NeurIPS 2026

NEURAL INFORMATION PROCESSING SYSTEMS

Quantizing with randomized Hadamard transforms

Efficient heuristic now proven

Ran Ben-Basat
University College London

William Kuszmaul
Carnegie Mellon University

Michael Mitzenmacher
Harvard University

Amit Portnoy
Independent Researcher

Shay Vargaftik
VMware Research by Broadcom

x2 Transforms

Ensure approximately Gaussian coordinates

x3 Transforms

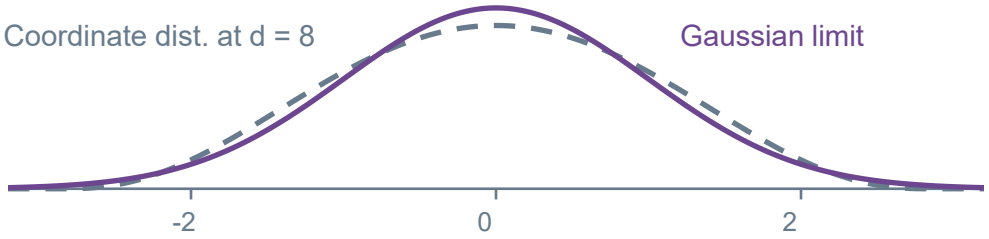
Ensure weak correlation

Uniform Random Rotations (URR) for quantization

Input vector $\boldsymbol{x} \in \mathbb{R}^d$ → Uniform random rotation $R\boldsymbol{x}$ → Normalize $\boldsymbol{u} = \frac{\sqrt{d}}{\|x\|} R\boldsymbol{x}$

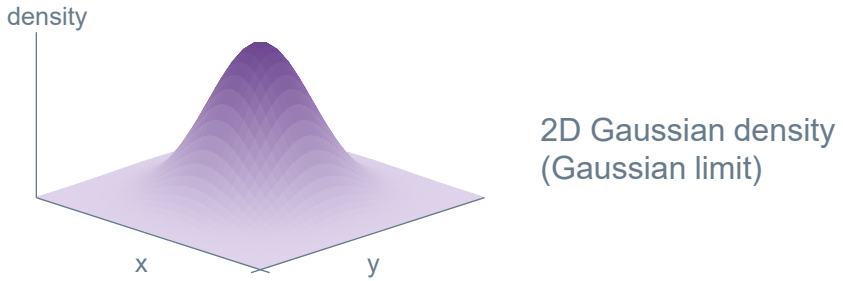
One coordinate

$$u_i \xrightarrow[d\to\infty]{dist.} \mathcal{N}(0,1)$$



A fixed number of coordinates

$$(u_{i_1}, \ldots, u_{i_k}) \xrightarrow[d\to\infty]{dist.} \mathcal{N}(0, I_k)$$



2D Gaussian density (Gaussian limit)

Scalar quantization

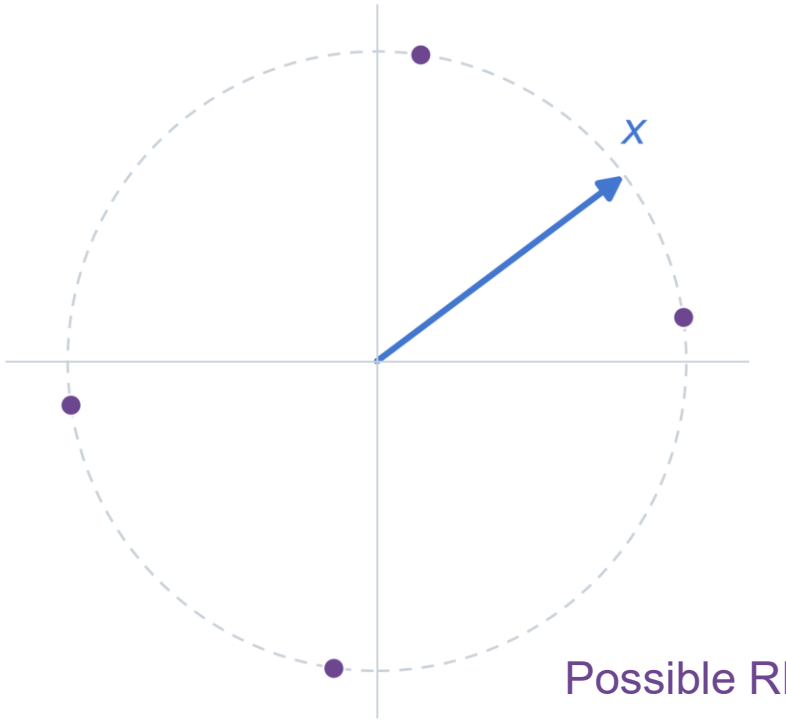
Optimized for a Gaussian

Vector Quantization (VQ)

Optimized for a joint Gaussian

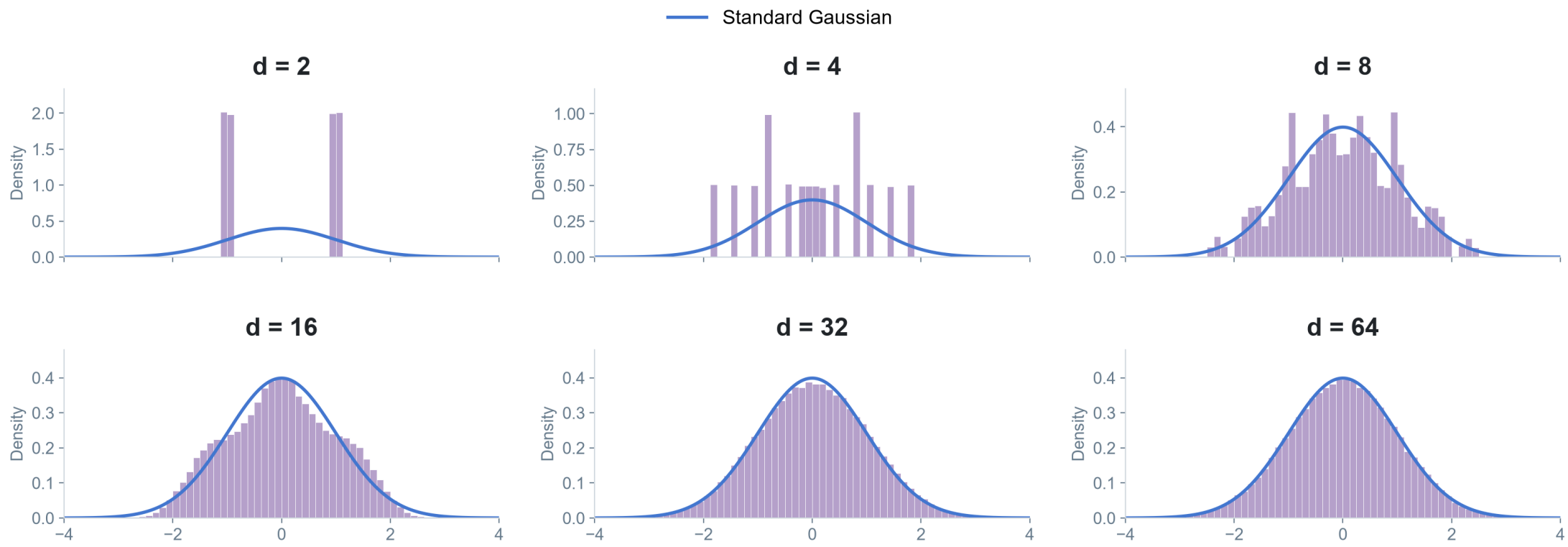
Approximating URR with the Randomized Hadamard Transforms (RHTs)

Uniform rotation: $\mathcal{O}(d^2)$ $\rightarrow$ RHT: $\mathcal{O}(d \log d)$



RHT on a random input

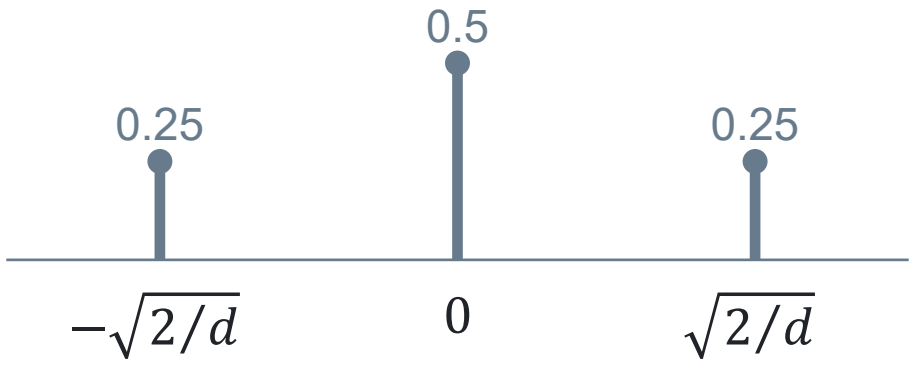
One random unit vector per dimension, held fixed; 100,000 repetitions.



RHT counterexample

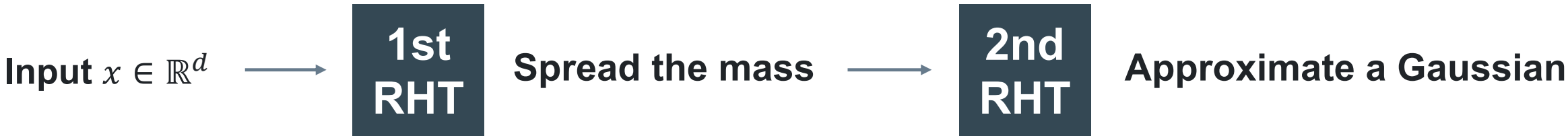
Two-sparse input: $x = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\right)$

The probability mass of each coordinate after one RHT:



Not Gaussian!

Scalar quantization with *two* RHTs



Every normalized coordinate approaches a standard Gaussian

$$\textit{Kolmogorov dist.} = \mathcal{O}(1/\sqrt{d}) \qquad \textit{1-Wasserstein dist.} = \mathcal{O}(1/\sqrt{d})$$

Gaussian-based scalar quantization error bounds transfer to two RHTs.

Examples:

DRIVE (NeurIPS '21)
EDEN (ICML '22)
QUIC-FL (ICML '24)

2 RHTs are not enough for vector quantization

Two-RHT counterexample



MSE: 1.473

Independent Gaussian



MSE: 0.727

Gaussian coordinates may not approach a joint Gaussian.

Three RHTs are sufficient for vector quantization

Input $x \in \mathbb{R}^d$ → **1st RHT** **Spread the mass**

→ **2nd RHT** **Approximate a Gaussian**

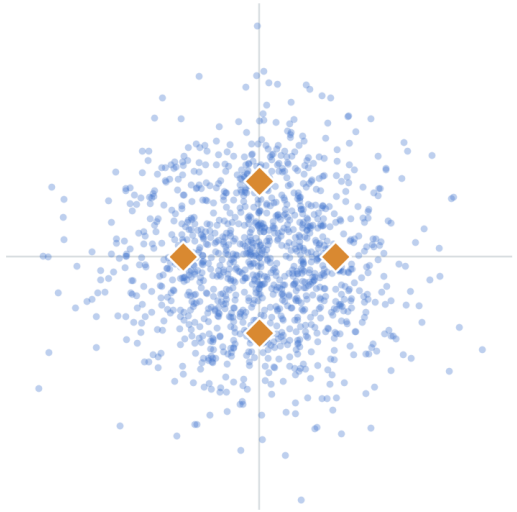
$\mathbb{E}(correlation^2)$ Can remain constant as dimension grows

→ **3rd RHT** **Decorrelate**

$$\mathbb{E}(correlation^2) = \mathcal{O}\left(\frac{1}{d}\right)$$

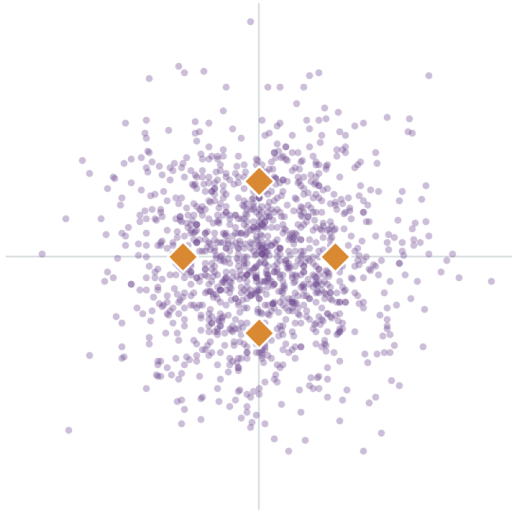
Same codebook, asymptotically the same error

Gaussian reference



Fixed quantization points

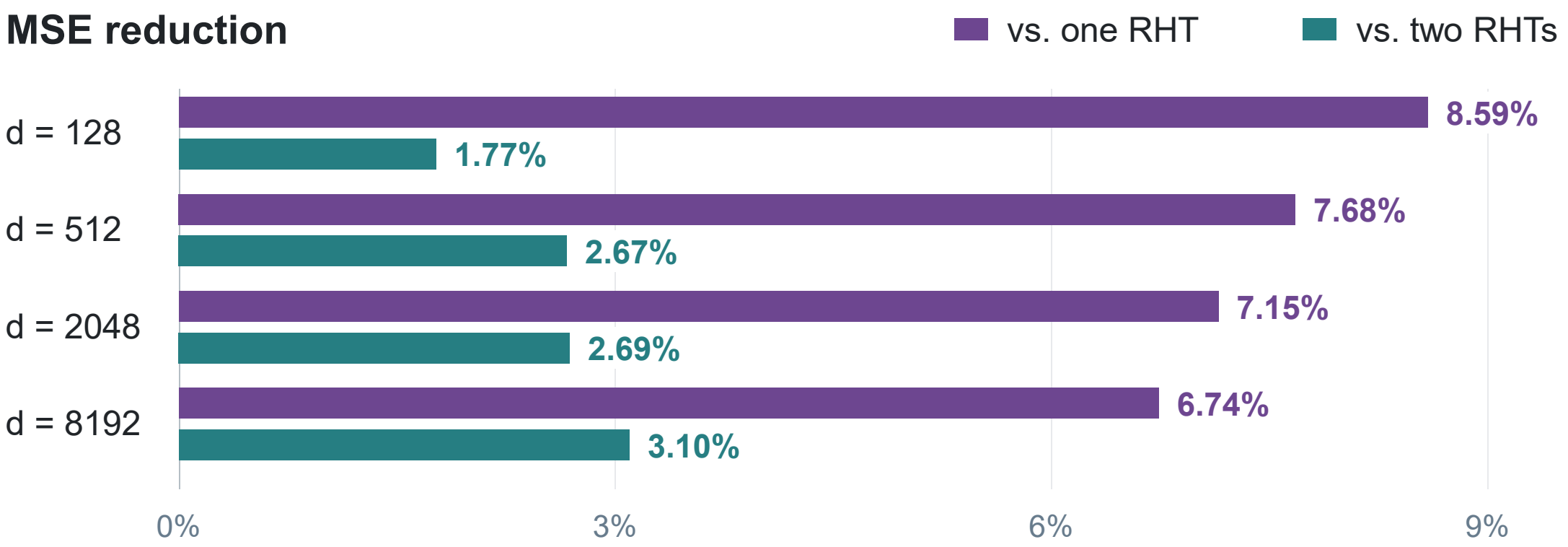
Three RHTs



Simulated, d = 256

$$|\mathrm{MSE}_{3\ \mathrm{RHTs}} - \mathrm{MSE}_{\mathrm{Gaussian}}| = \mathcal{O}\left(1/\sqrt{d}\right)$$

HIGGS (NAACL '25): lower error with three RHTs



32 consecutive random-sign entries; rest zero. 3 bits per coordinate. 2-coordinate blocks.

Higher is better

Inputs are rarely adversarial: checks to skip transforms

Scalar quantization

Normalized third absolute-moment mass

$$\frac{\sum_i |x_i|^3}{\left(\sum_i x_i^2\right)^{3/2}} \leq \eta_3$$

Pass → 1 RHT **Fail → 2 RHTs**

Vector quantization

Largest coordinate's share of energy

$$\frac{\max_i x_i^2}{\sum_i x_i^2} \leq \eta_\infty$$

Pass → 2 RHTs **Fail → 3 RHTs**

Both checks take $\mathcal{O}(d)$ time and $\mathcal{O}(1)$ space

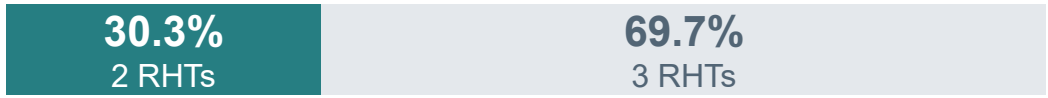
---

Experimental results over Qwen2.5-0.5B:

EDEN (ICML '22):



HIGGS VQ (NAACL '25):



NEURAL INFORMATION PROCESSING SYSTEMS

2 RHTs

Scalar quantization

Individual coordinates approach a standard Gaussian.

3 RHTs

Vector quantization

Block coordinates approach a standard multivariate Gaussian.

$\mathcal{O}(d)$-time tests select a sufficient number of transforms.