DRIVE: One-bit Distributed Mean Estimation

Shay Vargaftik *
VMware Research
shayv@vmware.com

Ran Ben Basat *
University College London
r.benbasat@cs.ucl.ac.uk

Amit Portnoy *
Ben-Gurion University
amitport@post.bgu.ac.il

Gal Mendelson
Stanford University
galmen@stanford.edu

Yaniv Ben-Itzhak
VMware Research
ybenitzhak@vmware.com

Michael Mitzenmacher
Harvard University
michaelm@eecs.harvard.edu

Abstract

We consider the problem where n clients transmit d-dimensional real-valued vectors using $d(1 + o(1))$ bits each, in a manner that allows the receiver to approximately reconstruct their mean. Such compression problems naturally arise in distributed and federated learning. We provide novel mathematical results and derive computationally efficient algorithms that are more accurate than previous compression techniques. We evaluate our methods on a collection of distributed and federated learning tasks, using a variety of datasets, and show a consistent improvement over the state of the art.

1 Introduction

In many computational settings, one wishes to transmit a d-dimensional real-valued vector. For example, in distributed and federated learning scenarios, multiple participants (a.k.a. *clients*) in distributed SGD send gradients to a parameter server that averages them and updates the model parameters accordingly [1]. In these applications and others (e.g., traditional machine learning methods such K-Means and power iteration [2] or other methods such as geometric monitoring [3]), sending *approximations* of vectors may suffice. Moreover, the vectors' dimension d is often large (e.g., in neural networks, d can exceed a billion [4, 5, 6]), so sending compressed vectors is appealing.

Indeed, recent works have studied how to send vector approximations using representations that use a small number of bits per entry (e.g., [2, 7, 8, 9, 10, 11]). Further, recent work has shown direct training time reduction from compressing the vectors to one bit per coordinate [12]. Most relevant to our work are solutions that address the distributed mean estimation problem. For example, [2] uses the randomized Hadamard transform followed by stochastic quantization (a.k.a. randomized rounding). When each of the n clients transmits $O(d)$ bits, their Normalized Mean Squared Error (NMSE) is bounded by $O\left(\frac{\log d}{n}\right)$. They also show a $O(\frac{1}{n})$ bound with $O(d)$ bits via variable-length encoding, albeit at a higher computational cost. The sampling method of [10] yields an $O\left(\frac{r \cdot R}{n}\right)$ NMSE bound using $d(1 + o(1))$ bits *in expectation*, where r is each coordinate's representation length and R is the normalized average variance of the sent vectors. Recently, researchers proposed to use Kashin's representation [11, 13, 14]. Broadly speaking, it allows representing a d-dimensional vector in (a higher) dimension $\lambda \cdot d$ for some $\lambda > 1$ using small coefficients. This results in an $O\left(\frac{\lambda^2}{(\sqrt{\lambda}-1)^4 \cdot n}\right)$ NMSE bound, where each client transmits $\lambda \cdot d(1 + o(1))$ bits [14].

We step back and focus on approximating d-dimensional vectors using $d(1 + o(1))$ bits (e.g., one bit per dimension and a lower order overhead). We develop novel biased and unbiased compression

*Equal Contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

techniques based on (uniform as well as structured) random rotations in high-dimensional spheres. Intuitively, after a rotation, the coordinates are identically distributed, allowing us to estimate each coordinate with respect to the resulting distribution. Our algorithms do not require expensive operations, such as variable-length encoding or computing the Kashin's representation, and are fast and easy to implement. We obtain an $O\left(\frac{1}{n}\right)$ NMSE bound using $d(1+o(1))$ bits, regardless of the coordinates' representation length, improving over previous works. Evaluation results indicate that this translates to a consistent improvement over the state of the art in different distributed and federated learning tasks.

2 Problem Formulation and Notation

1b - Vector Estimation. We start by formally defining the *1b - vector estimation* problem. A sender, called Buffy, gets a real-valued vector $x \in \mathbb{R}^d$ and sends it using a $d(1 + o(1))$ bits message (i.e., asymptotically *one bit per coordinate*). The receiver, called Angel, uses the message to derive an estimate $\widehat{x}$ of the original vector x. We are interested in the quantity $\|x - \widehat{x}\|_2^2$, which is the sum of squared errors (SSE), and its expected value, the Mean Squared Error (MSE). For ease of exposition, we hereafter assume that $x \neq 0$ as this special case can be handled with one additional bit. Our goal is to minimize the *vector*-NMSE (denoted *vNMSE*), defined as the normalized MSE, i.e., $\frac{\mathbb{E}\left[\|x-\widehat{x}\|_2^2\right]}{\|x\|_2^2}$.

1b - Distributed Mean Estimation. The above problem naturally generalizes to the *1b - Distributed Mean Estimation* problem. Here, we have a set of $n \in \mathbb{N}$ *clients* and a *server*. Each client $\mathfrak{c} \in \{1, \ldots, n\}$ has its own vector $x_{(\mathfrak{c})} \in \mathbb{R}^d$, which it sends using a $d(1+o(1))$-bits message to the server. The server then produces an estimate $\widehat{x_{\text{avg}}} \in \mathbb{R}^d$ of the average $x_{\text{avg}} = \frac{1}{n}\sum_{\mathfrak{c}=1}^{n} x_{(\mathfrak{c})}$ with the goal of minimizing its *NMSE*, defined as the average estimate's MSE normalized by the average norm of the clients' original vectors, i.e., $\frac{\mathbb{E}\left[\|x_{\text{avg}}-\widehat{x_{\text{avg}}}\|_2^2\right]}{\frac{1}{n}\cdot\sum_{\mathfrak{c}=1}^{n}\left\|x_{(\mathfrak{c})}\right\|_2^2}$.

Notation. We use the following notation and definitions throughout the paper:

Subscripts. x_i denotes the i'th *coordinate* of the vector x, to distinguish it from client $\mathfrak{c}$'s vector $x_{(\mathfrak{c})}$.

Binary-sign. For a vector $x \in \mathbb{R}^d$, we denote its binary-sign function as $\text{sign}(x)$, where $\text{sign}(x)_i = 1$ if $x_i \geqslant 0$ and $\text{sign}(x)_i = -1$ if $x_i < 0$.

Unit vector. For any (non-zero) real-valued vector $x \in \mathbb{R}^d$, we denote its normalized vector by $\breve{x} = \frac{x}{\|x\|_2}$. That is, $\breve{x}$ and x has the same direction and it holds that $\|\breve{x}\|_2 = 1$.

Rotation Matrix. A matrix $R \in \mathbb{R}^{d\times d}$ is a rotation matrix if $R^T R = I$. The set of all rotation matrices is denoted as $\mathcal{O}(d)$. It follows that $\forall R \in \mathcal{O}(d)\colon det(R) \in \{-1, 1\}$ and $\forall x \in \mathbb{R}^d\colon \|x\|_2 = \|Rx\|_2$.

Random Rotation. A random rotation $\mathcal{R}$ is a distribution over all random rotations in $\mathcal{O}(d)$. For ease of exposition, we abuse the notation and given $x \in \mathbb{R}^d$ denote the random rotation of x by $\mathcal{R}(x) = Rx$, where R is drawn from $\mathcal{R}$. Similarly, $\mathcal{R}^{-1}(x) = R^{-1}x = R^T x$ is the inverse rotation.

Rotation Property. A quantity that determines the guarantees of our algorithms is $\mathcal{L}^d_{\mathcal{R},x} = \frac{\|\mathcal{R}(\breve{x})\|_1^2}{d}$ (note the use of the L_1 norm). We show that rotations with high $\mathcal{L}^d_{\mathcal{R},x}$ values yield better estimates.

Shared Randomness. We assume that Buffy and Angel have access to shared randomness, e.g., by agreeing on a common PRNG seed. Shared randomness is studied both in communication complexity (e.g., [15]) and in communication reduction in machine learning systems (e.g., [2, 7]). In our context, it means that Buffy and Angel can generate the same random rotations without communication.

3 The DRIVE Algorithm

We start by presenting DRIVE (Deterministically RoundIng randomly rotated VEctors), a novel 1b - Vector Estimation algorithm. Later, we extend DRIVE to the 1b - Distributed Mean Estimation problem. In DRIVE, Buffy uses shared randomness to sample a rotation matrix $R \sim \mathcal{R}$ and rotates the vector $x \in \mathbb{R}^d$ by computing $\mathcal{R}(x) = Rx$. Buffy then calculates S, a scalar quantity we explain below. Buffy then sends $\left(S, \text{sign}(\mathcal{R}(x))\right)$ to Angel. As we discuss later, sending $\left(S, \text{sign}(\mathcal{R}(x))\right)$ requires $d(1 + o(1))$ bits. In turn, Angel computes $\widehat{\mathcal{R}(x)} = S \cdot \text{sign}(\mathcal{R}(x)) \in \{-S, +S\}^d$. It then

Algorithm 1 DRIVE

Buffy:
1: Compute $\mathcal{R}(x),\ S$.
2: Send $\big(S, \text{sign}(\mathcal{R}(x))\big)$ to Angel.

Angel:
1: Compute $\widehat{\mathcal{R}(x)} = S \cdot \text{sign}\big(\mathcal{R}(x)\big)$.
2: Estimate $\widehat{x} = \mathcal{R}^{-1}\big(\widehat{\mathcal{R}(x)}\big)$.

uses the shared randomness to generate the same rotation matrix and employs the inverse rotation, i.e., estimates $\widehat{x} = \mathcal{R}^{-1}(\widehat{\mathcal{R}(x)})$. The pseudocode of DRIVE appears in Algorithm 1.

The properties of DRIVE depend on the rotation $\mathcal{R}$ and the *scale parameter* S. We consider both uniform rotations, that provide stronger guarantees, and structured rotations that are orders of magnitude faster to compute. As for the scale $S = S(x, R)$, its exact formula determines the characteristics of DRIVE's estimate, e.g., having minimal vNMSE or being unbiased. The latter allows us to apply DRIVE to the 1b - Distributed Mean Estimation (Section 4.2) and get an NMSE that decreases proportionally to the number of clients.

We now prove a general result on the SSE of DRIVE that applies to any random rotation $\mathcal{R}$ and any vector $x \in \mathbb{R}^d$. In the following sections, we use this result to obtain the vNMSE when considering specific rotations and scaling methods as well as analyzing their guarantees.

Theorem 1. *The SSE of DRIVE is:* $\|x - \widehat{x}\|_2^2 = \|x\|_2^2 - 2 \cdot S \cdot \|\mathcal{R}(x)\|_1 + d \cdot S^2$.

Proof. The SSE in estimating $\mathcal{R}(x)$ using $\widehat{\mathcal{R}(x)}$ equals that of estimating x using $\widehat{x}$. Therefore,

$$\|x - \widehat{x}\|_2^2 = \|\mathcal{R}(x - \widehat{x})\|_2^2 = \|\mathcal{R}(x) - \mathcal{R}(\widehat{x})\|_2^2 = \left\|\mathcal{R}(x) - \widehat{\mathcal{R}(x)}\right\|_2^2$$
$$= \|\mathcal{R}(x)\|_2^2 - 2\left\langle \mathcal{R}(x), \widehat{\mathcal{R}(x)} \right\rangle + \left\|\widehat{\mathcal{R}(x)}\right\|_2^2 = \|x\|_2^2 - 2\left\langle \mathcal{R}(x), \widehat{\mathcal{R}(x)} \right\rangle + \left\|\widehat{\mathcal{R}(x)}\right\|_2^2 . \quad (1)$$

Next, we have that,

$$\left\langle \mathcal{R}(x), \widehat{\mathcal{R}(x)} \right\rangle = \sum_{i=1}^d \mathcal{R}(x)_i \cdot \widehat{\mathcal{R}(x)}_i = S \cdot \sum_{i=1}^d \mathcal{R}(x)_i \cdot \text{sign}\big(\mathcal{R}(x)_i\big) = S \cdot \|\mathcal{R}(x)\|_1 \ , \quad (2)$$

$$\left\|\widehat{\mathcal{R}(x)}\right\|_2^2 = \sum_{i=1}^d \widehat{\mathcal{R}(x)}_i^{\,2} = d \cdot S^2 \ . \quad (3)$$

Substituting Eq. (2) and Eq. (3) in Eq. (1) yields the result. □

4 DRIVE With a Uniform Random Rotation

We first consider the thoroughly studied uniform random rotation (e.g., [16, 17, 18, 19, 20]), which we denote by $\mathcal{R}_U$. The sampled matrix is denoted by $R_U \sim \mathcal{R}_U$, that is, $\mathcal{R}_U(x) = R_U \cdot x$. An appealing property of a uniform random rotation is that, as we show later, it admits a scaling that results in a low constant vNMSE even with unbiased estimates.

4.1 1b - Vector Estimation

Using Theorem 1, we obtain the following result. The result holds for any rotation, including $\mathcal{R}_U$.

Lemma 1. *For any $x \in \mathbb{R}^d$, DRIVE's SSE is minimized by $S = \frac{\|\mathcal{R}(x)\|_1}{d}$ (that is, $S = \frac{\|Rx\|_1}{d}$ is determined after $R \sim \mathcal{R}$ is sampled). This yields a vNMSE of $\frac{\mathbb{E}\left[\|x-\widehat{x}\|_2^2\right]}{\|x\|_2^2} = 1 - \mathbb{E}\left[\mathcal{L}^d_{\mathcal{R},x}\right]$.*

Proof. By Theorem 1, to minimize the SSE we require

$$\tfrac{\partial}{\partial S}\left(\|x\|_2^2 - 2 \cdot S \cdot \|\mathcal{R}(x)\|_1 + d \cdot S^2\right) = -2 \cdot \|\mathcal{R}(x)\|_1 + 2 \cdot d \cdot S = 0 \ ,$$

leading to $S = \frac{\|\mathcal{R}(x)\|_1}{d}$. Then, the SSE of DRIVE becomes:

$$\|x - \widehat{x}\|_2^2 = \|x\|_2^2 - 2 \cdot S \cdot \|\mathcal{R}(x)\|_1 + d \cdot S^2 = \|x\|_2^2 - 2 \cdot \tfrac{\|\mathcal{R}(x)\|_1^2}{d} + d \cdot \tfrac{\|\mathcal{R}(x)\|_1^2}{d^2}$$
$$= \|x\|_2^2 - \tfrac{\|\mathcal{R}(x)\|_1^2}{d} = \|x\|_2^2 - \tfrac{\|x\|_2^2 \cdot \|\mathcal{R}(\bar{x})\|_1^2}{d} = \|x\|_2^2 \left(1 - \tfrac{\|\mathcal{R}(\bar{x})\|_1^2}{d}\right) .$$

Thus, the normalized SSE is $\frac{\|x-\widehat{x}\|_2^2}{\|x\|_2^2} = 1-\mathcal{L}^d_{\mathcal{R},x}$. Taking expectation yields the result. □

Interestingly, for the uniform random rotation, $\mathcal{L}^d_{\mathcal{R}_U,x}$ follows the same distribution for all x. This is because, by the definition of $\mathcal{R}_U$, it holds that $\mathcal{R}_U(\breve{x})$ is distributed uniformly over the unit sphere for any x. Therefore DRIVE's vNMSE depends only on the dimension d. We next analyze the vNMSE attainable by the best possible S, as given in Lemma 1, when the algorithm uses $\mathcal{R}_U$ and is not required to be unbiased. In particular, we state the following theorem whose proof appears in Appendix A.1.

Theorem 2. *For any $x \in \mathbb{R}^d$, the vNMSE of DRIVE with $S = \frac{\|\mathcal{R}_U(x)\|_1}{d}$ is $\left(1-\frac{2}{\pi}\right)\left(1-\frac{1}{d}\right)$.*

4.2 1b - Distributed Mean Estimation

An appealing property of DRIVE with a uniform random rotation, established in this section, is that with a proper scaling parameter S, the estimate is unbiased. That is, for any $x \in \mathbb{R}^d$, our scale guarantees that $\mathbb{E}\left[\widehat{x}\right] = x$. Unbiasedness is useful when generalizing to the Distributed Mean Estimation problem. Intuitively, when n clients send their vectors, any biased algorithm would result in an NMSE that may not decrease with respect to n. For example, if they have the same input vector, the bias would remain after averaging. Instead, an unbiased encoding algorithm has the property that when all clients act (e.g., use different PRNG seeds) independently, the NMSE decreases proportionally to $\frac{1}{n}$.

Another useful property of uniform random rotation is that its distribution is unchanged when composed with other rotations. We use it in the following theorem's proof, given in Appendix A.2.

Theorem 3. *For any $x \in \mathbb{R}^d$, set $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$. Then DRIVE satisfies $\mathbb{E}[\widehat{x}] = x$.*

Now, we proceed to obtain vNMSE guarantees for DRIVE's unbiased estimate.

Lemma 2. *For any $x \in \mathbb{R}^d$, DRIVE with $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$ has a vNMSE of $\mathbb{E}\left[\frac{1}{\mathcal{L}^d_{\mathcal{R}_U,x}}\right]-1$.*

Proof. By Theorem 1, the SSE of the algorithm satisfies:

$$\|x\|_2^2 - 2\cdot S\cdot\|R_U\cdot x\|_1 + d\cdot S^2 = \|x\|_2^2 - 2\cdot\frac{\|x\|_2^2}{\|R_U\cdot x\|_1}\cdot\|R_U\cdot x\|_1 + d\cdot\left(\frac{\|x\|_2^2}{\|R_U\cdot x\|_1}\right)^2$$

$$= d\cdot\left(\frac{\|x\|_2^2}{\|x\|_2\|R_U\cdot\breve{x}\|_1}\right)^2 - \|x\|_2^2 = d\cdot\frac{\|x\|_2^2}{\|R_U\cdot\breve{x}\|_1^2} - \|x\|_2^2 = \|x\|_2^2\cdot\left(\left(\frac{d}{\|R_U\cdot\breve{x}\|_1^2}\right)-1\right).$$

Normalizing by $\|x\|_2^2$ and taking expectation over R_U concludes the proof. □

Our goal is to derive an upper bound on the above expression and thus upper-bound the vNMSE. Most importantly, we show that even though the estimate is unbiased and we use only a single bit per coordinate, the vNMSE does not increase with the dimension and is bounded by a small constant. In particular, in Appendix A.3, we prove the following:

Theorem 4. *For any $x \in \mathbb{R}^d$, the vNMSE of DRIVE with $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$ satisfies:*

(i) For all $d \geqslant 2$, it is at most 2.92. (ii) For all $d \geqslant 135$, it is at most $\frac{\pi}{2}-1+\sqrt{\frac{(6\pi^3-12\pi^2)\cdot\ln d+1}{d}}$.

This theorem yields strong bounds on the vNMSE. For example, the vNMSE is lower than 1 for $d \geqslant 4096$ and lower than 0.673 for $d \geqslant 10^5$. Finally, we obtain the following corollary,

Corollary 1. *For any $x \in \mathbb{R}^d$, the vNMSE tends to $\frac{\pi}{2}-1 \approx 0.571$ as $d \to \infty$.*

Recall that DRIVE's above scale S is a function of both x and the sampled R_U. An alternative approach is to *deterministically* set S to $\frac{\|x\|_2^2}{\mathbb{E}\left[\|\mathcal{R}_U(x)\|_1\right]}$. As we prove in Appendix A.4, the resulting scale is $\frac{\|x\|_2\cdot(d-1)\cdot\mathrm{B}(\frac{1}{2},\frac{d-1}{2})}{2d}$, where B is the Beta function. Interestingly, this scale no longer depends on x but only on its norm. In the appendix, we also prove that the resulting vNMSE is bounded by $\frac{\pi}{2}-1$ *for any* d. In practice, we find that the benefit is marginal.

Finally, with a vNMSE guarantee for the unbiased estimate by DRIVE, we obtain the following key result for the 1b - Distributed Mean Estimation problem, whose proof appears in Appendix A.5. We note that this result guarantees (e.g., see [21]) that distributed SGD, where the participants' gradients are compressed with DRIVE, converges at the same asymptotic rate as without compression.

Theorem 5. *Assume n clients, each with its own vector $x_{(\mathfrak{c})} \in \mathbb{R}^d$. Let each client independently sample $R_{U,\mathfrak{c}} \sim \mathcal{R}_U$ and set its scale to $\frac{\left\|x_{(\mathfrak{c})}\right\|_2^2}{\left\|R_{U,\mathfrak{c}} \cdot x_{(\mathfrak{c})}\right\|_1}$. Then, the server average estimate's NMSE satisfies: $\frac{\mathbb{E}\left[\left\|x_{avg} - \widehat{x_{avg}}\right\|_2^2\right]}{\frac{1}{n} \cdot \sum_{\mathfrak{c}=1}^{n} \left\|x_{(\mathfrak{c})}\right\|_2^2} = \frac{vNMSE}{n}$, where vNMSE is given by Lemma 2 and is bounded by Theorem 4.*

To the best of our knowledge, DRIVE is the first algorithm with a provable NMSE of $O(\frac{1}{n})$ for the 1b - Distributed Mean Estimation problem (i.e., with $d(1 + o(1))$ bits). In practice, we use only $d + O(1)$ bits to implement DRIVE. We use the $d(1 + o(1))$ notation to ensure compatibility with the theoretical results; see Appendix B for a discussion.

5 Reducing the vNMSE with DRIVE$^+$

To reduce the vNMSE further, we introduce the DRIVE$^+$ algorithm. In DRIVE$^+$, we also use a scale parameter, denoted $S^+ = S^+(x, R)$ to differentiate it from the scale S of DRIVE. Here, instead of reconstructing the rotated vector in a symmetric manner, i.e., $\widehat{\mathcal{R}(x)} \in S \cdot \{-1, 1\}^d$, we have that $\widehat{\mathcal{R}(x)} \in S^+ \cdot \{c_1, c_2\}^d$ where c_1, c_2 are computed using K-Means clustering with $K = 2$ over the d entries of the rotated vector $\mathcal{R}(x)$. That is, c_1, c_2 are chosen to minimize the SSE over any choice of two values. This does not increase the (asymptotic) time complexity over the random rotations considered in this paper as solving K-Means for the special case of one-dimensional data is deterministically solvable in $O(d \log d)$ (e.g., [22]). Notice that DRIVE$^+$ still requires $d(1 + o(1))$ bits as we communicate $(S^+ \cdot c_1, S^+ \cdot c_2)$ and a single bit per coordinate, indicating its nearest centroid. We defer the pseudocode and analyses of DRIVE$^+$ to Appendix C. We show that with proper scaling, for both the 1b - Vector Estimation and 1b - Distributed Mean Estimation problems, DRIVE$^+$ yields guarantees that are at least as strong as those of DRIVE.

6 DRIVE with a Structured Random Rotation

Uniform random rotation generation usually relies on QR factorization (e.g., see [23]), which requires $O(d^3)$ time and $O(d^2)$ space. Therefore, uniform random rotation can only be used in practice to rotate low-dimensional vectors. This is impractical for neural network architectures with many millions of parameters. To that end, we continue to analyze DRIVE and DRIVE$^+$ with the (randomized) Hadamard transform, a.k.a. *structured* random rotation [2, 24], that admits a fast *in-place*, parallelizable, $O(d \log d)$ time implementation [25, 26, 27]. We start with a few definitions.

Definition 1. *The Walsh-Hadamard matrix ([28]) $H_{2^k} \in \{+1, -1\}^{2^k \times 2^k}$ is recursively defined via: $H_{2^k} = \begin{pmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{pmatrix}$ and $H_1 = (1)$. Also, $(\frac{1}{\sqrt{d}} H) \cdot (\frac{1}{\sqrt{d}} H)^T = I$ and $det(\frac{1}{\sqrt{d}} H) \in [-1, 1]$.*

Definition 2. *Let R_H denote the rotation matrix $\frac{HD}{\sqrt{d}} \in \mathbb{R}^{d \times d}$, where H is a Walsh-Hadamard matrix and D is a diagonal matrix whose diagonal entries are i.i.d. Rademacher random variables (i.e., taking values uniformly in ± 1). Then $\mathcal{R}_H(x) = R_H \cdot x = \frac{1}{\sqrt{d}} H \cdot (x_1 \cdot D_{11}, \ldots, x_d \cdot D_{dd})^T$ is the randomized Hadamard transform of x and $\mathcal{R}_H^{-1}(x) = R_H^T \cdot x = \frac{DH}{\sqrt{d}} \cdot x$ is the inverse transform.*

6.1 1b - Vector Estimation

Recall that the vNMSE of DRIVE, when minimized using $S = \frac{\|\mathcal{R}(x)\|_1}{d}$, is $1 - \mathbb{E}\left[\mathcal{L}^d_{\mathcal{R},x}\right]$ (see Lemma 1). We now bound this quantity of DRIVE with a structured random rotation.

Lemma 3. *For any dimension $d \geqslant 2$ and vector $x \in \mathbb{R}^d$, the vNMSE of DRIVE with a structured random rotation and scale $S = \frac{\|\mathcal{R}_H(x)\|_1}{d}$ is: $1 - \mathbb{E}\left[\mathcal{L}^d_{\mathcal{R}_H,x}\right] \leqslant \frac{1}{2}$.*

Proof. Observe that for all i, $\mathbb{E}\left[|\mathcal{R}_H(x)_i|\right] = \mathbb{E}\left[\left|\sum_{j=1}^{d} \frac{x_j}{\sqrt{d}} H_{ij} D_{jj}\right|\right]$. Since $\{H_{ij}D_{jj} \mid j \in [d]\}$ are i.i.d. Rademacher random variables we can use the Khintchine inequality [29, 30] which implies that $\frac{1}{\sqrt{2d}} \cdot \|x\|_2 \leqslant \mathbb{E}\left[|\mathcal{R}_H(x)_i|\right] \leqslant \frac{1}{\sqrt{d}} \cdot \|x\|_2$ (see [31, 32] for simplified proofs). We conclude that:

$$\mathbb{E}\left[\mathcal{L}^d_{\mathcal{R}_H,x}\right] = \frac{1}{d} \cdot \mathbb{E}\left[\|\mathcal{R}_H(\breve{x})\|_1^2\right] \geqslant \frac{1}{d} \cdot \mathbb{E}\left[\|\mathcal{R}_H(\breve{x})\|_1\right]^2 \geqslant \frac{1}{d} \cdot \left(\sum_{i=1}^{d} \frac{1}{\sqrt{2d}}\right)^2 = \frac{1}{2} \,.$$

This bound is sharp since for $d \geqslant 2$ we have that $\mathcal{L}^d_{\mathcal{R}_H,x} = \frac{1}{2}$ for $x = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0)^T$. □

Observe that unlike for the uniform random rotation, $\mathbb{E}\left[\mathcal{L}^d_{\mathcal{R}_H,x}\right]$ depends on x. We also note that this bound of $\frac{1}{2}$ applies to DRIVE$^+$ (with scale $S^+ = 1$) as we show in Appendix C.2.

6.2 1b - Distributed Mean Estimation

For an arbitrary $x \in \mathbb{R}^d$ and $\mathcal{R}$, and in particular for $\mathcal{R}_H$, the estimates of DRIVE cannot be made unbiased. For example, for $x = (\frac{2}{3}, \frac{1}{3})^T$ we have that $\text{sign}(\mathcal{R}_H(x)) = (D_{11}, D_{11})^T$ and thus $\widehat{\mathcal{R}_H(x)} = S \cdot (D_{11}, D_{11})^T$. This implies that $\widehat{x} = \mathcal{R}_H^{-1}(\widehat{\mathcal{R}_H(x)}) = \frac{1}{\sqrt{2}} \cdot D \cdot H \cdot S \cdot (D_{11}, D_{11})^T = \sqrt{2} \cdot S \cdot D \cdot (D_{11}, 0)^T = \sqrt{2} \cdot S \cdot (D_{11}^2, 0)^T = (\sqrt{2} \cdot S, 0)^T$. Therefore, $\mathbb{E}[\widehat{x}] \neq x$ regardless of the scale.

Nevertheless, we next provide evidence for why when the input vector is high dimensional and admits finite moments, a structured random rotation performs similarly to a uniform random rotation, yielding all the appealing aforementioned properties. Indeed, it is a common observation that the distribution of machine learning workloads and, in particular, neural network gradients are governed by such distributions (e.g., lognormal [33] or normal [34, 35]).

We seek to show that at high dimensions, the distribution of $\mathcal{R}_H(x)$ is sufficiently similar to that of $\mathcal{R}_U(x)$. By definition, the distribution of $\mathcal{R}_U(x)$ is that of a uniformly at random distributed point on a sphere. Previous studies of this distribution for high dimensions (e.g., [36, 37, 38, 39]) have shown that individual coordinates of $\mathcal{R}_U(x)$ converge to the same normal distribution and that these coordinates are "weakly" dependent in the sense that the joint distribution of every $O(1)$-sized subset of coordinates is similar to that of independent normal variables for large d.

We hereafter assume that $x = (x_1, \ldots, x_d)$, where the x_is are i.i.d. and that $\mathbb{E}[x_j^2] = \sigma^2$ and $\mathbb{E}[|x_j|^3] = \rho < \infty$ for all j. We show that $\mathcal{R}_H(x)_i$ converges to the same normal distribution for all i. Let $F_{i,d}(x)$ be the cumulative distribution function (CDF) of $\frac{1}{\sigma} \cdot \mathcal{R}_H(x)_i$ and Φ be the CDF of the standard normal distribution. The following lemma, proven in Appendix D.1, shows the convergence.

Lemma 4. *For all i, $\mathcal{R}_H(x)_i$ converges to a normal variable:* $\sup_{x \in \mathbb{R}} |F_{i,d}(x) - \Phi(x)| \leqslant \frac{0.409 \cdot \rho}{\sigma^3 \sqrt{d}}$.

With this result, we continue to lay out evidence for the "weak dependency" among the coordinates. We do so by calculating the moments of their joint distribution in increasing subset sizes showing that these moments converge to those of independent normal variables. Previous work has shown that a structured random rotation on vectors with specific distributions results in "weakly dependent" normal variables. This line of research [40, 41, 42] utilized the Hadamard transform for a different purpose. Their goal was to develop a computationally cheap method to generate independent normally distributed variables from simpler (e.g., uniform) distributions. We apply their analysis to our setting.

We partially rely on the following observation that the Hadamard matrix satisfies.

Observation 1. *([40]) The Hadamard product (coordinate-wise product), $H_{\langle i \rangle} \circ H_{\langle \ell \rangle}$, of two rows $H_{\langle i \rangle}, H_{\langle \ell \rangle}$ in the Hadamard matrix yields another row at the matrix $H_{\langle i \rangle} \circ H_{\langle \ell \rangle} = H_{\langle 1+(i-1) \oplus (\ell-1) \rangle}$. Here, $(i-1) \oplus (\ell-1)$ is the bitwise xor of the $(\log d)$-sized binary representation of $(i-1)$ and $(\ell-1)$. It follows that $\sum_{j=1}^{d} H_{ij} H_{\ell j} = \sum_{j=1}^{d} (H_{\langle i \rangle} \circ H_{\langle \ell \rangle})_j = \sum_{j=1}^{d} H_{1+(i-1) \oplus (\ell-1), j}$.*

We now analyze the moments of the rotated variables, starting with the following observation. It follows from the sign-symmetry of D and matches the joint distribution of i.i.d. normal variables.

Observation 2. *All odd moments containing $\mathcal{R}_H(x)$ entries are 0. That is,*

$$\forall q \in \mathbb{N}, \forall i_1, \ldots, i_{2q+1} \in \{1, \ldots, d\} : \mathbb{E}\left[\mathcal{R}_H(x)_{i_1} \cdot \ldots \cdot \mathcal{R}_H(x)_{i_{2q+1}}\right] = 0.$$

Therefore, we need to examine only even moments. We start with showing that the second moments also match with the distribution of independent normal variables.

Lemma 5. *For all $i \neq \ell$ it holds that $\mathbb{E}\left[(\mathcal{R}_H \cdot x)_i \cdot (\mathcal{R}_H \cdot x)_\ell\right] = 0$, whereas $\mathbb{E}\left[(\mathcal{R}_H \cdot x)_i^2\right] = \sigma^2$.*

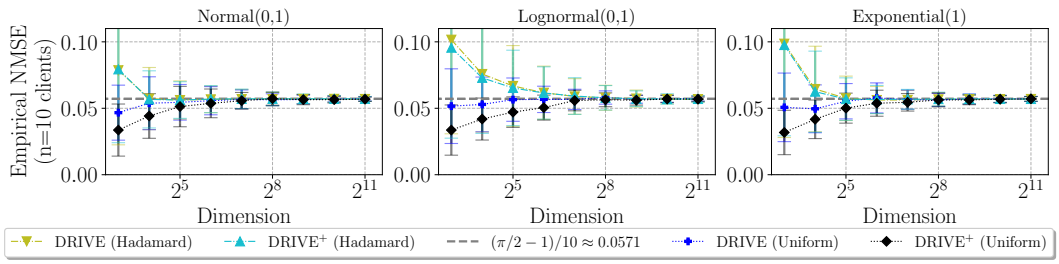

Figure 1: Distributed mean estimation comparison: each data point is averaged over 10^4 trials. In each trial, *the same* (randomly sampled) vector is sent by $n = 10$ clients.

Proof. Since $\{D_{jj} \mid j \in \{1,\dots,d\}\}$ are sign-symmetric and i.i.d., $\mathbb{E}\left[(\mathcal{R}_H \cdot x)_i \cdot (\mathcal{R}_H \cdot x)_\ell\right] = \mathbb{E}\left[\frac{1}{d}(\sum_{j=1}^d x_j H_{ij} D_{jj}) \cdot (\sum_{j=1}^d x_j H_{\ell j} D_{jj})\right] = \mathbb{E}\left[x_j^2\right] \cdot \frac{1}{d} \cdot \sum_{j=1}^d H_{ij} H_{\ell j}$. Notice that $\sum_{j=1}^d H_{1j} = d$ and $\sum_{j=1}^d H_{ij} = 0$ for all $i > 1$. Thus, by Observation 1 we get 0 if $i \neq \ell$ and σ^2 otherwise. □

We have established that the coordinates are pairwise uncorrelated. Similar but more involved analysis shows that the same trend continues under the assumption of the existence of x's higher moments. In Appendix D.2 we analyze the 4th moments showing that they indeed approach the 4th moments of independent normal variables with a rate of $\frac{1}{d}$; the reader is referred to [40] for further intuition and higher moments analysis. We therefore expect that using DRIVE and DRIVE$^+$ with Hadamard transform will yield similar results to that of a uniform random rotation at high dimensions and when the input vectors respect the finite moments assumption.

In addition to the theoretical evidence, in Figure 1, we show experimental results comparing the measured NMSE for the 1b - Distributed Mean Estimation problem with $n = 10$ clients (all given the *same* vector so biases do not cancel out) for DRIVE and DRIVE$^+$ using both uniform and structured random rotations over three different distributions. The results indicate that all variants yield similar NMSEs in reasonable dimensions, in line with the theoretical guarantee of Corollary 1 and Theorem 5.

7 Evaluation

We evaluate DRIVE and DRIVE$^+$, comparing them to standard and recent state-of-the-art techniques. We consider classic distributed learning tasks as well as federated learning tasks (e.g., where the data distribution is not i.i.d. and clients may change over time). All the distributed tasks are implemented over PyTorch [43] and all the federated tasks are implemented over TensorFlow Federated [44]. We focus our comparison on vector quantization algorithms and recent sketching techniques and exclude sparsification methods (e.g., [45, 46, 47, 48]) and methods that involve client-side memory since these can often work in conjunction with our algorithms.

Datasets. We use MNIST [49, 50], EMNIST [51], CIFAR-10 and CIFAR-100 [52] for image classification tasks; a next-character-prediction task using the Shakespeare dataset [53]; and a next-word-prediction task using the Stack Overflow dataset [54]. Additional details appear in Appendix E.1.

Algorithms. Since our focus is on the distributed mean estimation problem and its federated and distributed learning applications, we run DRIVE and DRIVE$^+$ with the unbiased scale quantities.[2]

We compare against several alternative algorithms: (1) *FedAvg* [1] that uses the full vectors (i.e., each coordinate is represented using a 32-bit float); (2) Hadamard transform followed by 1-bit stochastic quantization (SQ) [2, 10]; (3) Kashin's representation followed by 1-bit stochastic quantization [11]; (4) *TernGrad* [8], which clips coordinates larger than 2.5 times the standard deviation, then performs 1-bit stochastic quantization on the absolute values and separately sends their signs and the maximum coordinate for scale (we note that TernGrad is a low-bit variant of a well-known algorithm called

[2]For DRIVE the scale is $S = \frac{\|x\|_2^2}{\|\mathcal{R}(x)\|_1}$ (see Theorem 3). For DRIVE$^+$ the scale is $S^+ = \frac{\|x\|_2^2}{\|c\|_2^2}$, where $c \in \{c_1, c_2\}^d$ is the vector indicating the nearest centroid to each coordinate in $\mathcal{R}(x)$ (see Section 5).

Dimension (d)	Hadamard + 1-bit SQ	Kashin + 1-bit SQ	Drive (Uniform)	Drive$^+$ (Uniform)	Drive (Hadamard)	Drive$^+$ (Hadamard)
128	0.5308, *0.34*	0.2550, *2.12*	0.0567, *40.4*	**0.0547**, *40.7*	0.0591, *0.36*	0.0591, *0.72*
8,192	1.3338, *0.57*	0.3180, *3.42*	**0.0571**, *5088*	**0.0571**, *5101*	**0.0571**, *0.60*	**0.0571**, *1.06*
524,288	2.1456, *0.79*	0.3178, *4.69*	—	—	**0.0571**, *0.82*	**0.0571**, *1.35*
33,554,432	2.9332, *27.1*	0.3179, *332*	—	—	**0.0571**, *27.2*	**0.0571**, *37.8*

Table 1: Empirical NMSE and average per-vector encoding time (in milliseconds, on an RTX 3090 GPU) for distributed mean estimation with $n = 10$ clients (same as in Figure 1) and Lognormal(0,1) distribution. Each entry is a (NMSE, *time*) tuple and the most accurate result is highlighted in **bold**.

QSGD [9], and we use TernGrad since we found it to perform better in our experiments)[3]; and (5-6) *Sketched-SGD* [55] and *FetchSGD* [56], which are both count-sketch [57] based algorithms designed for distributed and federated learning, respectively.

We note that Hadamard with 1-bit stochastic quantization is our most fair comparison, as it uses the same number of bits as DRIVE$^+$ (and slightly more than DRIVE) and has similar computational costs. This contrasts with Kashin's representation, where both the number of bits and the computational costs are higher. For example, a standard TensorFlow Federated implementation (e.g., see "CLASS KASHINHADAMARDENCODINGSTAGE" hyperparameters at [58]) uses a minimum of 1.17 bits per coordinate, and three iterations of the algorithm resulting in five Hadamard transforms for each vector. Also, note that TernGrad uses an extra bit per coordinate for sending the sign. Moreover, the clipping performed by TernGrad is a heuristic procedure, which is orthogonal to our work.

For each task, we use a subset of datasets and the most relevant competition. Detailed configuration information and additional results appear in Appendix E. We first evaluate the vNMSE-Speed tradeoffs and then proceed to federated and distributed learning experiments.

vNMSE-Speed Tradeoff. Appearing in Table 1, the results show that our algorithms offer the lowest NMSE and that the gap increases with the dimension. As expected, DRIVE and DRIVE$^+$ with uniform rotation are more accurate for small dimensions but are significantly slower. Similarly, DRIVE is as accurate as DRIVE$^+$, and both are significantly more accurate than Kashin (by a factor of 4.4×-5.5×) and Hadamard (9.3×-51×) with stochastic quantization. Additionally, DRIVE is 5.7×-12× faster than Kashin and about as fast as Hadamard. In Appendix E.3 we discuss the results, give the complete experiment specification, and provide measurements on a commodity machine.

We note that the above techniques, including DRIVE, are more computationally expensive than linear-time solutions like TernGrad. Nevertheless, DRIVE's computational overhead becomes insignificant for modern learning tasks. For example, our measurements suggest that it can take 470 ms for computing the gradient on a ResNet18 architecture (for CIFAR100, batch size = 128, using NVIDIA GeForceGTX 1060 (6GB) GPU) while the encoding of DRIVE (Hadamard) takes 2.8 ms. That is, the overall computation time is only increased by 0.6% while the error reduces significantly. Taking the transmission and model update times into consideration would reduce the importance of the compression time further.

Federated Learning. We evaluate over four tasks: (1) EMNIST over customized CNN architecture with two convolutional layers with $\approx 1.2M$ parameters [59]; (2) CIFAR-100 over ResNet-18 [52] ; (3) a next-character-prediction task using the Shakespeare dataset [1]; (4) a next-word-prediction task using the Stack Overflow dataset [60]. Both (3) and (4) use LSTM recurrent models [61] with $\approx 820K$ and $\approx 4M$ parameters, respectively. We use code, client partitioning, models, hyperparameters, and validation metrics from the federated learning benchmark of [60].

The results are depicted in Figure 2. We observe that in all tasks, DRIVE and DRIVE$^+$ have accuracy that is competitive with that of the baseline, FedAvg. In CIFAR-100, TernGrad and DRIVE provide the best accuracy. For the other tasks, DRIVE and DRIVE$^+$ have the best accuracy, while the best alternative is either Kashin + 1-bit SQ or TernGrad, depending on the task. Hadamard + 1-bit SQ, which is the most similar to our algorithms (in terms of both band-

[3]When restricted to two quantization levels, TernGrad is identical to QSGD's max normalization variant with clipping (slightly better due to the ability to represent 0).

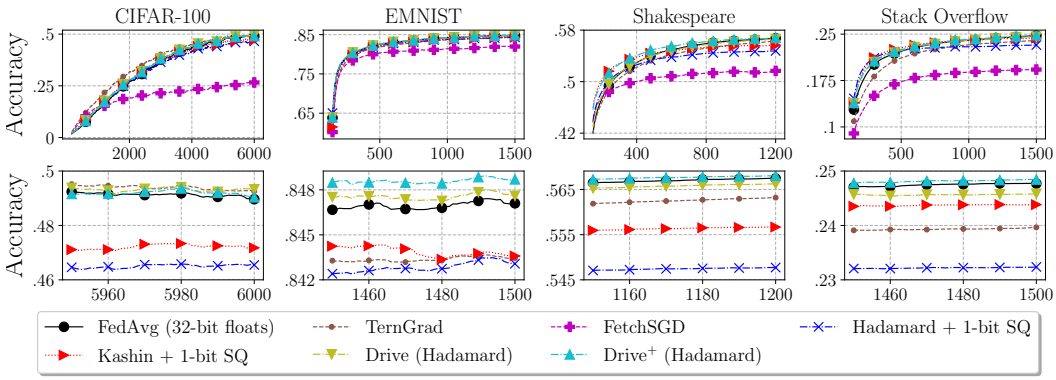

Figure 2: Accuracy per round on various federated learning tasks. Smoothing is done using a rolling mean with a window size of 150. The second row zooms-in on the last 50 rounds.

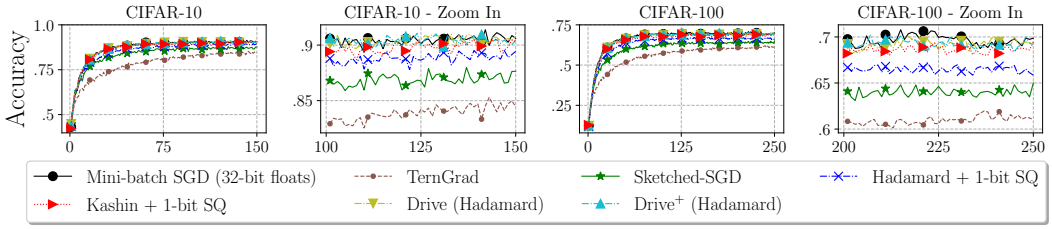

Figure 3: Accuracy per round on distributed learning tasks, with a zoom-in on the last 50 rounds.

width and compute), provides lower accuracy in all tasks. Additional details and hyperparameter configurations are presented in Appendix E.4.

Distributed CNN Training. We evaluate distributed CNN training with 10 clients in two configurations: (1) CIFAR-10 dataset with ResNet-9; (2) CIFAR-100 with ResNet-18 [52, 62]. In both tasks, DRIVE and DRIVE$^+$ have similar accuracy to FedAvg, closely followed by Kashin + 1-bit SQ. The other algorithms are less accurate, with Hadamard + 1-bit SQ being better than Sketched-SGD and TernGrad for both tasks. Additional details and hyperparameter configurations are presented in Appendix E.5. Figure 3 depicts the results.

Evaluation Summary. Overall, it is evident that DRIVE and DRIVE$^+$ consistently offer markedly favorable results in comparison to the alternatives in our setting. Kashin's representation appears to offer the best competition, albeit at somewhat higher computational complexity and bandwidth requirements. The lesser performance of the sketch-based techniques is attributed to the high noise of the sketch under such a low ($d(1+o(1))$ bits) communication requirement. This is because the number of counters they can use is too low, making too many coordinates map into each counter. In Appendix E.6, we also compare DRIVE and DRIVE$^+$ to state of the art techniques over K-Means and Power Iteration tasks for 10, 100, and 1000 clients, yielding similar trends.

Problem	Scale	Rotation	
	S	Uniform	Hadamard
1b - VE	$\frac{\|\mathcal{R}(x)\|_1}{d}$	vNMSE $= \left(1-\frac{2}{\pi}\right)\left(1-\frac{1}{d}\right)$	vNMSE $\leqslant \frac{1}{2}$
1b - DME	$\frac{\|x\|_2^2}{\|\mathcal{R}(x)\|_1}$	$(i)\ d \geqslant 2 \implies$ NMSE $\leqslant \frac{1}{n}\cdot 2.92$ $(ii)\ d \geqslant 135 \implies$ NMSE $\leqslant \frac{1}{n}\cdot\left(\frac{\pi}{2}-1+\sqrt{\frac{(6\pi^3-12\pi^2)\cdot\ln d+1}{d}}\right)$	—

Table 2: Summary of the proven error bounds for DRIVE.

8 Discussion

Proven Error Bounds. We summarize the proven error bounds in Table 2. Since DRIVE (Hadamard) is generally not unbiased (as discussed in Section 6.2), we cannot establish a formal guarantee for the 1b - DME problem when using Hadamard. It is a challenging research question whether there exists other structured rotations with low computational complexity and stronger guarantees.

Input Distribution Assumption. The distributed mean estimation analysis of our Hadamard-based variants is based on an assumption (Section 6.2) about the vector distributions. While machine learning workloads, and DNN gradients in particular (e.g., [33, 34, 35]), were observed to follow such distributions, this assumption may not hold for other applications.

For such cases, we note that DRIVE is compatible with the error feedback (EF) mechanism [63, 64] that ensured convergence and recovery of the convergence rate of non-compressed SGD. Specifically, as evident by Lemma 3, any scale $\frac{\|\mathcal{R}(x)\|_1}{d} \leqslant S \leqslant 2 \cdot \frac{\|\mathcal{R}(x)\|_1}{d}$ is sufficient to respect the *compressor* (i.e., *bounded variance*) assumption. For completeness, in Appendix E.2, we perform EF experiments comparing DRIVE and DRIVE$^+$ to other compression techniques that use EF.

Varying Communication Budget. Unlike some previous works, our algorithms' guarantees with more than one bit per coordinate are not established. It is thus an interesting future work to extend DRIVE to other communication budgets and understand what are the resulting guarantees. We refer the reader to [65] for initial steps towards that direction.

Entropy Encoding. Entropy encoding methods (such as Huffman coding) can further compress vectors of values when the values are not uniformly distributed. We have compared DRIVE against stochastic quantization methods using entropy encoding for the challenging setting for DRIVE where all vectors are the same (see Table 1 for further description). The results appear in Appendix E.7, where DRIVE still outperforms these methods. We also note that, when computation allows and when using DRIVE with multiple bits per entry, DRIVE can also be enhanced by entropy encoding techniques. We describe some initial results for this setting in [65].

Structured Data. When the data is highly sparse, skewed, or otherwise structured, one can leverage that for compression. We note that some techniques that exploit sparsity or structure can be use in conjunction with our techniques. For example, one may transmit only non-zero entries or Top-K entries while compressing these using DRIVE to reduce communication overhead even further.

Compatibility With Distributed All-Reduce Techniques. Quantization techniques, including DRIVE, may introduce overheads in the context of All-Reduce (depending on the network architecture and communication patterns). In particular, if every node in a cluster uses a different rotation, DRIVE will not allow for efficient in-path aggregation without decoding the vectors. Further, the computational overhead of the receiver increases by a $\log d$ factor as each vector has to be decoded separately before an average can be computed. It is an interesting future direction for DRIVE to understand how to minimize such potential overheads. For example, one can consider bucketizing co-located workers and apply DRIVE's quantization only for cross-rack traffic.

9 Conclusions

In this paper, we studied the vector and distributed mean estimation problems. These problems are applicable to distributed and federated learning, where clients communicate real-valued vectors (e.g., gradients) to a server for averaging. To the best of our knowledge, our algorithms are the first with a provable error of $O(\frac{1}{n})$ for the 1b - Distributed Mean Estimation problem (i.e., with $d(1 + o(1))$ bits). As shown in [14], any algorithm that uses $O(d)$ shared random bits (e.g., our Hadamard-based variant) has a vNMSE of $\Omega(1)$, i.e., DRIVE and DRIVE$^+$ are asymptotically optimal; additional discussion is given in Appendix F. Our experiments, carried over various tasks and datasets, indicate that our algorithms improve over the state of the art. All the results presented in this paper are fully reproducible by our source code, available at [27].

Acknowledgments and Disclosure of Funding

MM was supported in part by NSF grants CCF-2101140, CCF-2107078, CCF-1563710, and DMS-2023528. MM and RBB were supported in part by a gift to the Center for Research on Computation and Society at Harvard University. AP was supported in part by the Cyber Security Research Center at Ben-Gurion University of the Negev. We thank Moshe Gabel, Mahmood Sharif, Yuval Filmus and, the anonymous reviewers for helpful comments and suggestions.

References

[1] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[2] Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan. Distributed Mean Estimation With Limited Communication. In *International Conference on Machine Learning*, pages 3329–3337. PMLR, 2017.

[3] Yuval Alfassi, Moshe Gabel, Gal Yehuda, and Daniel Keren. A Distance-Based Scheme for Reducing Bandwidth in Distributed Geometric Monitoring. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021.

[4] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. Large Scale Distributed Deep Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[5] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[6] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Ran Ben-Basat, Michael Mitzenmacher, and Shay Vargaftik. How to Send a Real Number Using a Single Bit (and Some Shared Randomness). *arXiv preprint arXiv:2010.02331*, 2020.

[8] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. In *Advances in neural information processing systems*, pages 1509–1519, 2017.

[9] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient Sgd via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

[10] Jakub Konečnỳ and Peter Richtárik. Randomized Distributed Mean Estimation: Accuracy vs. Communication. *Frontiers in Applied Mathematics and Statistics*, 4:62, 2018.

[11] Sebastian Caldas, Jakub Konečný, H Brendan McMahan, and Ameet Talwalkar. Expanding the Reach of Federated Learning by Reducing Client Resource Requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[12] Youhui Bai, Cheng Li, Quan Zhou, Jun Yi, Ping Gong, Feng Yan, Ruichuan Chen, and Yinlong Xu. Gradient compression supercharged high-performance data parallel dnn training. In *The 28th ACM Symposium on Operating Systems Principles (SOSP 2021)*, 2021.

[13] Yurii Lyubarskii and Roman Vershynin. Uncertainty Principles and Vector Quantization. *IEEE Transactions on Information Theory*, 56(7):3491–3501, 2010.

[14] Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor. *arXiv preprint arXiv:2002.08958*, 2020.

[15] Ilan Newman. Private vs. Common Random Bits in Communication Complexity. *Information processing letters*, 39(2):67–71, 1991.

[16] Francesco Mezzadri. How to Generate Random Matrices From the Classical Compact Groups. *arXiv preprint math-ph/0609050*, 2006.

[17] RWM Wedderburn. Generating Random Rotations. *Research Report*, 1975.

[18] Richard M Heiberger. Generation of Random Orthogonal Matrices. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 27(2):199–206, 1978.

[19] Gilbert W Stewart. The Efficient Generation of Random Orthogonal Matrices With an Application to Condition Estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

[20] Martin A Tanner and Ronald A Thisted. Remark as r42: A Remark on as 127. Generation of Random Orthogonal Matrices. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2):190–192, 1982.

[21] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

[22] Allan Grønlund, Kasper Green Larsen, Alexander Mathiasen, Jesper Sindahl Nielsen, Stefan Schneider, and Mingzhou Song. Fast Exact K-Means, K-Medians and Bregman Divergence Clustering in 1D. *arXiv preprint arXiv:1701.07204*, 2017.

[23] Mario Lezcano Casado. GeoTorch's Documentation. A Library for Constrained Optimization and Manifold Optimization for Deep Learning in Pytorch. `https://geotorch.readthedocs.io/en/latest/_modules/geotorch/so.html`, 2021. accessed 17-May-21.

[24] Nir Ailon and Bernard Chazelle. The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[25] Bernard J. Fino and V. Ralph Algazi. Unified matrix treatment of the fast walsh-hadamard transform. *IEEE Transactions on Computers*, 25(11):1142–1146, 1976.

[26] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. Code available at: `https://github.com/uber-research/intrinsic-dimension`.

[27] Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. DRIVE open source code. `https://github.com/amitport/DRIVE-One-bit-Distributed-Mean-Estimation`, 2021.

[28] Kathy J Horadam. *Hadamard Matrices and Their Applications*. Princeton university press, 2012.

[29] Aleksandr Khintchine. Über Dyadische Brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

[30] S Szarek. On the Best Constants in the Khinchin Inequality. *Studia Mathematica*, 2(58):197–208, 1976.

[31] Yuval Filmus. Khintchine-Kahane using Fourier Analysis. *Posted at `http://www.cs.toronto.edu/yuvalf/KK.pdf`*, 2012.

[32] Rafał Latała and Krzysztof Oleszkiewicz. On the Best Constant In the Khinchin-Kahane Inequality. *Studia Mathematica*, 109(1):101–104, 1994.

[33] Brian Chmiel, Liad Ben-Uri, Moran Shkolnik, Elad Hoffer, Ron Banner, and Daniel Soudry. Neural Gradients Are Near-Lognormal: Improved Quantized and Sparse Training. *arXiv preprint arXiv:2006.08173*, 2020.

[34] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post-Training 4-Bit Quantization of Convolution Networks for Rapid-Deployment. *arXiv preprint arXiv:1810.05723*, 2018.

[35] Xucheng Ye, Pengcheng Dai, Junyu Luo, Xin Guo, Yingjie Qi, Jianlei Yang, and Yiran Chen. Accelerating CNN Training by Pruning Activation Gradients. In *European Conference on Computer Vision*, pages 322–338. Springer, 2020.

[36] Marcus Spruill et al. Asymptotic Distribution of Coordinates on High Dimensional Spheres. *Electronic communications in probability*, 12:234–247, 2007.

[37] Persi Diaconis and David Freedman. A Dozen de Finetti-Style Results in Search of a Theory. In *Annales de l'IHP Probabilités et statistiques*, volume 23, pages 397–423, 1987.

[38] Svetlozar T Rachev, L Ruschendorf, et al. Approximate Independence of Distributions on Spheres and Their Stability Properties. *The Annals of Probability*, 19(3):1311–1337, 1991.

[39] Adriaan J Stam. Limit Theorems for Uniform Distributions on Spheres in High-Dimensional Euclidean Spaces. *Journal of Applied probability*, 19(1):221–228, 1982.

[40] Charles M Rader. A New Method of Generating Gaussian Random Variables by Computer. Technical report, Massachusetts Institute of Technology, Lincoln Lab, 1969.

[41] David B Thomas. Parallel Generation of Gaussian Random Numbers Using the Table-Hadamard Transform. In *2013 IEEE 21st Annual International Symposium on Field-Programmable Custom Computing Machines*, pages 161–168. IEEE, 2013.

[42] Tamás Herendi, Thomas Siegl, and Robert F Tichy. Fast Gaussian Random Number Generation Using Linear Transformations. *Computing*, 59(2):163–181, 1997.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019.

[44] Google. TensorFlow Federated: Machine Learning on Decentralized Data. `https://www.tensorflow.org/federated`, 2020. accessed 25-Mar-20.

[45] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency, 2017.

[46] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary B. Charles, Dimitris S. Papailiopoulos, and Stephen Wright. ATOMO: Communication-efficient Learning via Atomic Sparsification. In *NeurIPS*, pages 9872–9883, 2018.

[47] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[48] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with Memory. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[49] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[50] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[51] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to Handwritten Letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[52] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features From Tiny Images. 2009.

[53] William Shakespeare. The Complete Works of William Shakespeare. `https://www.gutenberg.org/ebooks/100`.

[54] Stack Overflow Data. `https://www.kaggle.com/stackoverflow/stackoverflow`. accessed 01-Mar-21.

[55] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-Efficient Distributed Sgd With Sketching. *arXiv preprint arXiv:1903.04488*, 2019.

[56] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. FetchSGD: Communication-Efficient Federated Learning With Sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR, 2020.

[57] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[58] The TensorFlow Authors. TensorFlow Federated: Compression via Kashin's representation from Hadamard transform. `https://github.com/tensorflow/model-optimization/blob/9193d70f6e7c9f78f7c63336bd68620c4bc6c2ca/tensorflow_model_optimization/python/core/internal/tensor_encoding/stages/research/kashin.py#L92`. accessed 16-May-21.

[59] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings, 2019.

[60] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive Federated Optimization, 2020.

[61] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9:1735–1780, 1997.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[63] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-Bit Stochastic Gradient Descent and Its Application to Data-Parallel Distributed Training of Speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[64] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *International Conference on Machine Learning*, pages 3252–3261, 2019.

[65] Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Communication-efficient federated learning via robust distributed mean estimation. *arXiv preprint arXiv:2108.08842*, 2021.

[66] Mervin E Muller. A Note on a Method for Generating Points Uniformly on N-Dimensional Spheres. *Communications of the ACM*, 2(4):19–20, 1959.

[67] George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[68] Andreas Maurer et al. A Bound on the Deviation Probability for Sums of Non-Negative Random Variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):15, 2003.

[69] Paweł J Szabłowski. Uniform Distributions on Spheres in Finite Dimensional L_α and Their Generalizations. *Journal of multivariate analysis*, 64(2):103–117, 1998.

[70] Carl-Gustav Esseen. A Moment Inequality With an Application to the Central Limit Theorem. *Scandinavian Actuarial Journal*, 1956(2):160–170, 1956.

[71] Google. TensorFlow Model Optimization Toolkit. `https://www.tensorflow.org/model_optimization`, 2021. accessed 19-Mar-21.

[72] Apache license, version 2.0. `https://www.apache.org/licenses/LICENSE-2.0`. accessed 19-Mar-21.

[73] Pytorch license. `https://github.com/pytorch/pytorch/blob/aaccdc39965ade4b61b7852329739e777e244c25/LICENSE`. accessed 19-Mar-21.

[74] Creative Commons Attribution-Share Alike 3.0 license. `https://creativecommons.org/licenses/by-sa/3.0/`. accessed 01-Mar-21.

[75] The TensorFlow Authors. TensorFlow Federated: EncodedSumFactory class. `https://github.com/tensorflow/federated/blob/v0.19.0/tensorflow_federated/python/aggregators/encoded.py#L40-L142`. accessed 19-May-21.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 9.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not believe there is an inherit negative societal impact in this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the Supplemental Material. It will be released as open-source after publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Section 7 and Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Most experiments can be easily reproduced but require considerable execution time and compute resources. We did include error bars in Figure 1, where it was feasible.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 7 and Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 7 and Appendix E.
   (b) Did you mention the license of the assets? [Yes] See Section 7 and Appendix E.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the source code of our algorithms in the Supplemental Material. It will be released as open-source after publication.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We used well-known datasets whose creators use either publicly available data or received proper consent.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We used well-known datasets, which do not contain personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or conducted research with human subjects.

A Missing Proofs for DRIVE

A.1 Proof of Theorem 2

Theorem 2. *For any $x \in \mathbb{R}^d$, the vNMSE of DRIVE with $S = \frac{\|\mathcal{R}_U(x)\|_1}{d}$ is $\left(1-\frac{2}{\pi}\right)\left(1-\frac{1}{d}\right)$.*

Proof. Let $T \in \mathcal{S}^{d-1}$ be uniformly distributed on the unit sphere. By the definition of $\mathcal{R}_U$, for *any* $x \in \mathbb{R}^d$, $\mathcal{R}_U(\breve{x})$ and T follow the same distribution (i.e., $\mathcal{R}_U(\breve{x})$ is also uniformly distributed on the unit sphere). As was established in [66], a uniformly distributed point on a sphere $T \in \mathcal{S}^{d-1}$ can be obtained by first deriving a random vector $Z = (Z_1, \ldots, Z_d)$ where $Z_i \sim N(0,1), i \in [1, \ldots, d]$ are normally distributed i.i.d. random variables, and then normalizing its norm by $T = \frac{Z}{\|Z\|_2}$. We also make use of the fact that the norm and direction of a standard multivariate normal vector are independent. That is, T and $\|Z\|_2$ are independent. Therefore, we have that, $\|T\|_1^2 = \left\|\frac{Z}{\|Z\|_2}\right\|_1^2 = \frac{1}{\|Z\|_2^2} \cdot \|Z\|_1^2$ and thus $\|T\|_1^2 \cdot \|Z\|_2^2 = \|Z\|_1^2$. Taking expectation and rearranging yields,

$$\mathbb{E}\left[\|T\|_1^2\right] = \frac{\mathbb{E}\left[\|Z\|_1^2\right]}{\mathbb{E}\left[\|Z\|_2^2\right]} = \frac{\mathbb{E}\left[\sum_{i=1}^d Z_i^2 + \sum_{(i,j)\in\{d\times d\}, i\neq j} |Z_i| \cdot |Z_j|\right]}{d}$$

$$= \frac{\sum_{i=1}^d \mathbb{E}\left[Z_i^2\right] + \sum_{(i,j)\in\{d\times d\}, i\neq j} \mathbb{E}\left[|Z_i|\right] \cdot \mathbb{E}\left[|Z_j|\right]}{d} = \frac{d + d\cdot(d-1)\cdot(\sqrt{\frac{2}{\pi}})^2}{d} = 1 + (d-1)\frac{2}{\pi}\,.$$

Here we employed $\mathbb{E}\left[Z_i^2\right] = 1$ and that $|Z_i|$ and $|Z_j|$ are independent half-normal random variables for $i \neq j$ with $\mathbb{E}\left[|Z_i|\right] = \mathbb{E}\left[|Z_j|\right] = \sqrt{\frac{2}{\pi}}$. Finally, the resulting vNMSE depends only on the dimension and not the specific vector x. In particular, $1 - \mathbb{E}\left[\mathcal{L}^d_{\mathcal{R},x}\right] = 1 - \frac{1+(d-1)\frac{2}{\pi}}{d} = \left(1-\frac{2}{\pi}\right)\left(1-\frac{1}{d}\right)$. □

A.2 Proof of Theorem 3

Theorem 3. *For any $x \in \mathbb{R}^d$, set $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$. Then DRIVE satisfies $\mathbb{E}[\widehat{x}] = x$.*

Proof. For any $x \in \mathbb{R}^d$ denote $x' = (\|x\|_2, 0, \ldots, 0)^T$ and let $R_{x\to x'} \in \mathbb{R}^{d\times d}$ be a rotation matrix such that $R_{x\to x'} \cdot x = x'$. Further, denote $R_x = R_U R^{-1}_{x\to x'}$.

Using these definitions and observing that $\|\mathcal{R}_U(x)\|_1 = \langle R_U \cdot x,\ \text{sign}(R_U \cdot x)\rangle$ we obtain:

$$\begin{aligned}
\widehat{x} &= R^{-1}_{x\to x'} \cdot R_{x\to x'} \cdot \widehat{x} = R^{-1}_{x\to x'} \cdot R_{x\to x'} \cdot R_U^{-1} \cdot S \cdot \text{sign}\left(R_U \cdot x\right) \\
&= R^{-1}_{x\to x'} \cdot R_x^{-1} \cdot S \cdot \text{sign}\left(R_x \cdot R_{x\to x'} \cdot x\right) = S \cdot R^{-1}_{x\to x'} \cdot R_x^{-1} \cdot \text{sign}\left(R_x \cdot x'\right) \\
&= \frac{\|x\|_2^2}{\langle R_U \cdot x,\ \text{sign}(R_U \cdot x)\rangle} \cdot R^{-1}_{x\to x'} \cdot R_x^{-1} \cdot \text{sign}\left(R_x \cdot x'\right) \qquad (4) \\
&= R^{-1}_{x\to x'} \cdot \|x\|_2^2 \cdot \frac{R_x^{-1} \cdot \text{sign}\left(R_x \cdot x'\right)}{\langle R_x \cdot x',\ \text{sign}(R_x \cdot x')\rangle}\,.
\end{aligned}$$

Next, let C_i be a vector containing the values of the i'th column of R_x. Then, $R_x \cdot x' = \|x\|_2 \cdot C_0$ and $\text{sign}(R_x \cdot x') = \text{sign}(\|x\|_2 \cdot C_0) = \text{sign}(C_0)$. This means that,

$$\left\langle R_x \cdot x',\ \text{sign}(R_x \cdot x')\right\rangle = \|x\|_2 \cdot \langle C_0,\ \text{sign}(C_0)\rangle = \|x\|_2 \cdot \|C_0\|_1\,. \qquad (5)$$

Now, since $R_x^{-1} = R_x^T$, we have that,

$$R_x^{-1} \cdot \text{sign}\left(R_x \cdot x'\right) = \left(\|C_0\|_1, \langle C_1, \text{sign}(C_0)\rangle, \ldots, \langle C_{d-1}, \text{sign}(C_0)\rangle\right)^T. \qquad (6)$$

Using Eq. (5) and (6) in Eq. (4) yields

$$\widehat{x} = R^{-1}_{x\to x'} \cdot \|x\|_2 \cdot \left(1, \frac{\langle C_1, \text{sign}(C_0)\rangle}{\|C_0\|_1}, \ldots, \frac{\langle C_{d-1}, \text{sign}(C_0)\rangle}{\|C_0\|_1}\right)^T. \qquad (7)$$

Next, by drawing insight from (7), we show that the estimate of DRIVE is unbiased by constructing a symmetric algorithm to DRIVE whose reconstruction's expected value is identical to that of DRIVE, but with the additional property that the average of both reconstructions is exactly x for each specific run of the algorithm.

In particular, consider an algorithm DRIVE$'$ that operates exactly as DRIVE but, instead of directly using the sampled rotation matrix $R_U = R_x \cdot R^{-1}_{x\to x'}$ it calculates and uses the rotation matrix $R'_U = R_x \cdot I' \cdot R^{-1}_{x\to x'}$ where I' is identical to the d-dimensional identity matrix with the exception that $I'_{00} = -1$ instead of 1.

Now, since $R_U \sim \mathcal{R}_U$ and $R_{x\to x'}$ is a fixed rotation matrix, we have that $R_x = R_U \cdot R_{x\to x'} \sim \mathcal{R}_U$. In turn, this also means that $R_x \cdot I' \sim \mathcal{R}_U$ since I' is a fixed rotation matrix.

Consider a run of both algorithm where $\widehat{x}$ is the reconstruction of DRIVE for x with a sampled rotation R_U and $\widehat{x}'$ is the corresponding reconstruction of DRIVE$'$ for x with the rotation R'_U.

According to (7) it holds that: $\widehat{x} + \widehat{x}' = R^{-1}_{x\to x'} \cdot \|x\|_2 \cdot (2, 0, \ldots, 0)^T = 2 \cdot x$. This is because both runs are identical except that the first column of R_x (i.e., C_0) and $R_x \cdot I'$ have opposite signs. In particular, it holds that $\mathbb{E}\left[\widehat{x} + \widehat{x}'\right] = 2 \cdot x$. But, since $R_x \sim \mathcal{R}_U$ and $R_x \cdot I' \sim \mathcal{R}_U$, both algorithms have the same expected value. This yields $\mathbb{E}\left[\widehat{x}\right] = \mathbb{E}\left[\widehat{x}'\right] = x$. This concludes the proof. □

A.3 Proof of Theorem 4

The proof of the theorem follows from the following two lemmas. Lemma 6 relies on Chebyshev's inequality and allows us to bound the vNMSE for all d. Lemma 7 gives a sharper bound for large dimensions using the Bernstein's inequality.

Lemma 6. *For any constant k and dimension $d \geqslant 2$ such that $0 < k < \frac{d}{d-1} \cdot \sqrt{\frac{\pi}{\pi-3}} \cdot \frac{2}{\mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}$,*

$$\mathbb{E}\left[\frac{d}{\|\mathcal{R}(\breve{x})\|_1^2}\right] \leqslant \frac{d}{k^2} + \frac{\left(1 - \frac{1}{k^2}\right)}{\left(-k \cdot \sqrt{\frac{\pi-3}{\pi d}} + \frac{2\cdot\sqrt{d}}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}\right)^2}.$$

Proof. By Lemma 9 in Appendix A.4 we have that $\mathbb{E}\left[\|R_U \cdot \breve{x}\|_1\right] = \frac{2d}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}$. Thus,

$$\begin{aligned}\mathrm{Var}\left(\frac{\|R_U \cdot \breve{x}\|_1}{\sqrt{d}}\right) &= \mathbb{E}\left[\frac{\|R_U \cdot \breve{x}\|_1^2}{d}\right] - \left(\mathbb{E}\left[\frac{\|R_U \cdot \breve{x}\|_1}{\sqrt{d}}\right]\right)^2 \\ &= \frac{2}{\pi} + \frac{1 - \frac{2}{\pi}}{d} - \left(\frac{2 \cdot \sqrt{d}}{(d-1) \cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}\right)^2 \leqslant \frac{\pi - 3}{\pi d}.\end{aligned}$$

According to Chebyshev's inequality, $\Pr\left[-\frac{\|R_U\cdot\breve{x}\|_1}{\sqrt{d}} + \frac{2\cdot\sqrt{d}}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})} \geqslant k \cdot \sqrt{\frac{\pi-3}{\pi d}}\right] \leqslant \frac{1}{k^2}$ which is equivalent to $\Pr\left[\frac{\|R_U\cdot\breve{x}\|_1}{\sqrt{d}} \leqslant -k \cdot \sqrt{\frac{\pi-3}{\pi d}} + \frac{2\cdot\sqrt{d}}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}\right] \leqslant \frac{1}{k^2}$. Now, for any k that respects $-k \cdot \sqrt{\frac{\pi-3}{\pi d}} + \frac{2\cdot\sqrt{d}}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})} > 0$ it holds that $\Pr\left[\frac{d}{\|R_U\cdot\breve{x}\|_1^2} \geqslant \frac{1}{\left(-k\cdot\sqrt{\frac{\pi-3}{\pi d}} + \frac{2\cdot\sqrt{d}}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}\right)^2}\right] \leqslant \frac{1}{k^2}$.

In addition, observe that $\Pr\left[\frac{d}{\|R_U\cdot\breve{x}\|_1^2} > d\right] = 0$. This means that,

$$\mathbb{E}\left[\frac{d}{\|R_U \cdot \breve{x}\|_1^2}\right] \leqslant \frac{d}{k^2} + \left(1 - \frac{1}{k^2}\right) \cdot \left(-k \cdot \sqrt{\frac{\pi - 3}{\pi d}} + \frac{2 \cdot \sqrt{d}}{(d-1) \cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}\right)^{-2}.$$

This concludes the proof. □

Lemma 7. *For any vector $x \in \mathbb{R}^d$, a constant $0 < \epsilon < 1$, and any dimension $d \geqslant 2$, it holds that*

$$\mathbb{E}\left[\frac{d}{\|\mathcal{R}(\breve{x})\|_1^2}\right] \leqslant \frac{\pi}{2} \cdot (1 + \epsilon) + d \cdot exp\left(-d \cdot \frac{\epsilon^2}{16(\pi - 2)}\right).$$

Proof. By the definition of a uniform random rotation, it holds that $\frac{d}{\|\mathcal{R}(\check{x})\|_1^2}$ follows the same distribution as $\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}$ where $Z = (Z_1, \ldots, Z_d)$ and $Z_i \sim \mathcal{N}(0,1)$ are i.i.d. normal variables. Consider the two complementary events, $A_- = \|Z\|_1 \leqslant \beta \cdot d$ and $A_+ = \|Z\|_1 > \beta \cdot d$. Then,

$$\mathbb{E}\left[\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] = \mathbb{E}\left[\mathbb{1}_{A_+}\cdot\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] + \mathbb{E}\left[\mathbb{1}_{A_-}\cdot\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right].$$

We now treat each of the terms separately. For the first term we have,

$$\mathbb{E}\left[\mathbb{1}_{A_+}\cdot\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] \leqslant \frac{d}{(d\cdot\beta)^2}\cdot\mathbb{E}\left[\|Z\|_2^2\right] = \frac{1}{\beta^2}.$$

For the second term we have,

$$\mathbb{E}\left[\mathbb{1}_{A_-}\cdot\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] \leqslant d\cdot\mathbb{E}\left[\mathbb{1}_{A_-}\right] = d\cdot\Pr\left[\|Z\|_1 \leqslant \beta\cdot d\right].$$

Now, our goal is to bound the term $\Pr\left[\|Z\|_1 \leqslant \beta\cdot d\right]$. Denote $Y = (Y_1, \ldots, Y_d)$, where $Y_i = \sqrt{\frac{2}{\pi}} - |Z_i|$. Then, it holds that,

$$\Pr\left[\|Z\|_1 \leqslant \beta d\right] = \Pr\left[\|Z\|_1 - d\sqrt{\frac{2}{\pi}} \leqslant \beta d - d\sqrt{\frac{2}{\pi}}\right] = \Pr\left[\sum_{i=1}^{d} Y_i \geqslant d\left(\sqrt{\frac{2}{\pi}} - \beta\right)\right].$$

Further, for all i we have that $\mathbb{E}\left[Y_i\right] = 0$, $\mathbb{E}\left[Y_i\right]^2 = 1 - \frac{2}{\pi}$ and $Y_i \leqslant \sqrt{\frac{2}{\pi}}$. According to Bernstein's Inequality (from [67], or see [68, Theorem 3.1] for a simplified notation) this means that,

$$\Pr\left[\sum_{i=1}^{d} Y_i \geqslant d\left(\sqrt{\frac{2}{\pi}} - \beta\right)\right] \leqslant exp\left(\frac{-d^2\cdot(\sqrt{\frac{2}{\pi}} - \beta)^2}{2d\cdot(1-\frac{2}{\pi}) + d\cdot\frac{2}{3}\cdot\sqrt{\frac{2}{\pi}}\cdot\left(\sqrt{\frac{2}{\pi}} - \beta\right)}\right).$$

Setting $\beta = \sqrt{\frac{2-\epsilon}{\pi}}$ for some $\epsilon > 0$ yields,

$$\Pr\left[\sum_{i=1}^{d} Y_i \geqslant d\left(\sqrt{\frac{2}{\pi}} - \beta\right)\right] \leqslant exp\left(-d\cdot\frac{3\cdot\left(1-\sqrt{1-\frac{\epsilon}{2}}\right)^2}{3\pi - 4 - 2\sqrt{1-\frac{\epsilon}{2}}}\right).$$

Next, we obtain,

$$\mathbb{E}\left[\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] \leqslant \frac{\pi}{2}\cdot\frac{1}{1-\frac{\epsilon}{2}} + d\cdot exp\left(-d\cdot\frac{3\cdot\left(1-\sqrt{1-\frac{\epsilon}{2}}\right)^2}{3\pi - 4 - 2\sqrt{1-\frac{\epsilon}{2}}}\right).$$

We next use Taylor expansions to simplify the expression:

$$\begin{aligned}
\mathbb{E}\left[\frac{d\cdot\|Z\|_2^2}{\|Z\|_1^2}\right] &\leqslant \frac{\pi}{2}\cdot\frac{1}{1-\frac{\epsilon}{2}} + d\cdot exp\left(-d\cdot\frac{3\cdot\left(1-\sqrt{1-\frac{\epsilon}{2}}\right)^2}{3\pi - 4 - 2\sqrt{1-\frac{\epsilon}{2}}}\right) \\
&\leqslant \frac{\pi}{2}\cdot\frac{1}{1-\frac{\epsilon}{2}} + d\cdot exp\left(-d\cdot\frac{\left(\frac{\epsilon}{2}\right)^2}{4(\pi-2)}\right) \\
&\leqslant \frac{\pi}{2}\cdot\left(1 + 2\cdot\frac{\epsilon}{2}\right) + d\cdot exp\left(-d\cdot\frac{\left(\frac{\epsilon}{2}\right)^2}{4(\pi-2)}\right) \\
&= \frac{\pi}{2}\cdot(1+\epsilon) + d\cdot exp\left(-d\cdot\frac{\epsilon^2}{16(\pi-2)}\right).
\end{aligned}$$

This concludes the proof. □

Finally, we prove the theorem, which we restate here.

Theorem 4. *For any $x \in \mathbb{R}^d$, the vNMSE of DRIVE with $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$ satisfies:*
(i) For all $d \geqslant 2$, it is at most 2.92. (ii) For all $d \geqslant 135$, it is at most $\frac{\pi}{2} - 1 + \sqrt{\frac{(6\pi^3 - 12\pi^2)\cdot \ln d + 1}{d}}$.

Proof. Recall that, using Lemma 2, the vNMSE is $\mathbb{E}\left[\frac{d}{\|\mathcal{R}(\breve{x})\|_1^2}\right] - 1$. Therefore, the first part of the theorem follows from Lemma 6 with $k = \sqrt{d}$, which satisfies $\sqrt{k} < \frac{d}{d-1} \cdot \sqrt{\frac{\pi}{\pi - 3}} \cdot \frac{2}{\mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}$, as needed. The second part of the theorem follows from setting $\epsilon = \sqrt{\frac{(24\pi - 48)\ln d}{d}}$ in Lemma 7. Notice that for $d \geqslant 135$, we get $\epsilon < 1$, as required. Then:

$$\mathbb{E}\left[\frac{d}{\|\mathcal{R}(\breve{x})\|_1^2}\right] \leqslant \frac{\pi}{2} \cdot (1 + \epsilon) + d \cdot exp\left(-1.5 \ln d\right) = \frac{\pi}{2} + \sqrt{\frac{(6\pi^3 - 12\pi^2) \cdot \ln d + 1}{d}} \,.$$

This concludes the proof. □

A.4 Constant Scale for Unbiased Estimates in DRIVE with Uniform Random Rotations

In this appendix, we prove that it is possible to further lower the vNMSE while remaining unbiased by (deterministically) using $S = \frac{\|x\|_2^2}{\mathbb{E}\left[\|\mathcal{R}_U(x)\|_1\right]} = \frac{\|x\|_2}{\mathbb{E}\left[\|T\|_1\right]}$ where $T \in \mathcal{S}^{d-1}$ is uniformly at random distributed on the unit sphere (observe that the unbiasedness proof in Theorem A.2 holds for this constant scale). This is because $\frac{d}{\left(\mathbb{E}\left[\|T\|_1\right]\right)^2} - 1 \leqslant \mathbb{E}\left[\frac{d}{\|T\|_1^2}\right] - 1$. This gives a provable $\frac{\pi}{2} - 1$ vNMSE bound for any d. We start the analysis by proving two auxiliary lemmas. A generalization of these results is proved in [69, Eq. 3.3]. For completeness, we provide simplified proofs.

Lemma 8. *Let $T = (T_1, \ldots, T_d) \in \mathcal{S}^{d-1}$ be a point on the unit sphere, drawn uniformly at random. Then, the PDF of $|T_i|$ is given by,*

$$f_{|T_i|}(t) = \frac{2 \cdot (1 - t^2)^{\frac{d-1}{2} - 1}}{\mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}.$$

Proof. As was established in [66], a uniformly random point on a sphere T can be obtained by first deriving a random vector $Z = (Z_1, \ldots, Z_d)$ where $Z_i \sim N(0, 1), i \in [1, \ldots, d]$ are normally distributed i.i.d. random variables, and then normalizing its norm by $T = \frac{Z}{\|Z\|_2}$.

Therefore, the i'th coordinate of T is given as $T_i = \frac{Z_i}{\|Z\|_2}$. Since the coordinate distribution is sign-symmetric, we focus on its distribution in the interval $[0, 1]$. Observe that for $T_i \in [0, 1]$:

$$\begin{aligned}
\Pr\left[|T_i| \leqslant t\right] = \Pr\left[T_i^2 \leqslant t^2\right] = \Pr\left[\frac{Z_i^2}{\|Z\|_2^2} \leqslant t^2\right] = \Pr\left[\frac{Z_i^2}{\|Z\|_2^2 - Z_i^2} \leqslant t^2 \cdot \frac{\|Z\|_2^2}{\|Z\|_2^2 - Z_i^2}\right] \\
= \Pr\left[\frac{Z_i^2}{\|Z\|_2^2 - Z_i^2} \leqslant t^2 \cdot \left(1 + \frac{Z_i^2}{\|Z\|_2^2 - Z_i^2}\right)\right] = \Pr\left[\frac{Z_i^2}{\|Z\|_2^2 - Z_i^2} \leqslant \frac{t^2}{1 - t^2}\right].
\end{aligned}$$

Therefore $\Pr\left[|T_i| \leqslant t\right] = \Pr\left[\frac{(d-1)Z_i^2}{\|Z\|_2^2 - Z_i^2} \leqslant \frac{(d-1)t^2}{1-t^2}\right]$. Notice that $Z_i^2 \sim \chi^2(1)$ and $(\|Z\|_2^2 - Z_i^2) \sim \chi^2(d-1)$ are independent χ^2 random variables. Now denote $Y \triangleq \frac{(d-1)Z_i^2}{\|Z\|_2 - Z_i^2}$. Observe that Y follows a $F_{1,d-1}$ distribution. Therefore, the CDF of Y is then given by the following expression: $\Pr\left[Y \leqslant y\right] = I_{\frac{y}{y+d-1}}(\frac{1}{2}, \frac{d-1}{2})$, where I is the regularized incomplete Beta function. Substituting $y = \frac{(d-1)t^2}{1-t^2}$ yields $\Pr\left[|T_i| \leqslant t\right] = I_{t^2}(\frac{1}{2}, \frac{d-1}{2})$. In particular, taking the derivative yields $f_{|T_i|}(t) = \frac{2 \cdot (1-t^2)^{\frac{d-1}{2} - 1}}{\mathrm{B}(\frac{1}{2}, \frac{d-1}{2})}$. Interestingly, this PDF is similar to the Beta distribution (not to be confused with the B function). Indeed, one can verify that $T_i' = \frac{T_i + 1}{2}$ follows a Beta distribution with $T_i' \sim Beta(\frac{d-1}{2}, \frac{d-1}{2})$. □

With Lemma 8 at hand, we obtain the following result.

Lemma 9. *Let $T \in \mathcal{S}^{d-1}$ be a point on the unit sphere, drawn uniformly at random. Then,*

$$\mathbb{E}\left[\|T\|_1\right] = \frac{2d}{(d-1)\cdot \mathrm{B}(\frac{1}{2}, \frac{d-1}{2})} \cdot$$

Proof. Due to the linearity of expectation we have that $\mathbb{E}\left[\|T\|_1\right] = \sum_{i=1}^{d} \mathbb{E}\left[|T_i|\right]$. Thus, it is sufficient to calculate the expected value of a single element in the sum. By Lemma 8 we have that $\mathbb{E}\left[|T_i|\right] = \int_0^1 t \cdot \frac{2\cdot(1-t^2)^{\frac{d-1}{2}-1}}{\mathrm{B}(\frac{1}{2},\frac{d-1}{2})} dt = \frac{2}{(d-1)\cdot \mathrm{B}(\frac{1}{2},\frac{d-1}{2})}$. Summing over $i \in \{1,\ldots,d\}$] yields the result. □

We note that $\mathbb{E}\left[\|T\|_1\right] = \frac{2d}{(d-1)\cdot \mathrm{B}(\frac{1}{2},\frac{d-1}{2})} \geqslant \sqrt{\frac{2d}{\pi}}$. Finally, recall that our vNMSE is bounded by $\frac{d}{\mathbb{E}\left[\|T\|_1^2\right]} - 1$, which is therefore at most $\frac{\pi}{2} - 1$. In practice, while this deterministic approach gives a lower vNMSE for low dimensions, the benefit is marginal even for d values of several hundreds.

A.5 Proof of Theorem 5

Theorem 5. *Assume n clients, each with its own vector $x_{(\mathfrak{c})} \in \mathbb{R}^d$. Let each client independently sample $R_{U,\mathfrak{c}} \sim \mathcal{R}_U$ and set its scale to $\frac{\|x_{(\mathfrak{c})}\|_2^2}{\|R_{U,\mathfrak{c}} \cdot x_{(\mathfrak{c})}\|_1}$. Then, the server average estimate's NMSE satisfies: $\frac{\mathbb{E}\left[\|x_{avg} - \widehat{x_{avg}}\|_2^2\right]}{\frac{1}{n}\cdot\sum_{\mathfrak{c}=1}^{n}\|x_{(\mathfrak{c})}\|_2^2} = \frac{vNMSE}{n}$, where vNMSE is given by Lemma 2 and is bounded by Theorem 4.*

Proof. We have that,

$$\begin{aligned}
\mathbb{E}\left[\|x_{\text{avg}} - \widehat{x_{\text{avg}}}\|_2^2\right] &= \frac{1}{n^2}\cdot\mathbb{E}\left[\left\|\sum_{\mathfrak{c}=1}^{n} x_{(\mathfrak{c})} - \sum_{\mathfrak{c}=1}^{n}\widehat{x_{(\mathfrak{c})}}\right\|_2^2\right] = \frac{1}{n^2}\cdot\sum_{\mathfrak{c},\mathfrak{c}'}\mathbb{E}\left[\left\langle x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}, x_{(\mathfrak{c}')} - \widehat{x_{(\mathfrak{c}')}}\right\rangle\right] \\
&= \frac{1}{n^2}\cdot\sum_{\mathfrak{c}}\mathbb{E}\left[\left\langle x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}, x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}\right\rangle\right] \\
&\quad + \frac{1}{n^2}\cdot\sum_{\mathfrak{c}\neq\mathfrak{c}'}\mathbb{E}\left[\left\langle x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}, x_{(\mathfrak{c}')} - \widehat{x_{(\mathfrak{c}')}}\right\rangle\right] \\
&\underset{\langle 1\rangle}{=} \frac{1}{n^2}\sum_{\mathfrak{c}}\mathbb{E}\left[\left\|x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}\right\|_2^2\right] = \frac{1}{n^2}\cdot\sum_{\mathfrak{c}}\left\|x_{(\mathfrak{c})}\right\|_2^2\cdot\mathbb{E}\left[\frac{\left\|x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}\right\|_2^2}{\left\|x_{(\mathfrak{c})}\right\|_2^2}\right] \\
&= \frac{1}{n^2}\cdot\sum_{\mathfrak{c}}\left\|x_{(\mathfrak{c})}\right\|_2^2\cdot vNMSE = \frac{vNMSE}{n}\cdot\frac{\sum_{\mathfrak{c}}\left\|x_{(\mathfrak{c})}\right\|_2^2}{n}.
\end{aligned}$$

Here, since the $\mathbb{E}\left[\widehat{x_{(\mathfrak{c})}}\right] = x_{(\mathfrak{c})}$ for all i and since $\widehat{x_{(\mathfrak{c})}}$ and $\widehat{x_{(\mathfrak{c}')}}$ are independent for all $\mathfrak{c} \neq \mathfrak{c}'$, $\langle 1\rangle$ follows from $\mathbb{E}\left[\left\langle x_{(\mathfrak{c})} - \widehat{x_{(\mathfrak{c})}}, x_{(\mathfrak{c}')} - \widehat{x_{(\mathfrak{c}')}}\right\rangle\right] = \mathbb{E}\left[\left\langle x_{(\mathfrak{c})}, x_{(\mathfrak{c}')}\right\rangle\right] - \mathbb{E}\left[\left\langle x_{(\mathfrak{c})}, \widehat{x_{(\mathfrak{c}')}}\right\rangle\right] - \mathbb{E}\left[\left\langle \widehat{x_{(\mathfrak{c})}}, x_{(\mathfrak{c}')}\right\rangle\right] + \mathbb{E}\left[\left\langle \widehat{x_{(\mathfrak{c})}}, \widehat{x_{(\mathfrak{c}')}}\right\rangle\right] = 0$. □

With this result at hand, we can derive useful theoretical guarantees, especially for high-dimensional vectors (e.g., gradient updates in NN federated and distributed training). For example

Corollary 2. *For n clients with arbitrary vectors in $\mathbb{R}^d$ and $d \geqslant 10^5$, the NMSE is smaller than $\frac{0.673}{n}$.*

B Message Representation Length

Here we discuss the length of the message $\left(S, \text{sign}(\mathcal{R}(x))\right)$ that Buffy sends to Angel. The message includes the sign vector $\text{sign}(\mathcal{R}(x))$, which takes exactly d bits. It remains to consider how S is to be represented and how that affects the vNMSE of our algorithms. We first note that one must assume that the norm of the sent vector x is bounded, as otherwise no finite number of bits can be used for the transmission. This is justified as the coordinates of the vector are commonly represented by a

floating point representation with constant length (e.g., using 32 bits). In particular, this assumption implies that $\|x\|_2 = O(d)$, with each coordinate is bounded by a constant. It follows that all the quantities we consider for the scale satisfy $S = O(d)$. This is because the scale of Theorem 2 satisfies $\frac{\|\mathcal{R}(x)\|_1}{d} \leqslant \frac{\sqrt{d}\cdot\|\mathcal{R}(x)\|_2}{d} = \frac{\|x\|_2}{\sqrt{d}} = O(\sqrt{d})$ (where here the first step uses the Cauchy-Schwarz inequality), and similarly the scale of Theorem 3 satisfies $\frac{\|x\|_2^2}{\|\mathcal{R}(x)\|_1} \leqslant \frac{\|x\|_2^2}{\|\mathcal{R}(x)\|_2} = \|x\|_2 = O(d)$.

As shown in [2], this means that one can use $O(\log d)$ bits to represent S to within a $d^{-\Omega(1)}$ additive error, i.e., a factor that rapidly drops to zero as d grows. This increases the vNMSE for a single vector under DRIVE by at most $d^{-\Omega(1)}$, i.e., an additive factor that diminishes as d grows.

In the distributed mean estimation setting, where n clients send their vectors for averaging, [2] suggested using $O(\log(nd))$ bits. In fact, the dependency on n may be dropped. Specifically, $O(\log d)$ bits suffice if we encode S *in an unbiased manner*. One simple way to do so is to encode $\lfloor S \rfloor$ exactly (which requires $O(\log d)$ bits since $S = O(d)$) and to encode the remainder $r = S - \lfloor S \rfloor$ using $O(\log d)$-bit stochastic quantization. (That is, express r using the $O(\log d)$ bits but use randomized rounding for the last bit.) This approach still yields a vNMSE increase of $d^{-\Omega(1)}$ for each vector. As for the distributed mean estimation task, it follows that Theorem 5 still holds for DRIVE as the vectors are sent in an unbiased manner. Thus the NMSE increases by only $(nd)^{-\Omega(1)}$.

The above argument means that DRIVE can use $d + O(\log d) = d(1 + o(1))$ bit messages. While in practice implementations use 32- or 64-bit representations for S thus send $d + O(1)$ bit messages, we use the $d(1 + o(1))$ notation to ensure compatibility with the theoretical guarantees.

Similar arguments show that our improved DRIVE$^+$ algorithm, presented in Section 5, can also be encoded using $d + O(\log d) = d(1 + o(1))$ bits.

C Supplementary Material for DRIVE$^+$

In this section, we provide additional information about the DRIVE$^+$ algorithm.

C.1 Pseudocode

Here, we provide the pseudocode for DRIVE$^+$ in Algorithm 2:

Algorithm 2 DRIVE$^+$

Buffy:

1: Compute $\mathcal{R}(x)$.
2: Compute the $c_0, c_1 \in \mathbb{R}$ values that minimize
$$\sum_{i=1}^{d} \min\left\{(\mathcal{R}(x)_i - c_0)^2, (\mathcal{R}(x)_i - c_1)^2\right\}.$$
3: Compute $X \in \{0,1\}^d$ such that
$$X_i = \begin{cases} 0 & \text{if } |\mathcal{R}(x)_i - c_0| \leqslant |\mathcal{R}(x)_i - c_1| \\ 1 & \text{otherwise} \end{cases}.$$
4: Compute S^+ and $\overline{c}_0 = S^+ \cdot c_0,\ \overline{c}_1 = S^+ \cdot c_1$.
5: Send $\left(\overline{c}_0, \overline{c}_1, X\right)$ to Angel.

Angel:

1: Compute $\widehat{\mathcal{R}(x)} \in \{\overline{c}_0, \overline{c}_1\}^d$ such that
$$\widehat{\mathcal{R}(x)}_i = \overline{c}_{X_i}.$$
2: Estimate $\widehat{x} = \mathcal{R}^{-1}(\widehat{\mathcal{R}(x)})$.

We now provide the vNMSE guarantees for DRIVE$^+$.

C.2 1b - Vector Estimation with DRIVE$^+$

We provide the following lemma from which it follows that the vNMSE of DRIVE$^+$ with $S^+ = 1$ is upper-bounded by that of DRIVE when both algorithms are not required to be unbiased.

Lemma 10. *For any $R \sim \mathcal{R}$ and any $x \in \mathbb{R}^d$, the SSE of DRIVE$^+$ with $S^+ = 1$ is upper-bounded by the SSE of DRIVE for any choice of S.*

Proof. By definition, the values c_0, c_1 respect that $\sum_{i=1}^{d} \min\left\{(\mathcal{R}(x)_i - c_0)^2, (\mathcal{R}(x)_i - c_1)^2\right\} \leqslant \sum_{i=1}^{d} \min\left\{(\mathcal{R}(x)_i - S)^2, (\mathcal{R}(x)_i + S)^2\right\}$ for any choice of S. Now, recall that the SSE in estimating $\mathcal{R}(x)$ using $\widehat{\mathcal{R}(x)}$ equals that of estimating x using $\widehat{x}$. This concludes the proof. □

C.3 1b - Distributed Mean Estimation with DRIVE$^+$

In this subsection, we prove that DRIVE$^+$ can be made unbiased, when using the uniform random rotation $\mathcal{R}_U$, by the right choice of scale S^+, and that its vNMSE is upper bounded by that of DRIVE.

Let $c \in \{c_0, c_1\}^d$ where $c_0, c_1 \in \mathbb{R}$ minimize the term $\sum_{i=1}^{d}(\mathcal{R}(x)_i - c_i)^2$. Namely, c_0, c_1 are the solution to the 2-means optimization over the entries of the rotated vector.

Theorem 6. *For any $x \in \mathbb{R}^d$ set $S^+ = \frac{\|x\|_2^2}{\|c\|_2^2}$. Then, $\mathbb{E}[\widehat{x}] = x$. Thus, Theorem 5 applies to DRIVE$^+$.*

Proof. For ease of exposition, for any vector $y \in \mathbb{R}^d$, we denote by $c(y)$ the vector that minimizes the term $\sum_{i=1}^{d}(y_i - c_i)^2$ where $c(y) \in \{c_0, c_1\}^d$ and $c_0, c_1 \in \mathbb{R}$.

For the vector $\mathcal{R}(x)$ we drop the notation and simply use c. Namely, $c \in \{c_0, c_1\}^d$ where $c_0, c_1 \in \mathbb{R}$ is a vector that minimizes the term $\sum_{i=1}^{d}(\mathcal{R}(x)_i - c_i)^2$.

The proof follows similar lines to that of Theorem 3.

For any $x \in \mathbb{R}^d$ denote $x' = (\|x\|_2, 0, \ldots, 0)^T$ and let $R_{x \to x'} \in \mathbb{R}^{d \times d}$ be a rotation matrix such that $R_{x \to x'} \cdot x = x'$. Further, denote $R_x = R_U R_{x \to x'}^{-1}$. Using these definitions we have that,

$$\begin{aligned}
\widehat{x} &= R_{x \to x'}^{-1} \cdot R_{x \to x'} \cdot \widehat{x} = R_{x \to x'}^{-1} \cdot R_{x \to x'} \cdot R_U^{-1} \cdot S^+ \cdot c\left(R_U \cdot x\right) \\
&= R_{x \to x'}^{-1} \cdot R_x^{-1} \cdot S^+ \cdot c\left(R_x \cdot R_{x \to x'} \cdot x\right) = S^+ \cdot R_{x \to x'}^{-1} \cdot R_x^{-1} \cdot c\left(R_x \cdot x'\right) \\
&= R_{x \to x'}^{-1} \cdot \frac{\|x\|_2^2 \cdot R_x^{-1} \cdot c\left(R_x \cdot x'\right)}{\|c\|_2^2} .
\end{aligned}$$

Again, let C_i be a vector containing the values of the i'th column of R_x. Then, $R_x \cdot x' = \|x\|_2 \cdot C_0$ and therefore $c(R_x \cdot x') = c(\|x\|_2 \cdot C_0) = \|x\|_2 \cdot c(C_0)$. We obtain,

$$R_x^{-1} \cdot c\left(R_x \cdot x'\right) = \|x\|_2 \cdot \left(\|c(C_0)\|_2^2, \langle C_1, c(C_0)\rangle, \ldots, \langle C_{d-1}, c(C_0)\rangle\right)^T ,$$

Next, we have that,

$$\widehat{x} = R_{x \to x'}^{-1} \cdot \|x\|_2 \cdot \left(\frac{\|x\|_2^2 \cdot \|c(C_0)\|_2^2}{\|c\|_2^2}, \frac{\|x\|_2^2 \cdot \langle C_1, c(C_0)\rangle}{\|c\|_2^2}, \ldots, \frac{\|x\|_2^2 \cdot \langle C_{d-1}, c(C_0)\rangle}{\|c\|_2^2}\right)^T . \quad (8)$$

Observe that $\|x\|_2^2 \cdot \|c(C_0)\|_2^2 = \|c(R_x \cdot x')\|_2^2 = \|c(R_U \cdot x)\|_2^2 = \|c\|_2^2$. This means that the first coordinate in the above vector is 1.

We continue the proof by using the same construction as in the proof of Theorem 3.

In particular, consider an algorithm DRIVE $'$ that operates exactly as DRIVE but, instead of directly using the sampled rotation matrix $R_U = R_x \cdot R_{x \to x'}^{-1}$ it calculates and uses the rotation matrix $R'_U = R_x \cdot I' \cdot R_{x \to x'}^{-1}$ where I' is identical to the d-dimensional identity matrix with the exception that $I'_{00} = -1$ instead of 1.

Now, since $R_U \sim \mathcal{R}_U$ and $R_{x \to x'}$ is a fixed rotation matrix, we have that $R_x \sim \mathcal{R}_U$. In turn, this also means that $R_x \cdot I' \sim \mathcal{R}_U$ since I' is a fixed rotation matrix.

Consider a run of both algorithm where $\widehat{x}$ is the reconstruction of DRIVE for x with a sampled rotation R_U and $\widehat{x}'$ is the corresponding reconstruction of DRIVE $'$ for x with the rotation R'_U.

According to (8) it holds that: $\widehat{x}+\widehat{x}' = R_{x\to x'}^{-1}\cdot\|x\|_2\cdot(2,0,\ldots,0)^T = 2\cdot x$. This is because both runs are identical except that the first column of R_x and $R_x\cdot I'$ have opposite signs and thus the resulting centroids $c(C_0)$ also flip signs in the run of DRIVE $'$. In particular, it holds that $\mathbb{E}\left[\widehat{x}+\widehat{x}'\right] = 2\cdot x$. But, since $R_x \sim \mathcal{R}_U$ and $R_x\cdot I' \sim \mathcal{R}_U$, both algorithms have the same expected value. This yields $\mathbb{E}\left[\widehat{x}\right] = \mathbb{E}\left[\widehat{x}'\right] = x$. This concludes the proof. □

To prove the SSE of DRIVE$^+$ is lower-bounded by that of DRIVE, we use the following observation.

Observation 3. *For $i\in\{0,1\}$, let $I_i = \{j\in\{1,\ldots,d\} \mid |\mathcal{R}(x)_j - c_i| \leqslant |\mathcal{R}(x)_j - c_{1-i}|\}$ denote the set of points closer to c_i. Then $c_i = \frac{\sum_{j\in I_i}\mathcal{R}(x)_j}{|I_i|}$.*

We are now ready to prove the bound on the vNMSE DRIVE$^+$.

Lemma 11. *For any d and R_U, the SSE of DRIVE$^+$ with $S^+ = \frac{\|x\|_2^2}{\|c\|_2^2}$ is upper-bounded by the SSE of DRIVE with $S = \frac{\|x\|_2^2}{\|\mathcal{R}(x)\|_1}$. Thus, Theorem 4 and Corollary 1 apply to DRIVE$^+$.*

Proof. Recall that $c\in\{c_0,c_1\}^d$, where $c_0,c_1\in\mathbb{R}$, minimizes the term $\sum_{i=1}^d(\mathcal{R}(x)_i - c_i)^2$ and that the quantization SSE of the rotated vector equals the SSE of estimating x using $\widehat{x}$. For DRIVE$^+$, this means that the SSE is given by

$$\begin{aligned}\left\|\mathcal{R}(x) - S^+\cdot c\right\|_2^2 &= \|\mathcal{R}(x)\|_2^2 - 2\cdot S^+\cdot\langle\mathcal{R}(x),c\rangle + \left\|S^+\cdot c\right\|_2^2\\ &= \|x\|_2^2 - 2\cdot S^+\cdot\langle\mathcal{R}(x),c\rangle + (S^+)^2\cdot\|c\|_2^2\\ &= \|x\|_2^2 - 2\cdot S^+\cdot\|c\|_2^2 + (S^+)^2\cdot\|c\|_2^2\,.\end{aligned}$$

According to Observation 3, the inner produce in the first coordinate satisfies:

$$\langle\mathcal{R}(x),c\rangle = \sum_{j\in I_0}\mathcal{R}(x)_j\cdot c_0 + \sum_{j\in I_1}\mathcal{R}(x)_j\cdot c_1 = c_0^2\cdot|I_0| + c_1^2\cdot|I_1| = \|c\|_2^2\,.$$

We we will show that for any vector $x\in\mathbb{R}^d$ and any R_U, the SSE of DRIVE$^+$ with $S^+ = \frac{\|x\|_2^2}{\|c\|_2^2}$ is upper bounded by the SSE of DRIVE with $S = \frac{\|x\|_2^2}{\|\mathcal{R}_U(x)\|_1}$. Both results immediately follow.

From the above and Theorem 1, it is sufficient to show that $\|c\|_2^2\left[(S^+)^2 - 2S^+\right] \leqslant -2\cdot S\cdot\|\mathcal{R}(x)\|_1 + S^2\cdot d$, which, by substituting S and S^+, is equivalent to $\frac{\|x\|_2^4}{\|c\|_2^2} \leqslant \frac{\|x\|_2^4}{\|\mathcal{R}(x)\|_1^2}$ or $\|c\|_2 \leqslant \|\mathcal{R}(x)\|_1$. The last inequality immediately holds since $\|c\|_2 \leqslant \|c\|_1$ and $\|c\|_1 \leqslant \|\mathcal{R}(x)\|_1$ by the triangle inequality. □

Similarly to DRIVE, one can deterministically set $S^+ = \frac{\|x\|_2^2}{\mathbb{E}\left[\|c\|_2^2\right]}$ and gain a $\frac{\pi}{2}-1$ upper bound on the vNMSE for any dimension d. Notice that this requires calculating this expression; for any dimension d, one can approximate this quantity once to within arbitrary precision using a Monte-Carlo simulation. Nevertheless, as mentioned, this is less favorable as it introduces undesirable computational overhead for non uniform rotations.

C.4 Implementation Notes

We note that it is possible to introduce a tradeoff between the additional complexity introduced by the computation of the centroids and the reduction in vMNSE compared to DRIVE. This is achieved by initializing the centroids symmetrically to that of DRIVE and continuing to apply iterations of Lloyd's algorithm. Each such additional iteration introduces more calculations but reduces the SSE.

In high dimensions, we find that the SSE of DRIVE and DRIVE$^+$ similar for vectors whose distribution admits finite moments (they converge to the same value). In low dimensions, we often find that after 2-3 such iterations, further improvement is negligible. Since Lloyd's algorithm admits an efficient GPU implementation, these extra iterations may offer a better vMNSE to computation tradeoff than optimally finding the centroids using less GPU-friendly implementations.

D Supplemental Results for Structured Random Rotations

We next provide supplementary results for DRIVE and DRIVE$^+$ with a structured random rotation.

D.1 Convergence of the Rotated Coordinates to a Normal Distribution

For completeness, we restate Lemma 4:

Lemma 4. *For all i, $\mathcal{R}_H(x)_i$ converges to a normal variable:* $\sup_{x\in\mathbb{R}} |F_{i,d}(x) - \Phi(x)| \leqslant \frac{0.409\cdot\rho}{\sigma^3\sqrt{d}}$.

Proof. By definition $\frac{1}{\sqrt{d}} \cdot \mathcal{R}_H(x)_i = \frac{1}{d} \cdot \sum_{j=1}^{d} x_j H_{ij} D_{jj}$ for all i. Due to the sign symmetry of D_{jj}, it holds that $x_j H_{ij} D_{jj}$ are zero-mean i.i.d. random variables. Also, $\mathbb{E}\left[x_j H_{ij} D_{jj}\right]^2 = \sigma^2$ and $\mathbb{E}\left[|x_j H_{ij} D_{jj}|\right]^3 = \rho$. Now, denote by $F_{i,d}$ be the cumulative distribution function (CDF) of $\frac{1}{\sigma} \cdot \mathcal{R}_H(x)_i$. Using the Berry–Esseen theorem [70], we obtain than $\sup_{x\in\mathbb{R}} |F_{i,d}(x) - \Phi(x)| \leqslant \frac{0.409\cdot\rho}{\sigma^3\sqrt{d}}$, where Φ the CDF of the standard normal distribution. This concludes the proof. □

D.2 Analysis of the 4th Moment When Using a Structured Random Rotation

Further assume that $\mathbb{E}[x_j^4] = M_4 < \infty$. Then, we have that,

$$\begin{aligned}
&\mathbb{E}\left[(\mathcal{R}_H \cdot x)_{i_1} \cdot (\mathcal{R}_H \cdot x)_{i_2} \cdot (\mathcal{R}_H \cdot x)_{i_3} \cdot (\mathcal{R}_H \cdot x)_{i_4}\right] \\
&= \frac{1}{d^2} \sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d} \mathbb{E}\left[x_j H_{i_1 j} D_{jj} \cdot x_k H_{i_2 k} D_{kk} \cdot x_l H_{i_3 l} D_{ll} \cdot x_m H_{i_4 m} D_{mm}\right] \\
&= \frac{1}{d^2} \sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d} H_{i_1 j} H_{i_2 k} H_{i_3 l} H_{i_4 m} \cdot \mathbb{E}\left[x_j x_k x_l x_m \cdot D_{jj} D_{kk} D_{ll} D_{mm}\right].
\end{aligned}$$

There are only two cases for which the expectation is not zero. Where there are two different indexes (where the resulting expectation is σ^4) and where all four are identical (where the resulting expectation is M_4). Therefore,

$$\begin{aligned}
&\mathbb{E}\left[(\mathcal{R}_H \cdot x)_{i_1} \cdot (\mathcal{R}_H \cdot x)_{i_2} \cdot (\mathcal{R}_H \cdot x)_{i_3} \cdot (\mathcal{R}_H \cdot x)_{i_4}\right] \\
&= \frac{1}{d^2} \cdot \sigma^4 \cdot \left[\sum_{j=1}^{d} H_{1+(i_1-1)\oplus(i_2-1),j} \cdot \left[\sum_{k=1,k\neq j}^{d} H_{1+(i_3-1)\oplus(i_4-1),k}\right]\right] \\
&\quad + \frac{1}{d^2} \cdot \sigma^4 \cdot \left[\sum_{j=1}^{d} H_{1+(i_1-1)\oplus(i_3-1),j} \cdot \left[\sum_{k=1,k\neq j}^{d} H_{1+(i_2-1)\oplus(i_4-1),k}\right]\right] \\
&+ \frac{1}{d^2} \cdot \sigma^4 \cdot \left[\sum_{j=1}^{d} H_{1+(i_1-1)\oplus(i_4-1),j} \cdot \left[\sum_{k=1,k\neq j}^{d} H_{1+(i_2-1)\oplus(i_3-1),k}\right]\right] \\
&\quad + \frac{1}{d^2} \cdot M_4 \cdot \sum_{j=1}^{d} H_{1+(i_1-1)\oplus(i_2-1)\oplus(i_3-1)\oplus(i_4-1),j}
\end{aligned}$$

Now we can calculate all the fourth moments.

There are four cases to consider: (1) for $i_1 = i_2 = i_3 = i_4$ we obtain $3 \cdot \sigma^4 \cdot \frac{d\cdot(d-1)}{d^2} + \frac{d}{d^2} \cdot M_4 = 3 \cdot \sigma^4 + O(\frac{1}{d})$; (2) for all pair equalities (e.g., $i_1 = i_2$ and $i_3 = i_4$ but $i_1 \neq i_2$) we obtain $\sigma^4 \cdot \frac{d\cdot(d-1)}{d^2} + \frac{d}{d^2} \cdot M_4 = \sigma^4 + O(\frac{1}{d})$; (3) when having a unique entry and at least a pair that equals (e.g., $i_1 = i_2, i_2 \neq i_4$ and $i_3 \neq i_4$ regardless of whether $i_2 = i_3$) we immediately obtain zero (recall that all odd moments are zero); (4) for all distinct entries it may be the case that $1 + (i_1 - 1) \oplus (i_2 - 1) \oplus (i_3 - 1) \oplus (i_4 - 1) = 1$, therefore we obtain $\frac{d}{d^2} \cdot M_4 = O(\frac{1}{d})$.

These indeed approach, with a rate of $\frac{1}{d}$, the 4'th moments of *independent* normal random variables.

E Additional Simulation Details and Results

In this section we provide additional details and simulation results. The source code and full reproduction instructions are in the supplementary material archive.

E.1 Preexisting Code and Datasets Description

Code. Our federated learning experiments use the TensorFlow Federated [44] library with the Hadamard rotation and the Kashin's representation code taken from the TensorFlow Model Optimization Toolkit [71]. Tensorflow and its subpackages are licensed under the Apache License, Version 2.0 [72]. The rest of the experiments use PyTorch [43], which is licensed under a BSD-style license [73].

Datasets. Next we provide a quick overview of the datasets used:

MNIST [49, 50]. Includes 70,000 examples of 28×28 grayscale labeled images of handwritten digits (10 classes).

EMNIST [51]. Extends MNIST and includes 749,068 examples of 28×28 grayscale labeled images of 62 handwritten characters.

CIFAR-10 and *CIFAR-100* [52]. Both consist of 60,000 examples of 32×32 images with 3 color channels, in each dataset the images are labeled into 10 and 100 classes, respectively.

Shakespeare [53]. Contains 18,424 examples, where each example is a contiguous set of lines spoken by a character in a play by Shakespeare.

Stack Overflow [54]. Consists of 152,404,765 examples, each containing the text of either an answer or a question. The data consists of the body text of all questions and answers

For the federated learning experiments, following previous works in the field [1, 59], the EMNIST, Shakespeare, and Stack Overflow datasets are partitioned naturally to create an heterogeneous unbalanced client dataset distribution. Specifically, they are partitioned by character writer, play speaker, and Stack Overflow user, respectively. In the CIFAR-100 federated experiment, the examples are split among 500 clients using a random heterogeneity-inducing process described in [60, Appendix C.1].

All datasets were released under CC BY-SA 3.0 [74] or similar licenses, the exact details are in the cited references.

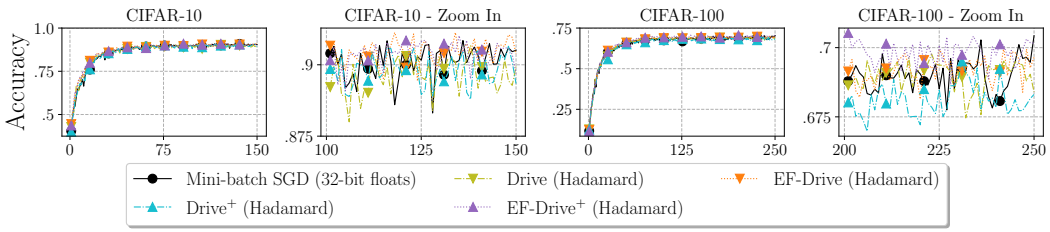

Figure 4: Distributed CNN EF experiments comparing DRIVE and DRIVE$^+$ with and without EF. Accuracy per round on distributed learning tasks, with a zoom-in on the last 50 rounds.

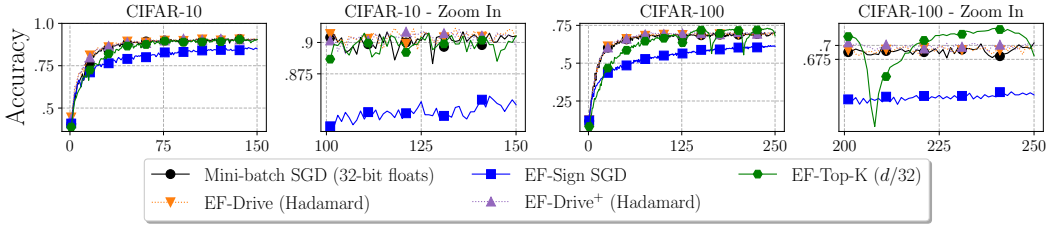

Figure 5: Distributed CNN EF experiments comparing DRIVE and DRIVE$^+$ with and techniques utilizing EF. Accuracy per round on distributed learning tasks, with a zoom-in on the last 50 rounds.

E.2 Error-Feedback Experiments

We next conduct error feedback (EF) experiments for the distributed CNN training. For DRIVE (and similarly for DRIVE$^+$), we use the following formula to set the scale, $S = \min\{2 \cdot \frac{\|\mathcal{R}_H(x)\|_1}{d}, \frac{\|x\|_2^2}{\|\mathcal{R}_H(x)\|_1}\}$. This scale ensures that the *compressor assumption* [64, Assumption A] holds but tries to make use of the "unbiased" scale whenever possible. In fact, in our CNN simulations, we did not encounter a single iteration in which $2 \cdot \frac{\|\mathcal{R}_H(x)\|_1}{d} \leqslant \frac{\|x\|_2^2}{\|\mathcal{R}_H(x)\|_1}$.

Figure 4 shows that with and without EF, DRIVE and DRIVE$^+$ result in similar performance. Also, as evident in Figure 5, DRIVE and DRIVE$^+$ show favorable performance in comparison to EF-Sign SGD and Top-K. Here, we configured the Top-K algorithm with $K = d/32$; note that this requires more than one bit per coordinate, as the sender needs to encode the indices of the sent coordinates in addition to their values.

E.3 Additional vNMSE-Speed Tradeoff Details and Results

Experiment Configuration. The experiments were executed over Intel Core i9-10980XE CPU (18 cores, 3.00 GHz, and 24.75 MB cache), 128 GB RAM, NVIDIA GeForce RTX 3090 GPU, Ubuntu 20.04.2 LTS operating system, and CUDA release 11.1, V11.1.105. We draw 100 vectors[4] from a Lognormal(0,1) distribution and measure 100 encodings of each vector. Similarly to Figure 1, each of $n = 10$ clients are given the same vector at each iteration. Due to its high runtime, we only evaluate the uniform random rotation on the low ($d \in \{128, 8192\}$) dimensions.

Experiment Results Summary. The results, given in Table 1, show that DRIVE and DRIVE$^+$ are the most accurate algorithms and their NMSE converges to the theoretical NMSE of $\frac{\pi/2-1}{10} \approx 0.0571$ even in dimensions as small as 2^{13}. In even lower dimensions, such as $d = 128$, DRIVE$^+$ is more accurate than DRIVE, albeit slightly slower. Running DRIVE and DRIVE$^+$ with uniform rotation instead of Hadamard yield even lower NMSE, but also takes more time. DRIVE is also almost as fast as Hadamard with 1-bit SQ [2] and is 12x faster than Kashin with 1-bit SQ. We conclude that DRIVE with Hadamard offers the best tradeoff in all cases, except when the dimensions is small, in which case one may choose to use DRIVE$^+$ and possibly uniform random rotation.

Dimension (d)	Hadamard + 1-bit SQ	Kashin + 1-bit SQ	Drive (Uniform)	Drive$^+$ (Uniform)	Drive (Hadamard)	Drive$^+$ (Hadamard)
128	0.5308, *0.93*	0.2550, *6.06*	0.0567, *14.95*	**0.0547**, *18.25*	0.0591, *0.98*	0.0591, *2.07*
8,192	1.3338, *1.53*	0.3180, *10.9*	**0.0571**, *2646*	**0.0571**, *2672*	**0.0571**, *1.58*	**0.0571**, *2.93*
524,288	2.1456, *3.76*	0.3178, *30.8*	—	—	**0.0571**, *3.78*	**0.0571**, *5.72*
33,554,432	2.9332, *40.9*	0.3179, *1714*	—	—	**0.0571**, *43.0*	**0.0571**, *252*

Table 3: A *commodity machine* comparison of empirical NMSE and average per-vector encoding time (in milliseconds) for distributed mean estimation with $n = 10$ clients (same as in Figure 1) and Lognormal(0,1) distribution. Each entry is a (NMSE, *time*) tuple and the best result is highlighted in **bold**.

Speed Measurements on a Commodity Machine. The experiment in Table 1 in the paper was performed on a high-end server with a NVIDIA GeForce RTX 3090 GPU. For completeness, we also run measurements on a commodity machine. Specifically, we re-execute the experiments over Intel Core i7-7700 CPU (8 cores, 3.60 GHz, and 8 MB cache), 64 GB RAM, NVIDIA GeForce GTX 1060 (6GB) GPU, Windows 10 (build 18363.1556) operating system and CUDA release 10.2, V10.2.89. The results, shown in Table 3 show a similar trend. Again, as expected, DRIVE and DRIVE$^+$ with uniform rotation are more accurate for small dimensions but are significantly slower. On this machine, DRIVE is 6.18×-39.8× faster than Kashin and, again, about as fast as Hadamard. All NMSE measurements, as expected yielded the same results as in Table 1.

[4] except for $d = 2^{25}$, where we use 10 encodings of each vector due to the fact that in such high dimension, random vectors are concentrated close to their expectation.

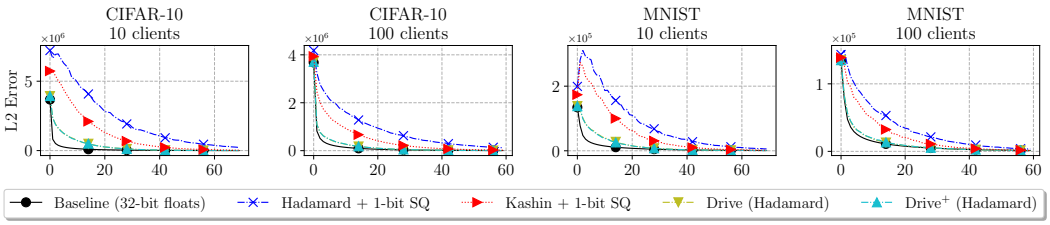

Figure 6: L2 error per round for distributed power iteration.

E.4 Federated Learning Experiments Configuration

The experiments were executed over a university cluster that contains several types of NVIDIA GPUs (TITAN X (Pascal), TITAN Xp, GeForce GTX 1080 Ti, GeForce RTX 2080 Ti, and GeForce RTX 3090) with no specific hardware guarantee.

Task	Clients per round	Rounds	Batch size	Client lr	Server lr
EMNIST	10	1500	20	0.1	1
CIFAR-100	10	6000	20	0.1	0.1
Shakespeare	10	1200	4	1	1
Stack Overflow	50	1500	16	0.3	1

Table 4: Summary of main hyperparameters for federated learning tasks.

Table 4 contains a summary of the hyperparameters we used. Those generally correspond to the best setup described in [60, Appendix D] for FedAvg except for the following changes: (1) We increased the number of rounds of the EMNIST task from 4000 to 6000 to show a clear convergence; (2) We used a server momentum of 0.9 (named *FedAvgM* in [60]) for the Stack Overflow task since otherwise, this task produces noisy results due to its extreme unbalancedness in the number of samples per client; and (3) For FetchSGD on EMNIST and Stack Overflow we set the server learning rate to 0.3 as it offered improved performance.

We compress each layer separably due to the implementation of the libraries we use and algorithms we compare to (specifically, see "CLASS ENCODEDSUMFACTORY" [75] in TensorFlow Federated). This layer-wise compression reduces the dimension in practice and explains the difference seen between DRIVE and DRIVE$^+$ performance. Additionally, all compression algorithms skip layers with less than 10,000 parameters since those are usually either normalization or last fully-connected layers, which are more cost-effective to send as-is.

E.5 Distributed Learning Experiments Configuration

The experiments were executed over Intel Core i9-10980XE CPU (18 cores, 3.00 GHz, and 24.75 MB cache), 128 GB RAM, NVIDIA GeForce RTX 3090 GPU, and Ubuntu 20.04.2 LTS OS.

For the distributed CNN training we have used an SGD optimizer with a Cross entropy loss criterion, a momentum of 0.9, and a weight decay of 5e-4. The learning rate of all algorithms was set to 0.1 for the CIFAR-10 over ResNet-9 experiment and 0.05 for the CIFAR-100 over ResNet-18 experiments. TernGrad was an exception and its learning rate was set to 0.01 and 0.005 as it offered improved performance. The batch size in all experiments and clients was set to 128.

E.6 K-Means and Power Iteration Simulation Results

Distributed Power Iteration. Power Iteration is a simple method for approximating the dominant Eigenvalues and Eigenvectors of a matrix. It is often used as a sub-routine in more complicated tasks such as Principal Component Analysis. In our distributed setting, at each epoch, the server broadcasts the current estimated top eigenvector to all clients. Then, each client: (1) updates the top eigenvector based on its local data; (2) computes its diff to the current estimated top eigenvector; (3) compresses the diff vector; (4) sends it to the server, possibly scaled by a *learning rate* to ensure

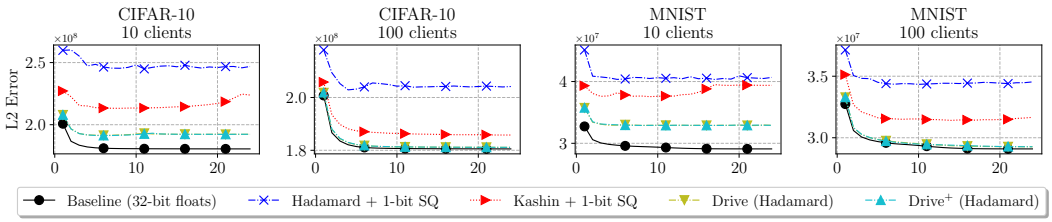

Figure 7: L2 error per round for distributed K-Means.

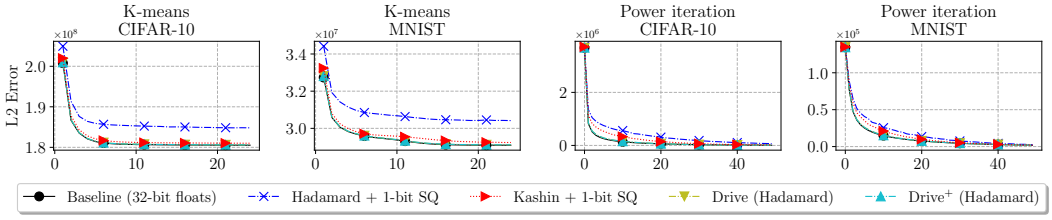

Figure 8: Distributed K-Means and power iteration of CIFAR-10 and MNIST with 1000 clients.

convergence to due the high estimation variance. Figure 6 presents the results in a setting with $n = 10$ and $n = 100$ clients and a learning rate of 0.1.

In both datasets, and for both $n = 10$ and $n = 100$ clients, DRIVE and DRIVE$^+$ have lower error and faster convergence than other compression algorithms. Our convergence is faster when using a larger number of clients, which matches the theoretical analysis which shows that the estimates of DRIVE (Hadamard) and DRIVE$^+$ (Hadamard) are (nearly) unbiased for such a high dimension.

Distributed K-Means. In this distributed learning task, the server coordinates a run of Lloyd's algorithm over a distributed dataset. At each training round, the server broadcasts the centroids' coordinates to the clients. Then, each client: (1) updates these coordinates based on its local data; (2) compresses them; (3) sends them to the server. The centroid weights (i.e., the number of observations assigned to each centroid) require fewer bits and are essential for K-Means to converge quickly; therefore, the weights are sent without compression.

Figure 7 presents the results with $n = 10$ and $n = 100$ clients. As shown, DRIVE and DRIVE$^+$ have a lower error than other compression algorithms. While none of the compression methods is competitive with the baseline when using $n = 10$ clients, DRIVE and DRIVE$^+$ are the only algorithms that have comparable accuracy for $n = 100$.

Increasing the Number of Clients. Figure 8 depicts simulation results for the distributed K-Means and distributed Power Iteration experiments with $n = 1000$ clients. The results indicate a similar trend. DRIVE and DRIVE$^+$ offer the lowest L2 error among the low-bandwidth techniques, and they are now closer to the uncompressed baseline due to the larger number of clients, which further reduces the estimation variance.

E.7 Comparison With Entropy Encoding

We conduct an experiment for distributed mean estimation with $n = 10$ clients (same as in Figure 1 and Table 1) with the Lognormal(0,1) distribution. We determine how many bits per coordinate are required by a compression method that uses stochastic quantization followed by entropy encoding to reach the same NMSE as that of DRIVE (that uses a single bit per coordinate). We examine two methods: (1) standard stochastic quantization followed by Huffman encoding; (2) stochastic quantization with levels chosen for entropy encoding followed by Huffman encoding, as proposed in [2] (see Section 4). We refer to this second variant as Enhanced SQ. We introduced an increasing number of quantization levels for each of these methods until the resulting empirical NMSE was similar to that of DRIVE. We repeat each experiment 100 times and report the mean value. The results summarized in Table 5 indicate that these methods require at least 1.261 bits per coordinate.

Dimension (d)	DRIVE (Hadamard) NMSE (1-bit)	Required bits/coordinates	
		SQ + Huffman	Enhanced SQ + Huffman [2]
128	0.0591	1.284	1.261
8,192	0.0571	1.335	1.295
524,288	0.0571	1.324	1.298
33,554,432	0.0571	1.321	1.301

Table 5: Empirical NMSE for distributed mean estimation with $n = 10$ clients (same as in Figure 1 and Table 1) with the Lognormal(0,1) distribution. We find the number of bits per coordinate that are required by the entropy encoding methods to reach the same empirical NMSE as that of DRIVE (that uses a single bit per coordinate).

There are other uses of entropy encoding techniques for machine learning tasks as suggested by [2, 9]. Such methods often introduce higher computational costs. We leave further comparisons of DRIVE and entropy encoding for specific problems for future work.

F Lower Bounds

It is proven in [14] that *any* unbiased algorithm must have a 1b - Vector Estimation vNMSE of at least $\frac{1}{3}$ and any biased algorithm incurs a vNMSE of at least $\frac{1}{4}$. However, it does not appear that their proof accounts for algorithms in which Buffy and Angel may have shared randomness, as is the case for all algorithms proposed in our paper. The reason is that in [14] it is assumed that if Buffy sends b bits, Angel only has 2^b possible estimates, one for each of the 2^b possible messages. However, if shared randomness is allowed, there are more possible values for $\widehat{x}$. In fact, if the number of shared random bits is not restricted, the number of different possible estimates can be unbounded. Additionally, it is shown in [7] that for a single number ($d = 1$), using shared randomness allows algorithms to have an MSE lower than the optimal algorithm when no such randomness is allowed.

Nonetheless, their result implies that any algorithm that uses $O(d)$ shared random bits (as does our Hadamard-based variants) has a vNMSE of $\Omega(1)$, i.e., DRIVE and DRIVE$^+$ are asymptotically optimal. The reason is that Buffy could generate (private) random bits and send them as part of the message to mimic shared random bits. For example, in our Hadamard variant, Buffy could send the random diagonal matrix D using d additional bits. This implies that any biased algorithm with at most d shared random bits must have a vNMSE of at least $\frac{1}{16}$ and any unbiased solution must have a vNMSE of at least $\frac{1}{15}$.