PAIRWISE-INDEPENDENT DITHERING FOR SINGLE-STAGE HADAMARD QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantizing high-dimensional vectors is fundamental to similarity search, distributed learning, and model compression. Feng, Indyk, Kapralov, Krachun, and Prokhorov established sharp guarantees for an unbiased dithered quantizer based on a randomized Hadamard transform (Feng et al., 2026, NeurIPS). Their $1/d$-scale inner-product estimator, however, uses a second randomized transform and residual quantization, increasing both communication and the leading constant in the proved bound. We show that this extra stage is unnecessary: pairwise-independent dithers across Hadamard coordinates suffice. The resulting unbiased single-stage estimator uses b bits per coordinate and achieves

$$\mathbb{E}|\langle y, \widehat{x} - x\rangle|^2 \le \left(\frac{3\pi\sqrt{3}}{2} + o(1)\right) \frac{\|y\|_2^2}{d\,4^b},$$

as $b \to \infty$, with a dimension-free $o(1)$ term uniform over unit inputs and fixed queries. Compared with the two-stage construction of Feng et al., it eliminates the residual-stage $O(d)$-bit payload and reduces the leading upper-bound constant by a factor of approximately 5.93.

1 INTRODUCTION

Vector quantization represents a high-dimensional vector by a short binary string while retaining the geometric information needed by downstream tasks. At b bits per coordinate, a vector quantizer consists of an encoder $\mathsf{Enc} : \mathbb{R}^d \to \{0,1\}^{bd}$ and a decoder $\mathsf{Dec} : \{0,1\}^{bd} \to \mathbb{R}^d$; on input $x \in \mathbb{R}^d$, it returns the reconstruction $\widehat{x} = \mathsf{Dec}(\mathsf{Enc}(x))$. The encoder and decoder may use shared randomness, which is not counted toward the representation length. The goal is to preserve geometric information about x, such as reconstruction accuracy, distances, and inner products, while keeping the representation small. This primitive is used in similarity search (Charikar, 2002; Jégou et al., 2011; Datar et al., 2004; Dong et al., 2008; Gao & Long, 2024; Gao et al., 2025), federated learning (Vargaftik et al., 2021; 2022; Ben-Basat et al., 2024), efficient transformers and KV-cache compression (Zandieh et al., 2026), and model compression (Gholami et al., 2022; Ashkboos et al., 2024; Tseng et al., 2024).

A widely used data-oblivious template has the form $\mathsf{Enc}(x) = \mathrm{quant}(Rx)$, where R is a random rotation matrix and $\mathrm{quant}(\cdot)$ is often a scalar quantizer applied coordinatewise (Charikar, 2002; Datar et al., 2004; Dong et al., 2008; Ashkboos et al., 2024). Broader variants replace the scalar rule by structured vector codebooks (Andoni et al., 2015; Tseng et al., 2024). The rotation distribution and the codebook are chosen independently of the dataset, so no data-dependent training is required. Moreover, the method is amenable to analysis because each coordinate of Rx is approximately normally distributed when R is a uniformly random rotation (Feng et al., 2026). Variants of this template include DRIVE and EDEN for distributed mean estimation (Vargaftik et al., 2021; 2022), RaBitQ for distance and inner-product estimation (Gao & Long, 2024; Gao et al., 2025), and TurboQuant (Zandieh et al., 2026).

Despite these advantages, applying a dense random rotation matrix R requires $\Theta(d^2)$ time, which can be prohibitively expensive in high dimensions. A standard way to reduce this cost is to use randomized Hadamard transforms (Ailon & Chazelle, 2009; Sarlós, 2006; Woolfe et al., 2008; Krahmer & Ward, 2011; Halko et al., 2011). Specifically, one replaces R with HD, where H is a normal-

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.

ized Hadamard matrix and D is a diagonal matrix of independent random signs. For the Walsh–Hadamard choice, the resulting transform can be applied in $O(d \log d)$ time using the fast Walsh–Hadamard algorithm. However, this speedup comes at an analytical cost: HD is discrete, and each coordinate of HDx is an input-dependent weighted sum of the random signs; for worst-case inputs, this sum need not be approximately normal (Feng et al., 2026).

Feng, Indyk, Kapralov, Krachun, and Prokhorov (2026) overcame this difficulty for vector distortion by combining a single randomized Hadamard transform with a dithered, Gaussian-companded scalar quantizer. At b bits per coordinate, their reconstruction is exactly unbiased and satisfies, uniformly over unit inputs,

$$\mathbb{E}\|\widehat{x} - x\|_2^2 \le \left(\frac{\pi\sqrt{3}}{2} + o(1)\right) 4^{-b},$$

as $b \to \infty$. For fixed-query inner products, their separate construction uses a closely related two-stage scheme: after an initial randomized Hadamard quantization, it projects the reconstruction onto the unit ball and quantizes the residual after a fresh randomized Hadamard transform. The residual stage yields the desired $1/d$ scaling, but it also adds communication and enlarges the leading constant in the proved bound.

In this work, we show that the residual stage is unnecessary. In the base quantizer, one scalar dither is shared across all Hadamard coordinates, so the coordinate errors can remain correlated even though each is centered. We replace the shared dither by pairwise-independent dithers. Conditioned on D, the errors are then centered and pairwise independent, so every off-diagonal term in the inner-product second moment vanishes. Thus pairwise independence suffices for the desired second-moment bound.

1.1 Related work

Random rotations and coordinate-wise quantization. Random projection followed by scalar quantization is classical in similarity estimation and locality-sensitive hashing (Charikar, 2002; Datar et al., 2004; Dong et al., 2008). In distributed mean estimation, Suresh et al. (2017) combined a randomized Hadamard transform with stochastic quantization, while DRIVE and EDEN developed rotation-based schemes that include unbiased reconstruction and flexible rates (Vargaftik et al., 2021; 2022). Thus this template, including unbiased variants, predates TurboQuant. In vector search, RaBitQ and its multi-bit extension give rigorous distance and inner-product guarantees, including an asymptotically optimal space–error tradeoff (Gao & Long, 2024; Gao et al., 2025), consistent with the general compression bounds of Alon & Klartag (2017). TurboQuant later combined a dense-rotation MSE quantizer with residual quantization (Zandieh et al., 2026). Its relationship to DRIVE/EDEN and RaBitQ is discussed in (Ben-Basat et al., 2026a; Gao et al., 2026).

Dithering. Randomized offsets control bias and input dependence of quantization noise (Schuchman, 1964; Zamir & Feder, 1992). Combined with random projections, dithered scalar grids support similarity and distance estimation (Datar et al., 2004; Dong et al., 2008). Feng et al. (2026) share one scalar offset across all transformed coordinates. We retain their scalar quantizer and change only the dependence among offsets.

Fast structured rotations. Randomized Hadamard transforms are central to fast embedding and randomized linear-algebra methods (Ailon & Chazelle, 2009; Halko et al., 2011; Krahmer & Ward, 2011; Tropp, 2011). Approximation of dense random rotations by one or more such transforms has been studied for a single transform (Cherapanamjeri & Nelson, 2022), two structured blocks (Zilca & Mendelson, 2026), and multiple transforms with blockwise dependence control (Ben-Basat et al., 2026b). Direct single-transform quantization analyses appear in distributed mean estimation (Ben-Basat et al., 2024) and, using dithering, in the sharp high-rate result of Feng et al. (2026).

Relation to Feng et al. We use the unbiased Gaussian-companded scalar quantizer of Feng et al. (2026), including its reconstruction map, averaging identity, and scalar error estimates. Our new ingredients are the pairwise-independent dither family, conditional covariance diagonalization, and the weighted mixed-Rademacher estimate for the remaining diagonal terms.

1.2 Our result

We work first in a dimension d admitting a normalized Hadamard matrix; in particular, one may take d to be a power of two. Let $B = 2^b$. The encoder and decoder share the signs in D and two independent random variables $V_0, V_1 \sim \mathrm{Unif}[0,1)$. For coordinate $i \in \{1, \dots, d\}$, define

$$U_i = (V_0 + iV_1) \bmod 1. \tag{1.1}$$

The family $(U_i)_{i=1}^d$ is pairwise independent and each U_i is uniform on $[0,1)$. The quantizer applies the scalar unbiased reconstruction rule with dither U_i to the ith coordinate of $\sqrt{d}HDx$ and transmits one symbol from $\{0, \dots, B-1\}$ per coordinate.

Theorem 1.1 (Main theorem). *Let d admit a normalized Hadamard matrix, let $b \geq 2$ be an integer, and set $B = 2^b$. For every fixed $x \in \mathbb{S}^{d-1}$ and $y \in \mathbb{R}^d$, the single-stage quantizer in Algorithm 1, with D independent of (V_0, V_1), satisfies*

$$\mathbb{E}[\widehat{x}] = x, \tag{1.2}$$

and

$$\mathbb{E}|\langle y, \widehat{x} - x\rangle|^2 \leq \left(\frac{3\pi\sqrt{3}}{2} + o_B(1)\right) \frac{\|y\|_2^2}{dB^2}. \tag{1.3}$$

The term $o_B(1)$ tends to zero as $B \to \infty$ uniformly over all dimensions d admitting a normalized Hadamard matrix, all $x \in \mathbb{S}^{d-1}$, and all nonzero $y \in \mathbb{R}^d$. The communicated representation consists of exactly db bits, excluding the shared randomness.

The two-stage inner-product construction of Feng et al. (2026, Theorem 1.2) has proved leading constant $13\left(\frac{\pi\sqrt{3}}{2} + 1\right) \approx 48.369$ and can be encoded using at most $db + \left(3 + \frac{1}{2\ln 2}\right) d + O(\log(b + \log d))$ bits. Thus our single-stage estimator removes the residual payload. Relative to this stated upper bound, its communication guarantee is smaller by approximately $3.72d$ bits, up to lower-order terms. It also reduces the proved leading constant from approximately 48.369 to 8.162, an improvement by a factor of approximately 5.93, while preserving the vector mean-squared distortion guarantee of the base quantizer.

1.3 Technical overview

We use the scalar quantizer introduced by Feng et al. (2026). It first maps a scalar t to quantile space via $F(t) = \Phi(t/\sqrt{3})$ and then quantizes $F(t)$ on a uniformly shifted B-level grid. The decoder maps the selected grid point back to the real line using a modified inverse-compander G_B. If $\mathsf{Q}_U(t)$ denotes the reconstruction with dither U, the construction satisfies the exact averaging identity $\mathbb{E}_U[\mathsf{Q}_U(t)] = t$ and, in the central region, the variance bound

$$\mathbb{E}_U\left[(\mathsf{Q}_U(t) - t)^2\right] \leq \left(\frac{\pi}{2} + o_B(1)\right) B^{-2} e^{t^2/3}.$$

A coarser global bound handles the tails. Feng et al. use a single scalar dither across all Hadamard coordinates. In their separate inner-product construction, the first-stage reconstruction is projected onto the unit ball and the resulting residual is quantized after a fresh randomized Hadamard transform; the fresh transform provides the decorrelation needed for the $1/d$ variance reduction. Our construction instead builds this decorrelation into the unbiased single-stage quantizer: we replace the shared dither by the pairwise-independent family $(U_i)_{i=1}^d$ and use the reconstruction directly.

The role of pairwise independence becomes clear upon expanding the inner-product error. Set $z = \sqrt{d}HDx$, $e_i = \mathsf{Q}_{U_i}(z_i) - z_i$, and, for $y \neq 0$, $w_i = \sqrt{d}(HDy)_i / \|y\|_2$. Orthogonality gives

$$\langle y, \widehat{x} - x\rangle = \frac{\|y\|_2}{d} \sum_{i=1}^d w_i e_i.$$

Once D is fixed, the quantities z_i, w_i are deterministic and e_i depends only on U_i. The averaging identity centers each error, while pairwise independence eliminates the cross terms: $\mathbb{E}[e_i \mid D] = 0$ and $\mathbb{E}[e_i e_j \mid D] = 0$ for $i \neq j$. Since the target is a second moment, no higher-order independence is needed. Averaging over D therefore gives

$$\mathbb{E}|\langle y, \widehat{x} - x\rangle|^2 = \frac{\|y\|_2^2}{d^2} \sum_{i=1}^d \mathbb{E}[w_i^2 e_i^2].$$

There are d summands and a prefactor d^{-2}, so a dimension-free $O(B^{-2})$ bound for each summand yields the desired factor $1/d$.

The remaining issue is therefore a weighted scalar estimate. For a fixed coordinate, (z_i, w_i) has the same joint law as $X = \sum_j a_j \varepsilon_j$ and $Y = \sum_j c_j \varepsilon_j$, where the two sums use the same Rademacher signs, $\|a\|_2 = \|c\|_2 = 1$, $a_j = \sqrt{d} H_{ij} x_j$, and $c_j = \sqrt{d} H_{ij} y_j / \|y\|_2$. The weight Y^2 carries the query direction and prevents us from applying the unweighted scalar bound directly. On the central event, the variance estimate above produces the mixed moment $\mathbb{E}[Y^2 e^{X^2/3}]$. On the complement, the global bound, Markov's inequality applied to the tilted exponential-moment estimate, and layer-cake integration give a uniform $o_B(1)B^{-2}$ contribution. Together these estimates yield

$$\mathbb{E}\left[Y^2 \big(\mathsf{Q}_U(X) - X\big)^2\right] \le \left(\frac{\pi}{2}\mathbb{E}[Y^2 e^{X^2/3}] + o_B(1)\right) B^{-2}.$$

The sums X and Y are generally dependent because they use the same signs. Nevertheless, a direct exponential-moment calculation for the underlying Rademacher signs, followed by a Gaussian-mixture representation of $e^{X^2/3}$, gives

$$\mathbb{E}[Y^2 e^{\lambda X}] \le (1+\lambda^2) e^{\lambda^2/2} \quad \Longrightarrow \quad \mathbb{E}[Y^2 e^{X^2/3}] \le 3\sqrt{3}.$$

Each diagonal term is therefore at most $\left(3\pi\sqrt{3}/2 + o_B(1)\right) B^{-2}$, and summing over the d coordinates proves Equation (1.3).

2 Preliminaries

Throughout, $x \in \mathbb{S}^{d-1}$ is a fixed input and $y \in \mathbb{R}^d$ is a fixed query. For an integer bit width $b \ge 2$, set $B = 2^b$ and $\delta = (B-1)^{-1}$. Unless stated otherwise, expectations are over all randomness in the construction.

2.1 Randomized Hadamard transforms

A normalized Hadamard matrix is a matrix $H \in \mathbb{R}^{d\times d}$ satisfying $H_{ij} \in \{\pm d^{-1/2}\}$ and $H^\top H = I_d$. The normalized Walsh–Hadamard matrix provides such an H whenever d is a power of two. Let $D = \mathrm{diag}(\varepsilon_1, \dots, \varepsilon_d)$, where $\varepsilon_1, \dots, \varepsilon_d$ are independent random signs. We refer to HD as a randomized Hadamard transform.

An arbitrary ambient dimension n can be handled by zero-padding the input to $d = 2^{\lceil \log_2 n \rceil} < 2n$, applying the scheme in $\mathbb{R}^d$, and discarding the final $d - n$ coordinates after decoding; the query is zero-padded for the error analysis. The resulting representation uses $db < 2nb$ bits, and the directional error bound holds with the padded dimension d.

2.2 Gaussian-companded scalar quantization

Let Φ denote the standard Gaussian distribution function. Following Feng et al. (2026), define the compander

$$F(t) = \Phi\left(\frac{t}{\sqrt{3}}\right), \qquad F'(t) = \frac{1}{\sqrt{6\pi}} e^{-t^2/6}.$$

The identity

$$\frac{1}{12F'(t)^2} = \frac{\pi}{2} e^{t^2/3} \tag{2.1}$$

will determine the leading scalar error constant.

Let G_B be the reconstruction map constructed by Feng et al. (2026, Lemma A.8): $G_B : \left[-\frac{\delta}{2}, 1+\frac{\delta}{2}\right] \to \mathbb{R}$. It satisfies the averaging identity

$$\frac{1}{\delta}\int_{r-\delta/2}^{r+\delta/2} G_B(s)\,\mathrm{d}s = F^{-1}(r), \qquad r \in (0,1).$$

For $U \sim \mathrm{Unif}[0,1)$ and $t \in \mathbb{R}$, define the transmitted index and scalar reconstruction by

$$J_U(t) = \lfloor (B-1)F(t) - U \rfloor + 1, \qquad \mathsf{Q}_U(t) = G_B\left(\frac{J_U(t) + U - 1/2}{B-1}\right).$$

Since $F(t) \in (0,1)$, the index $J_U(t)$ lies in $\{0,\dots,B-1\}$. For fixed t, the argument of G_B is distributed as $F(t)+V\delta$, where $V \sim \mathrm{Unif}[-1/2,1/2]$. The averaging identity therefore gives

$$\mathbb{E}_U[\mathsf{Q}_U(t)] = t \qquad (t \in \mathbb{R}). \tag{2.2}$$

For this reconstruction rule, the following estimates are direct consequences of Lemma A.8 and the proof of Lemma A.9 in Feng et al. (2026).

Proposition 2.1 (Scalar error bounds)**.** *Let*

$$\psi_B(t) = \mathbb{E}_U\big[(\mathsf{Q}_U(t)-t)^2\big], \qquad M_B = \sqrt{5\log B}.$$

Then the following bounds hold.

1. *Uniformly for $|t| \le M_B$,*

$$\psi_B(t) \le (1+o_B(1))\frac{\delta^2}{12F'(t)^2} = (1+o_B(1))\frac{\pi\delta^2}{2}e^{t^2/3}. \tag{2.3}$$

2. *There are absolute constants C_1, C_2 such that, for every $t \in \mathbb{R}$,*

$$\psi_B(t) \le C_1\left(1+t^2+\frac{\delta^2}{F'(t)^2}\right) \le C_2\left(1+t^2+\delta^2 e^{t^2/3}\right). \tag{2.4}$$

3 The quantizer and its error decomposition

We now define the single-stage quantizer and derive the conditional moment identity that reduces its analysis to a weighted estimate for Rademacher sums.

3.1 Pairwise-independent dithers

We retain the scalar quantization rule from Section 2, but replace the single dither shared across Hadamard coordinates by a pairwise-independent family. Two independent uniform random variables suffice to generate it.

Lemma 3.1 (Affine dither construction)**.** *Let V_0, V_1 be independent random variables, each uniform on $[0,1)$ and independent of D. For $i = 1,\dots,d$, define*

$$U_i = (V_0 + iV_1) \bmod 1, \tag{3.1}$$

where $a \bmod 1$ denotes the fractional part of a. Each U_i is uniform on $[0,1)$, and U_i and U_j are independent whenever $i \neq j$.

Proof. Conditional on $V_1 = v$, the variable U_i is a modulo-one translate of V_0 and is therefore uniform on $[0,1)$. Since this conditional distribution does not depend on v, the variables U_i and V_1 are independent. Now fix $i \neq j$ and set $W = ((j-i)V_1) \bmod 1$. The variable W is uniform on $[0,1)$: after partitioning $[0,1)$ into $|j-i|$ equal intervals, the map $v \mapsto ((j-i)v) \bmod 1$ covers $[0,1)$ once on each interval, with the orientation reversed when $j < i$. Since U_i is independent of V_1, it is also independent of W. Moreover, $U_j = (U_i + W) \bmod 1$. Conditioning on U_i leaves W uniform, so the right-hand side is conditionally uniform on $[0,1)$. Its distribution does not depend on U_i, proving that U_i and U_j are independent. □

The dither sequence can be generated in $O(d)$ time: after computing $U_1 = (V_0 + V_1) \bmod 1$, use $U_{i+1} = (U_i + V_1) \bmod 1$.

3.2 Single-stage encoder and decoder

Applying the scalar rule coordinatewise with the dithers above gives Algorithm 1. The scheme coincides with the base vector quantizer of Feng et al. (2026), except that its shared scalar dither is replaced by the affine family in Lemma 3.1. All randomness sampled in the algorithm is shared by the encoder and decoder.

Algorithm 1 Pairwise-independent single-stage Hadamard quantizer

Require: Dimension d, integer bit width $b \geq 2$, normalized Hadamard matrix $H \in \mathbb{R}^{d\times d}$

Set $B \leftarrow 2^b$.
Sample independent Rademacher signs $\varepsilon_1, \ldots, \varepsilon_d$ and set $D \leftarrow \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_d)$.
Sample independent $V_0, V_1 \sim \mathrm{Unif}[0, 1)$, independently of D.
for $i = 1, \ldots, d$ **do**
    $U_i \leftarrow (V_0 + iV_1) \bmod 1$.
end for
procedure $\mathsf{Enc}(x \in \mathbb{S}^{d-1})$
    $z \leftarrow \sqrt{d}\, HDx$.
    for $i = 1, \ldots, d$ **do**
        $j_i \leftarrow \lfloor (B-1)F(z_i) - U_i \rfloor + 1$.
    end for
    return $j = (j_1, \ldots, j_d) \in \{0, \ldots, B-1\}^d$.
end procedure
procedure $\mathsf{Dec}(j)$
    for $i = 1, \ldots, d$ **do**
        $q_i \leftarrow G_B\big((j_i + U_i - 1/2)/(B-1)\big)$.
    end for
    $\widehat{x} \leftarrow DH^\top (q/\sqrt{d})$.
    return $\widehat{x}$.
end procedure

Since each j_i belongs to an alphabet of size $B = 2^b$, fixed-length encoding uses exactly b bits per coordinate. The transmitted message therefore has exactly db bits; the shared randomness is not counted.

3.3 Conditional first and second moments

Define the transformed input and scalar errors by

$$z_i = \sqrt{d}(HDx)_i, \qquad e_i = \mathsf{Q}_{U_i}(z_i) - z_i. \tag{3.2}$$

For $y \neq 0$, define the normalized transformed query coefficients

$$w_i = \frac{\sqrt{d}(HDy)_i}{\|y\|_2}. \tag{3.3}$$

Proposition 3.2 (Conditional unbiasedness)**.** *For every fixed $x \in \mathbb{S}^{d-1}$, $\mathbb{E}[\widehat{x} \mid D] = x$.*

Proof. Coordinatewise scalar unbiasedness gives $\mathbb{E}[q \mid D] = z = \sqrt{d}HDx$. Hence

$$\mathbb{E}[\widehat{x} \mid D] = DH^\top \frac{\mathbb{E}[q \mid D]}{\sqrt{d}} = DH^\top HDx = x. \qquad \square$$

Proposition 3.3 (Conditional second moment)**.** *Conditioned on D, the errors $e_1, \ldots, e_d$ are centered and pairwise independent. Consequently, for every fixed $y \neq 0$,*

$$\mathbb{E}\left[|\langle y, \widehat{x} - x\rangle|^2 \,\middle|\, D\right] = \frac{\|y\|_2^2}{d^2} \sum_{i=1}^{d} w_i^2 \psi_B(z_i). \tag{3.4}$$

Proof. The decoder gives

$$\widehat{x} - x = DH^\top \left(\frac{q - z}{\sqrt{d}}\right) = DH^\top \left(\frac{e}{\sqrt{d}}\right),$$

and therefore

$$\langle y, \widehat{x} - x\rangle = \frac{\|y\|_2}{d} \sum_{i=1}^{d} w_i e_i. \tag{3.5}$$

Once D is fixed, every z_i and w_i is deterministic. The error e_i depends only on U_i. By Lemma 3.1, the errors are therefore pairwise independent under the remaining randomness. Moreover, Equation (2.2) gives $\mathbb{E}[e_i \mid D] = 0$. For $i \neq j$, it follows that

$$\mathbb{E}[e_i e_j \mid D] = \mathbb{E}[e_i \mid D]\mathbb{E}[e_j \mid D] = 0 \qquad (i \neq j).$$

Since $\mathbb{E}[e_i^2 \mid D] = \psi_B(z_i)$, expanding the square in Equation (3.5) proves Equation (3.4). □

We next express the diagonal terms in Equation (3.4) as weighted errors of Rademacher sums. Fix $y \neq 0$ and a row i of H, and set $\eta_j = \sqrt{d} H_{ij}\varepsilon_j, \qquad j = 1, \dots, d.$ These are independent Rademacher signs, and

$$z_i = \sum_{j=1}^{d} x_j \eta_j, \qquad w_i = \sum_{j=1}^{d} \frac{y_j}{\|y\|_2} \eta_j. \tag{3.6}$$

Both coefficient vectors in Equation (3.6) have Euclidean norm one. Averaging Equation (3.4) over D therefore gives

$$\mathbb{E}|\langle y, \widehat{x} - x\rangle|^2 = \frac{\|y\|_2^2}{d^2} \sum_{i=1}^{d} \mathbb{E}_D[w_i^2 \psi_B(z_i)], \tag{3.7}$$

where every summand has the form $\mathbb{E}[Y^2 \psi_B(X)]$ for two normalized Rademacher sums built from the same signs.

4 Scalar error for Rademacher sums

Throughout this section, $\eta_1, \dots, \eta_d$ are independent Rademacher random variables,

$$X = \sum_{j=1}^{d} a_j \eta_j, \qquad Y = \sum_{j=1}^{d} c_j \eta_j, \qquad \|a\|_2 = \|c\|_2 = 1,$$

and $U \sim \mathrm{Unif}[0, 1)$ is independent of the signs. The reduction in Equation (3.7) leads to the following estimate.

Lemma 4.1 (Weighted scalar estimate)**.** *Uniformly over the dimension and all unit coefficient vectors,*

$$\mathbb{E}\left[Y^2(\mathsf{Q}_U(X) - X)^2\right] \leq \left(\frac{\pi}{2}\mathbb{E}[Y^2 e^{X^2/3}] + o_B(1)\right) B^{-2}. \tag{4.1}$$

Consequently,

$$\mathbb{E}\left[Y^2(\mathsf{Q}_U(X) - X)^2\right] \leq \left(\frac{3\pi\sqrt{3}}{2} + o_B(1)\right) B^{-2}. \tag{4.2}$$

We first establish the exponential moment bounds used in its proof.

4.1 Exponential moment bounds

Lemma 4.2 (Exponential moment bounds)**.** *For every $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[Y^2 e^{\lambda X}\right] \leq (1 + \lambda^2) e^{\lambda^2/2}. \tag{4.3}$$

Moreover,

$$\mathbb{E}\left[Y^2 e^{X^2/3}\right] \leq 3\sqrt{3}, \tag{4.4}$$

and every normalized Rademacher sum X satisfies

$$\mathbb{E} e^{X^2/3} \leq \sqrt{3}. \tag{4.5}$$

Proof. The moment generating function of X satisfies

$$M(\lambda) := \mathbb{E} e^{\lambda X} = \prod_{j=1}^{d} \cosh(\lambda a_j) \leq e^{\lambda^2/2}. \tag{4.6}$$

Set $t_j = \tanh(\lambda a_j)$. Independence of the signs and direct expansion of Y^2 give

$$\frac{\mathbb{E}[Y^2 e^{\lambda X}]}{M(\lambda)} = \sum_j c_j^2(1-t_j^2) + \left(\sum_j c_j t_j\right)^2 \le 1 + \sum_j t_j^2 \le 1+\lambda^2,$$

where we used Cauchy–Schwarz and $|\tanh u| \le |u|$. Together with Equation (4.6), this proves Equation (4.3). Let $g \sim \mathcal{N}(0,1)$ be independent of the Rademacher variables. The Gaussian moment generating function gives $e^{X^2/3} = \mathbb{E}_g e^{\sqrt{2/3}\, gX}$. Tonelli's theorem and Equation (4.3), applied conditionally on g with $\lambda = \sqrt{2/3}\, g$, yield

$$\mathbb{E}\left[Y^2 e^{X^2/3}\right] = \mathbb{E}_g \mathbb{E}\left[Y^2 e^{\sqrt{2/3}\, gX}\right] \le \mathbb{E}_g\left[\left(1+\frac{2}{3}g^2\right) e^{g^2/3}\right].$$

For a standard Gaussian, $\mathbb{E}e^{g^2/3} = \sqrt{3}, \qquad \mathbb{E}[g^2 e^{g^2/3}] = 3\sqrt{3}$. This proves Equation (4.4). Finally,

$$\mathbb{E}e^{X^2/3} = \mathbb{E}_g \mathbb{E} e^{\sqrt{2/3}\, gX} \le \mathbb{E}_g e^{g^2/3} = \sqrt{3}$$

by Equation (4.6), proving Equation (4.5). □

4.2 Proof of the weighted estimate

Proof of Lemma 4.1. Recall

$$\psi_B(t) = \mathbb{E}_U[(\mathsf{Q}_U(t) - t)^2], \qquad M_B = \sqrt{5\log B}.$$

We split $\mathbb{E}[Y^2\psi_B(X)]$ over the central event $\{|X| \le M_B\}$ and its complement. On the central event, Equations (2.1) and (2.3) give

$$\begin{aligned} B^2\mathbb{E}\left[Y^2\psi_B(X)\mathbf{1}_{\{|X|\le M_B\}}\right] &\le (1+o_B(1))\frac{\pi}{2}\mathbb{E}\left[Y^2 e^{X^2/3}\mathbf{1}_{\{|X|\le M_B\}}\right] \\ &\le (1+o_B(1))\frac{\pi}{2}\mathbb{E}[Y^2 e^{X^2/3}], \end{aligned} \tag{4.7}$$

where $B^2\delta^2 = 1 + O(B^{-1})$. The bound in Equation (4.4) makes the resulting $o_B(1)$ uniform. It remains to show that the tail contribution is $o(B^{-2})$ uniformly. For $s > 0$, Markov's inequality and Equation (4.3), with $\lambda = s$, imply

$$\mathbb{E}[Y^2\mathbf{1}_{\{X>s\}}] \le e^{-s^2}\mathbb{E}[Y^2 e^{sX}] \le (1+s^2)e^{-s^2/2}.$$

The same argument applied to $-X$ gives

$$T(s) := \mathbb{E}[Y^2\mathbf{1}_{\{|X|>s\}}] \le 2(1+s^2)e^{-s^2/2}. \tag{4.8}$$

Layer-cake integration gives, for $M > 0$,

$$\mathbb{E}[Y^2X^2\mathbf{1}_{\{|X|>M\}}] = M^2T(M) + 2\int_M^\infty sT(s)\,\mathrm{d}s \le C(1+M^4)e^{-M^2/2}, \tag{4.9}$$

and

$$\mathbb{E}[Y^2 e^{X^2/3}\mathbf{1}_{\{|X|>M\}}] = e^{M^2/3}T(M) + \int_M^\infty \frac{2s}{3}e^{s^2/3}T(s)\,\mathrm{d}s \le C(1+M^4)e^{-M^2/6}. \tag{4.10}$$

The constant C is absolute in both displays. Taking $M = M_B$ in Equation (4.9), combining it with Equation (4.8), and using $M_B^2 = 5\log B$, gives

$$B^2\mathbb{E}\left[Y^2(1+X^2)\mathbf{1}_{\{|X|>M_B\}}\right] = o_B(1). \tag{4.11}$$

Similarly, Equation (4.10) gives

$$\mathbb{E}\left[Y^2 e^{X^2/3}\mathbf{1}_{\{|X|>M_B\}}\right] = o_B(1),$$

uniformly over the coefficient vectors and dimension. The global scalar estimate in Equation (2.4) therefore yields

$$\begin{aligned} B^2\mathbb{E}\left[Y^2\psi_B(X)\mathbf{1}_{\{|X|>M_B\}}\right] &\le CB^2\mathbb{E}\left[Y^2(1+X^2)\mathbf{1}_{\{|X|>M_B\}}\right] \\ &\quad + CB^2\delta^2\mathbb{E}\left[Y^2e^{X^2/3}\mathbf{1}_{\{|X|>M_B\}}\right] \\ &= o_B(1). \end{aligned}$$

Combining this with Equation (4.7) proves Equation (4.1). The consequence Equation (4.2) follows from Equation (4.4). □

Corollary 4.3 (Unweighted scalar estimate)**.** *Let $\eta_1, \dots, \eta_d$ be independent Rademacher variables, let $X = \sum_j a_j\eta_j$ with $\|a\|_2 = 1$, and let $U \sim \mathrm{Unif}[0,1)$ be independent of the signs. Then, uniformly over the dimension and all unit coefficient vectors,*

$$\mathbb{E}[(\mathsf{Q}_U(X) - X)^2] \le \left(\frac{\pi\sqrt{3}}{2} + o_B(1)\right) B^{-2}. \tag{4.12}$$

Proof. Let η_0 be a Rademacher sign independent of $U, \eta_1, \dots, \eta_d$, and write $\widetilde{X} = X + 0\eta_0$ and $Y = \eta_0$. These are normalized Rademacher sums built from the same signs, and $Y^2 = 1$. The claim follows from Equations (4.1) and (4.5). □

5 Proof of the main theorem

Proof of Theorem 1.1. Taking expectation over D in Proposition 3.2 gives $\mathbb{E}[\widehat{x}] = x$. The case $y = 0$ is immediate, so assume $y \neq 0$. Since U_i is uniform and independent of D,

$$\mathbb{E}_D[w_i^2\psi_B(z_i)] = \mathbb{E}_{D,U_i}\left[w_i^2(\mathsf{Q}_{U_i}(z_i) - z_i)^2\right].$$

Applying Equation (4.2) to the Rademacher pair in Equation (3.6) therefore gives, uniformly in i, d, x, y,

$$\mathbb{E}_D[w_i^2\psi_B(z_i)] \le \left(\frac{3\pi\sqrt{3}}{2} + o_B(1)\right) B^{-2}. \tag{5.1}$$

Substituting Equation (5.1) into Equation (3.7) and summing these d bounds proves Equation (1.3). The communication claim follows because each transmitted index belongs to an alphabet of size $B = 2^b$ and hence uses exactly b bits. The d indices therefore require exactly db bits; the shared randomness is not transmitted. □

Corollary 5.1 (Vector mean-squared error)**.** *Under the hypotheses of Theorem 1.1,*

$$\mathbb{E}\|\widehat{x} - x\|_2^2 \le \left(\frac{\pi\sqrt{3}}{2} + o_B(1)\right) B^{-2}. \tag{5.2}$$

The convergence is uniform over all d admitting a normalized Hadamard matrix and all $x \in \mathbb{S}^{d-1}$.

Proof. Orthogonality and Equation (3.2) give $\|\widehat{x} - x\|_2^2 = \frac{1}{d}\sum_{i=1}^d e_i^2$. For every fixed row i, the first identity in Equation (3.6) represents z_i as a normalized Rademacher sum under the randomness in D. Moreover, U_i is uniform and independent of D, so Corollary 4.3 applies to each coordinate. Averaging proves Equation (5.2). Thus the affine dither family matches the vector-distortion guarantee of Feng et al. (2026). □

Conclusion In this paper, we have shown that a single randomized Hadamard transform suffices for unbiased fixed-query inner-product estimation with $1/d$ variance reduction. Starting from the Gaussian-companded scalar quantizer of Feng et al. (2026), we retain its reconstruction rule and replace the shared dither by a pairwise-independent family generated from two shared uniform random variables. Conditional decorrelation and a weighted Rademacher bound give a uniform high-rate mean-squared-error bound of $O(\|y\|_2^2/(d4^b))$ using exactly b bits per coordinate under shared randomness, while preserving the vector-distortion guarantee. Compared with the two-stage construction, our scheme removes the residual payload and reduces the proved leading upper-bound constant by a factor of approximately 5.93.

AI USE STATEMENT

The proof was initially obtained using a research harness based on Gemini 3.1 Pro. The authors have verified the proof and edited it for clarity of presentation, and take responsibility for the final version. Generative AI tools were also used to assist with drafting and language editing of the manuscript.

REFERENCES

Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009. doi: 10.1137/060673096.

Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, pp. 639–650. IEEE Computer Society, 2017. doi: 10.1109/FOCS.2017.65.

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal LSH for angular distance. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1225–1233. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/2823f4797102ce1a1aec05359cc16dd9-Abstract.html.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100213–100240. Curran Associates, Inc., 2024. doi: 10.52202/079017-3180. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Ran Ben-Basat, Shay Vargaftik, Amit Portnoy, Gil Einziger, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Accelerating federated learning with quick distributed mean estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3410–3442. PMLR, 2024. URL https://proceedings.mlr.press/v235/ben-basat24a.html.

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:2604.18555, 2026a. URL https://arxiv.org/abs/2604.18555.

Ran Ben-Basat, William Kuszmaul, Michael Mitzenmacher, Amit Portnoy, and Shay Vargaftik. Quantizing with randomized Hadamard transforms: Efficient heuristic now proven. arXiv:2605.06014, 2026b. URL https://arxiv.org/abs/2605.06014.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 380–388. Association for Computing Machinery, 2002. doi: 10.1145/509907.509965.

Yeshwanth Cherapanamjeri and Jelani Nelson. Uniform approximations for randomized Hadamard transforms with applications. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 659–671. Association for Computing Machinery, 2022. doi: 10.1145/3519935.3519961.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pp. 253–262. Association for Computing Machinery, 2004. doi: 10.1145/997817.997857.

Wei Dong, Moses Charikar, and Kai Li. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 123–130. Association for Computing Machinery, 2008. doi: 10.1145/1390334.1390358.

Ying Feng, Piotr Indyk, Michael Kapralov, Dmitry Krachun, and Boris Prokhorov. Provable quantization with randomized Hadamard transform. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. URL https://arxiv.org/abs/2605.13810. To appear.

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, May 2024. doi: 10.1145/3654970.

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, June 2025. doi: 10.1145/3725413.

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:2604.19528, 2026. URL https://arxiv.org/abs/2604.19528.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In George K. Thiruvathukal, Yung-Hsiang Lu, Jaeyoun Kim, Yiran Chen, and Bo Chen (eds.), *Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence*, pp. 291–326. Chapman and Hall/CRC, 2022. doi: 10.1201/9781003162810-13.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57.

Felix Krahmer and Rachel Ward. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011. doi: 10.1137/100810447.

Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 143–152. IEEE Computer Society, 2006. doi: 10.1109/FOCS.2006.37.

Leonard Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, December 1964. doi: 10.1109/TCOM.1964.1088973.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3329–3337. PMLR, 2017. URL https://proceedings.mlr.press/v70/suresh17a.html.

Joel A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1–2):115–126, 2011. doi: 10.1142/S1793536911000787.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

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. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0397758f8990c1b41b81b43ac389ab9f-Abstract.html.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21984–22014. PMLR, 2022. URL https://proceedings.mlr.press/v162/vargaftik22a.html.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, November 2008. doi: 10.1016/j.acha.2007.12.002.

Ram Zamir and Meir Feder. On universal quantization by randomized uniform/lattice quantizers. *IEEE Transactions on Information Theory*, 38(2):428–436, March 1992. doi: 10.1109/18.119699.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tO3ASKZlok.

Tomer Zilca and Gal Mendelson. Approximating uniform random rotations by two-block structured Hadamard rotations in high dimensions. arXiv:2604.23418, 2026. URL https://arxiv.org/abs/2604.23418.