b=2	$+6.59\mathrm{e}{-3}$	$+1.30\mathrm{e}{-3}$	$+1.75\mathrm{e}{-4}$	$+2.82\mathrm{e}{-5}$	$+2.05\mathrm{e}{-5}$
one-RHT, b=3	$+2.15\mathrm{e}{-2}$	$+2.08\mathrm{e}{-2}$	$+2.06\mathrm{e}{-2}$	$+2.06\mathrm{e}{-2}$	$+2.05\mathrm{e}{-2}$
two-RHT, b=3	$+6.76\mathrm{e}{-3}$	$+1.41\mathrm{e}{-3}$	$+4.24\mathrm{e}{-4}$	$+8.03\mathrm{e}{-5}$	$+1.74\mathrm{e}{-5}$

Table 5: Worst-case MSE over five fixed-input families relative to TurboQuant (dense Haar with Lloyd–Max, exact = 1); "dither" is Alg. 1 of Feng et al. (2026), whose dense-Haar row is exact (`project/dither_head_to_head.py`).

	$d = 64$				$d = 256$				$d = 1024$				$d = 4096$			
method	b=1	b=2	b=3	b=4	b=1	b=2	b=3	b=4	b=1	b=2	b=3	b=4	b=1	b=2	b=3	b=4
two-RHT + Lloyd–Max (ours)	1.029	1.073	1.224	1.409	1.006	1.014	1.045	1.041	1.002	1.004	1.015	1.009	1.001	1.001	1.002	1.002
one-RHT + Lloyd–Max	1.228	2.183	1.861	1.407	1.215	2.209	1.755	1.044	1.211	2.216	1.727	1.008	1.211	2.219	1.717	1.002
one-RHT + dither (Feng et al.)	2.059	1.865	1.611	1.253	2.037	1.854	1.617	1.224	2.043	1.813	1.572	1.222	2.018	1.791	1.772	1.216
dense Haar + dither (exact)	2.055	1.832	1.466	1.249	2.037	1.809	1.434	1.224	2.032	1.804	1.426	1.219	2.031	1.802	1.424	1.217

C.1 GPU ROTATION BENCHMARK

Table 10 times a Triton FWHT against cuBLAS dense GEMM on a Tesla P100 (`project/gpu_rotation_bench.py`). Launched naively, the 24-launch FWHT costs 1.2–1.5 ms regardless of size and beats GEMM only at $d = 4096$, $B = 1024$. With CUDA-graph capture it is slower at $d = 1024$, $B = 256$ ($0.80\times$), at parity at $d = 1024$, $B = 1024$ ($1.09\times$), and 1.6–$1.9\times$ and 2.9–$3.6\times$ faster at $d = 2048$ and 4096. The timed kernel applies one more sign layer than the two-RHT, so it slightly overstates the two-RHT cost. Tensor-core GPUs and a single-kernel implementation are untested; faster dense GEMMs would move the break-even point to larger d.

Table 6: Inner-product estimator of Theorem 3 on fixed equal-magnitude 4-sparse inputs ($d = 256$, 2000 inputs, 500 random queries, 10 rotation seeds; IP MSE in units of 10^{-4}; `project/qprod_experiment.py`).

b	mean bias		IP MSE (mean $\pm$ std over seeds)	
	dense Haar	two-RHT	dense Haar	two-RHT
2	$-6.9\text{e}-6$	$-2.1\text{e}-5$	22.213 ± 0.199	22.400 ± 0.297
3	$-3.3\text{e}-6$	$-2.5\text{e}-5$	7.188 ± 0.074	7.219 ± 0.142
4	$-4.3\text{e}-6$	$-1.3\text{e}-5$	2.093 ± 0.024	2.185 ± 0.078

Table 7: ANN recall@10 (mean $\pm$ std over 5 rotation seeds; `project/ann_benchmark.py`). Bits per vector are $128b$, plus 16 for the arms marked (+16). LM is the shared Lloyd–Max codebook of P_{128}.

data	method bits/vector	b=2 256	b=3 384	b=4 512	b=6 768	b=8 1024
GloVe-400K	dense Haar + LM (TurboQuant)	0.504 ± 0.003	0.666 ± 0.006	0.801 ± 0.005	0.940 ± 0.001	0.983 ± 0.001
	two-RHT + LM (ours)	0.504 ± 0.003	0.669 ± 0.006	0.802 ± 0.006	0.939 ± 0.003	0.983 ± 0.001
	one-RHT + LM	0.506 ± 0.004	0.663 ± 0.003	0.797 ± 0.005	0.937 ± 0.003	0.983 ± 0.002
	one-RHT + dither (Feng et al.)	0.443 ± 0.003	0.608 ± 0.002	0.769 ± 0.006	0.934 ± 0.003	0.983 ± 0.001
	mean-subtracted dense Haar + LM (+16)	0.546 ± 0.003	0.709 ± 0.002	0.832 ± 0.003	0.949 ± 0.001	0.986 ± 0.000
	mean-subtracted two-RHT + LM (+16)	0.547 ± 0.004	0.707 ± 0.003	0.830 ± 0.001	0.949 ± 0.002	0.986 ± 0.001
	EDEN, two-RHT, unbiased scale (+16)	0.605 ± 0.005	0.777 ± 0.004	0.877 ± 0.002	0.963 ± 0.001	0.990 ± 0.001
	fp16 (1600 bits)			0.999		
SIFT1M	dense Haar + LM (TurboQuant)	0.169 ± 0.002	0.304 ± 0.007	0.498 ± 0.011	0.824 ± 0.015	0.947 ± 0.004
	two-RHT + LM (ours)	0.176 ± 0.002	0.302 ± 0.012	0.498 ± 0.022	0.819 ± 0.021	0.947 ± 0.008
	one-RHT + LM	0.172 ± 0.005	0.308 ± 0.009	0.516 ± 0.011	0.836 ± 0.004	0.953 ± 0.003
	one-RHT + dither (Feng et al.)	0.117 ± 0.002	0.205 ± 0.007	0.436 ± 0.009	0.826 ± 0.010	0.951 ± 0.002
	mean-subtracted dense Haar + LM (+16)	0.242 ± 0.005	0.432 ± 0.004	0.631 ± 0.008	0.880 ± 0.006	0.965 ± 0.002
	mean-subtracted two-RHT + LM (+16)	0.240 ± 0.008	0.436 ± 0.003	0.637 ± 0.006	0.883 ± 0.005	0.966 ± 0.002
	EDEN, two-RHT, unbiased scale (+16)	0.430 ± 0.006	0.665 ± 0.006	0.813 ± 0.005	0.944 ± 0.003	0.980 ± 0.001
	fp16 (2048 bits)			0.993		

D REPRODUCIBILITY: SCRIPTS AND RUN FILES

All scripts are in `project/`, run on CPU (Python 3.14, NumPy 2.5, SciPy 1.18, PyTorch 2.13, Transformers 5.17) unless noted, fix their seeds, and write the listed run file to `project/runs/`. `make_tables.py` and `make_figures.py` regenerate every table and figure from these files.

Result	Script	Run file
Fig. 2(a), Tables 3 and 4	`exact_gap.py`	`exact_gap.json`
Proposition 3 (Γ_k)	`one_rht_limit.py`	`one_rht_limit.json`
Lemma 2 ratio check	`kink_lemma_check.py`	`kink_lemma_check.json`
Fig. 2(b), Table 5	`dither_head_to_head.py`	`dither_head_to_head.json`
Table 6	`qprod_experiment.py`	`qprod_d256.json`
Fig. 3, Table 7	`ann_benchmark.py`	`ann_benchmark.json`
Sparse gradients (Section 4.3)	`sparse_grad_study.py`	`sparse_grad_study.json`
Fig. 1	`make_figures.py`	`exact_gap.json`
Tables 2, 8 and 9, KV seeds	`run_kv_suite.py`	`kv_*.json`
Table 10 (GPU)	`gpu_rotation_bench.py`	`gpu_rotation_bench_p100.json`[†]

[†]Recovered from the printed output of the original P100 run; the raw run file of that run was not retained.

`ann_benchmark.py` reads GloVe 6B and SIFT1M (TEXMEX or the ann-benchmarks HDF5 mirror) from `project/data/`; `kv_benchmark.py` downloads the models and WikiText-2 from the Hugging Face Hub.

Table 8: KV-cache perplexity (WikiText-2 test, 16×1024 tokens; `project/kv_benchmark.py`). Bits per coordinate include side information. One rotation seed; see Section 4.4 for three seeds at $b = 2$.

method	bits/coord	Qwen2.5-0.5B	TinyLlama-1.1B
fp32 cache	32	14.74	8.44
int, per token (no rotation)	5.00	49.33	8.56
int, KIVI-style (keys per channel)	5.00	15.02	8.49
dense Haar + LM (TurboQuant)	4.25	255.13	8.66
two-RHT + LM (ours)	4.25	199.70	8.60
one-RHT + LM	4.25	157.32	8.59
centered + dense Haar + LM	4.25	25.20	8.62
centered + two-RHT + LM (ours)	4.25	25.69	8.70
int, per token (no rotation)	4.00	386.94	9.26
int, KIVI-style (keys per channel)	4.00	15.87	8.66
dense Haar + LM (TurboQuant)	3.25	396.26	9.31
two-RHT + LM (ours)	3.25	386.29	9.39
one-RHT + LM	3.25	429.29	9.21
centered + dense Haar + LM	3.25	74.15	9.58
centered + two-RHT + LM (ours)	3.25	62.30	9.72
int, per token (no rotation)	3.00	3446.69	44.81
int, KIVI-style (keys per channel)	3.00	31.87	11.25
dense Haar + LM (TurboQuant)	2.25	916.93	18.12
two-RHT + LM (ours)	2.25	1287.57	20.75
one-RHT + LM	2.25	881.73	16.53
centered + dense Haar + LM	2.25	303.77	49.48
centered + two-RHT + LM (ours)	2.25	322.17	52.96

Table 9: Qwen2.5-0.5B at $b = 4$: perplexity when only the keys, only the values, or both are quantized.

method ($b = 4$)	keys only	values only	both
group-32 int, no rotation	50.52	14.84	49.33
dense Haar + LM (TurboQuant)	245.27	14.87	255.13
two-RHT + LM (ours)	208.22	14.83	199.70
fp32 cache		14.74	

Table 10: GPU rotation latency (ms, fp16, median of CUDA-event timings) for $B \times d$ inputs on a Tesla P100.

d	B	GEMM	FWHT	FWHT (graph)	speedup (graph)
1024	256	0.115	1.155	0.143	0.80×
1024	1024	0.349	1.212	0.320	1.09×
2048	256	0.341	1.314	0.212	1.61×
2048	1024	1.175	1.346	0.620	1.90×
4096	64	0.477	1.429	0.163	2.93×
4096	256	1.316	1.394	0.371	3.55×
4096	1024	4.468	1.475	1.273	3.51×