Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor

Mher Safaryan Egor Shulgin Peter Richtárik

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

20 / 2 / 2020

Abstract

In order to mitigate the high communication cost in distributed and federated learning, various vector compression schemes, such as quantization, sparsification and dithering, have become very popular. In designing a compression method, one aims to communicate as few bits as possible, which minimizes the cost per communication round, while at the same time attempting to impart as little distortion (variance) to the communicated messages as possible, which minimizes the adverse effect of the compression on the overall number of communication rounds. However, intuitively, these two goals are fundamentally in conflict: the more compression we allow, the more distorted the messages become. We formalize this intuition and prove an *uncertainty principle* for randomized compression operators, thus quantifying this limitation mathematically, and *effectively providing lower bounds on what might be achievable with communication compression.* Motivated by these developments, we call for the search for the optimal compression operator. In an attempt to take a first step in this direction, we construct a new unbiased compression method inspired by the Kashin representation of vectors, which we call *Kashin compression (KC)*. In contrast to all previously proposed compression mechanisms, we prove that KC enjoys a *dimension independent* variance bound with an explicit formula even in the regime when only a few bits need to be communicate per each vector entry. We show how KC can be provably and efficiently combined with several existing optimization algorithms, in all cases leading to communication complexity improvements on previous state of the art.

Contents

1 Introduction

In the quest for high accuracy machine learning models, both the size of the model and consequently the amount of data necessary to train the model have been hugely increased over time (Schmidhuber, 2015; Vaswani et al., 2019). Because of this, performing the learning process on a single machine is often infeasible. In a typical scenario of distributed learning, the training data (and possibly the model as well) is spread across different machines and thus the process of training is done in a distributed manner (Bekkerman et al., 2011; Vogels et al., 2019). Another scenario, most common to federated learning (Konečný et al., 2016; McMahan et al., 2017; Karimireddy et al., 2019a), is when training data is inherently distributed across a large number of mobile edge devices due to data privacy concerns.

1.1 Communication bottleneck

In all cases of distributed learning and federated learning, information (e.g. current stochastic gradient vector or current state of the model) communication between computing nodes is inevitable, which forms the primary bottleneck of such systems (Zhang et al., 2017; Lin et al., 2018). This issue is especially apparent in federated learning, where computing nodes are devices with essentially inferior power and the network bandwidth is considerably slow (Li et al., 2019).

There are two general approaches to address/tackle this problem. One line of research dedicated to so-called local methods suggests to do more computational work before each communication in the hope that those would increase the worth/impact/value of the information to be communicated (Goyal et al., 2017; Wangni et al., 2018; Stich, 2018; Khaled et al., 2020). An alternative approach investigates inexact/lossy information compression strategies which aim to send approximate but relevant information encoded with less number of bits. In this work we focus on the second approach of *compressed learning*. Research in this latter stream splits into two orthogonal directions. To explore savings in communication, various (mostly randomized) compression operators have been proposed and analyzed such as random sparsification (Konečný & Richtárik, 2018; Wangni et al., 2018), top-k sparsification (Alistarh et al., 2018), standard random dithering (Goodall, 1951; Roberts, 1962; Alistarh et al., 2017), natural dithering (Horváth et al., 2019a), ternary quantization (Wen et al., 2017), and scaled sign quantization (Karimireddy et al., 2019b; Bernstein et al., 2018, 2019; Liu et al., 2019). Table 1 summarizes the most common compression methods with their variances and the number of encoding bits.

In designing a compression operator, one aims to (i) encode the compressed information with as few bits as possible, which minimizes the cost per communication round, and (ii) introduce as little noise (variance) to the communicated messages as possible, which minimizes the adverse effect of the compression on the overall iteration complexity.

Table 1: Compression operators in $\mathbb{U}(\omega)$ and $\mathbb{B}(\alpha)$.

Compression Method	Unbiased?	Variance ω	Variance α	Bits b (in *binary32*)
Random sparsification	YES	$\frac{d}{k} - 1 \approx \mathcal{O}(\frac{d}{k})$		$32k + \log_2 \binom{d}{k}$
Top-k sparsification	NO		$1 - \frac{k}{d}$	$32k + \log_2 \binom{d}{k}$
Standard Dithering	YES	$\min(\frac{\sqrt{d}}{s}, \frac{d}{s^2}) \approx \mathcal{O}(\frac{\sqrt{d}}{s})$		$31 + d \log_2(2s+1)$
Natural Dithering	YES	$\min(\frac{\sqrt{d}}{2^{s-1}}, \frac{d}{2^{2-2s}}) \approx \mathcal{O}(\frac{\sqrt{d}}{2^{s-1}})$		$31 + d \log_2(2s+1)$
Ternary Quantization	YES	$\sqrt{d} - 1 \approx \mathcal{O}(\sqrt{d})$		$31 + d \log_2 3$
Scaled Sign Quantization	NO		$1 - \frac{1}{d}$	$31 + d$
Kashin Compression (new)	YES	$\left({}^{10\sqrt{\lambda}}/_{\sqrt{\lambda}-1}\right)^4 \approx \mathcal{O}(1)$		$31 + \log_2 3 \cdot \lambda d$[1]

Table 2: Iteration complexities of different learning algorithms with respect to the variance of compression.

Optimization Algorithm	Objective Function	Iteration complexity
Compressed GD (Khirirat et al., 2018)	smooth, strongly convex	$\mathcal{O}\left(\kappa(\omega+1)\log {}^1/_\varepsilon\right)$
DIANA (Horváth et al., 2019b)	smooth, strongly convex	$\mathcal{O}\left((\kappa + \omega\frac{\kappa}{n} + \omega)\log {}^1/_\varepsilon\right)$
Distributed SGD (Horváth et al., 2019a)	smooth, non-convex	$\mathcal{O}\left((\omega+1)^2\, {}^1/_{\varepsilon^2}\right)$
DoublSqueeze (Tang et al., 2019)	smooth, non-convex	$\mathcal{O}\left({}^1/_{\varepsilon^2} + \frac{1}{1-\alpha^2}\, {}^1/_{\varepsilon^{1.5}}\right)$

1.2 Compressed learning

In order to utilize these compression methods efficiently, a lot of research has been devoted to the study of learning algorithms with compressed communication. Obviously, the presence of compression in a learning algorithm affects the training process and since compression operator encodes the original information approximately, it should be anticipated to increase the number of communication rounds. Table 2 highlights four gradient-type compressed learning algorithms with their corresponding setup and iteration complexity:

(i) distributed Gradient Descent (GD) with compressed gradients (Khirirat et al., 2018),

(ii) distributed Stochastic Gradient Descent (SGD) with gradient quantization and compression variance reduction (Horváth et al., 2019b),

(iii) distributed SGD with bi-directional gradient compression (Horváth et al., 2019a), and

(iv) distributed SGD with gradient compression and twofold error compensation (Tang et al., 2019).

In all cases, the iteration complexity depends on the variance (ω or α) of the underlying compression scheme and grows as more compression is applied. For this reason, we are interested in compression methods which save in communication by using less bits and minimize iteration complexity by introducing lower variance. However, intuitively and also evidently from Table 1, these two goals are in fundamental conflict, i.e. requiring fewer bits to be communicated in each round introduces higher variance, and demanding small variance forces more bits to be communicated.

1.3 Contributions

The contributions of our work are:

- **Uncertainty Principle.** We formalize this intuitive trade-off and *prove an uncertainty principle for randomized compression operators*, which quantifies this limitation mathematically with the inequality

$$\boxed{\alpha \cdot 4^{b/d} \geq 1}\,, \tag{1}$$

[1] In fact, the number of encoding bits depends on the quantization operator used in KC. Mentioned formula is for ternary quantization.

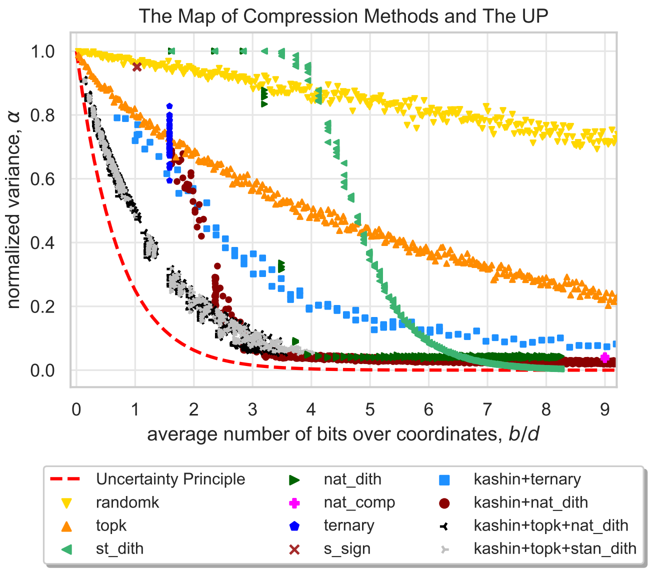


Figure 1: Comparison of the most common compression methods based on their normalized variance $\alpha \in [0, 1]$ and the average number of encoding bits per coordinate. Each color represents one compression method, each marker indicates one particular $d = 10^3$ dimensional vector randomly generated from Gaussian distribution, which subsequently gets compressed by the compression operator mentioned in the legend. Dashed red line shows the lower of bound of the uncertainty principle (1). *The Kashin compressor and the uncertainty principle are the key contributions of this paper.*

where $\alpha \in [0, 1]$ is the normalized variance / contraction factor associated with the compression operator (Definition 1), b is the number of bits required to encode the compressed vector and d is the dimension of the vector to be compressed. The notion of Uncertainty Principle (UP) for compression operators is introduced and theoretically proved in this paper. It is a universal property of compressed communication, completely independent of the optimization algorithm and the problem that distributed training is trying to solve. We visualize this fascinating principle in Figure 1, where we computed many possible combinations of parameters α and $^b/_d$ for various compression methods. The dashed red line indicating the lower bound (1) bounds all possible combinations of all compression operators, thus validating the obtained uncertainty principle for randomized compression operators.

- **Kashin Compression.** Motivated by this principle, we then focus on the search for the optimal compression operator. In an attempt to take a first step in this direction, we design a new unbiased compression operator inspired by Kashin representation of vectors (Kashin, 1977), which we call Kashin Compression (KC). In contrast to all previously proposed compression methods, we prove that KC enjoys a *dimension independent variance bound* even in a severe compression regime when only a few bits per coordinate can be communicated. We give an explicit formula for the variance bound and show how KC can be provably and efficiently combined with several existing optimization algorithms, in all cases leading to communication complexity improvements on previous state of the art. We believe that KC has the potential to play a role in the discovery of an optimal compression method, perhaps when composed with some other operators, such as dithering.

- **Experimental Validations.** In our experiments, we observed the superiority of KC in terms of communication savings and stabilization property when compared against a vast array of compressors proposed in the literature. In particular, Figure 1 justifies that KC combined with Top-k sparsification and dithering operators yields a compression method which almost closes the gap to the UP. Kashin's representation has been used heuristically in federated learning (Caldas et al., 2019) to mitigate the communication cost. In contrast to this work, we generate the initial tight frame of KC randomly as suggested by the theory, and tune the parameters accordingly. Moreover, we consider combinations of KC and other compression techniques such as ternary quantization, Top-k sparsification and dithering. We believe KC should be of high interest in federated and distributed learning.

2 Uncertainty principle for compression operators

In general, an uncertainty principle refers to any type of mathematical inequality expressing some fundamental trade-off between two measurements. The classical Heisenberg's uncertainty principle in quantum mechanics (Heisenberg, 1927) shows the trade-off between the position and momentum of a particle. In harmonic analysis, the uncertainty principle limits the localization of values of a function and its Fourier transform at the same time (Havin & Jöricke, 1994). Alternatively in the context of signal processing, signals cannot be simultaneously localized in both time domain and frequency domain (Gabor, 1946). The uncertainty principle in communication deals with the quite intuitive trade-off between information compression (encoding bits) and approximation error (variance), namely more compression forces heavier distortion to communicated messages and tighter approximation requires less information compression.

In this section, we present our UP for communication compression revealing the trade-off between encoding bits of compressed information and the variance produced by compression operator. First, we describe our UP for a general class of biased compressions. Afterwards, we specialize it to the class of unbiased compressions.

2.1 UP for biased compressions

We work with the class of biased compression operators which are contractive.

Definition 1 (Biased Compressions) *Let $\mathbb{B}(\alpha)$ be the class of biased (and possibly randomized) compression operators $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ with $\alpha \in [0,1]$ contractive property, i.e. for any $x \in \mathbb{R}^d$*

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \le \alpha \|x\|_2^2. \tag{2}$$

The parameter α can be seen as the normalized variance of the compression operator. Note that the compression $\mathcal{C}$ does not need to be randomized to belong to this class. For instance, Top-k sparsification operator satisfies (2) without the expectation for $\alpha = 1 - {}^{k}\!/_{d}$. Next, we formalize our uncertainty principle for the class $\mathbb{B}(\alpha)$.

Theorem 1 *Let $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ be any compression operator from $\mathbb{B}(\alpha)$ and b be the total number of bits needed to encode the compressed vector $\mathcal{C}(x)$ for any $x \in \mathbb{R}^d$. Then the following form of uncertainty principle holds*

$$\alpha \cdot 4^{b/d} \ge 1. \tag{3}$$

One can view the *binary32* and *binary64* floating-points formats as biased compression methods for the actual real numbers (i.e. $d = 1$), using only 32 and 64 bits respectively to represent a single number. Intuitively, these formats have their precision (i.e. $\sqrt{\alpha}$) limits and the uncertainty principle (3) shows that the precision cannot be better than 2^{-32} for *binary32* format and 2^{-64} for *binary64* format. Thus, any floating-point format representing a single number with r bits has precision constraint of 2^{-r}, where the base 2 stems from the binary nature of the bit.

Furthermore, notice that compression operators can achieve zero variance in some settings, e.g. ternary or scaled sign quantization when $d = 1$ (see Table 1). On the other hand, the UP (3) implies that the normalized variance $\alpha > 0$ for any finite bits b. The reason for this inconsistency comes from the fact that, for instance, the *binary32* format encodes any number with 32 bits and the error 2^{-32} is usually ignored in practice. We can adjust our UP to any digital format, using r bits per single number, as

$$\left(\alpha + 4^{-r}\right) \cdot 4^{b/d} \ge 1. \tag{4}$$

2.2 UP for unbiased compressions

We now specialize our UP to the class of unbiased compressions. First, we recall the definition of unbiased compression operators with a given variance.

Definition 2 (Unbiased Compressions) *Denote by $\mathbb{U}(\omega)$ the class of unbiased compression operators $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ with variance $\omega > 0$, that is, for any $x \in \mathbb{R}^d$*

$$\mathbb{E}\left[\mathcal{C}(x)\right] = x, \qquad \mathbb{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \le \omega \|x\|_2^2. \tag{5}$$

To establish an uncertainty principle for $\mathcal{C} \in \mathbb{U}(\omega)$, we show that all unbiased compression operators with the proper scaling factor are included in $\mathbb{B}(\alpha)$.

Lemma 1 *If $\mathcal{C} \in \mathbb{U}(\omega)$, then $\frac{1}{\omega+1}\mathcal{C} \in \mathbb{B}(\frac{\omega}{\omega+1})$.*

Using this inclusion, we can apply Theorem 1 to the class $\mathbb{U}(\omega)$ and derive an uncertainty principle for unbiased compression operators.

Theorem 2 *Let $\mathcal{C} \colon \mathbb{R}^d \to \mathbb{R}^d$ be any unbiased compression operator with variance $\omega \geq 0$ and b be the total number of bits needed to encode the compressed vector $\mathcal{C}(x)$ for any $x \in \mathbb{R}^d$. Then the uncertainty principle takes the form*

$$\frac{\omega}{\omega+1} \cdot 4^{b/d} \geq 1. \tag{6}$$

3 Compression with regular polytopes

Here we describe an unbiased compression scheme based on regular polytopes. With this particular compression we illustrate that it is possible for unbiased compressions to have dimension independent variance bounds and at the same time communicate a few bits per coordinate.

Let $x \in \mathbb{R}^d$ be the vector that we need to communicate. First, we project the vector on the unit sphere

$$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \colon \|x\|_2 = 1\},$$

thus separating the magnitude $\|x\|_2 \in \mathbb{R}$ from the direction $x/\|x\|_2 \in \mathbb{S}^{d-1}$. The magnitude is a dimension independent scalar value and we can transfer it cheaply, say by 32 bits. To encode the unit vector $x/\|x\|_2$ we approximate the unit sphere by regular polytopes and then randomize over the vertices of the polytope. Polytopes can be seen as generalizations of planar polygons in high dimensions. Formally, let P_m be a regular polytope with vertices $\{v_1, v_2, \dots, v_m\} \subset \mathbb{R}^d$ such that it contains the unit sphere, i.e. $\mathbb{S}^{d-1} \subset P_m$, and all vertices are on the sphere of radius $R > 1$. Then, any unit vector $v \in \mathbb{S}^{d-1}$ can be expressed as a convex combination $\sum_{k=1}^m w_k v_k$ with some non-negative weights $w_k = w_k(x)$. Equivalently, v can be expressed as an expectation of a random vector over v_k with probabilities w_k. Therefore, the direction $x/\|x\|_2$ could be encoded with roughly $\log m$ bits and the variance ω of compression will depend on the approximation, more specifically $\omega = R^2 - 1$. Kochol (2004, 1994) gave a constructive proof on approximation of the $d-$dimensional unit sphere by regular polytopes with $m \geq 2d$ vertices for which $\omega = \mathcal{O}\left(\frac{d}{\log {m/d}}\right)$. So, choosing the number of vertices to be $m = 2^d$, we get an unbiased compression operator with $\mathcal{O}(1)$ variance (independent of dimension d) and with 1 bit per coordinate encoding.

However, this method does not seem to be practical as 2^d vertices of the polytope either need to be stored or computed each time they are used, which is infeasible for large dimensions.

4 Compression with Kashin's representation

In this section we introduce the notion of Kashin's representation, the algorithm of Lyubarskii & Vershynin (2010) on computing it efficiently and then describe the quantization step.

4.1 Representation systems

The most common way of compressing a given vector $x \in \mathbb{R}^d$ is to use its *orthogonal representation* with respect to the standard basis $(e_i)_{i=1}^d$ in $\mathbb{R}^d$:

$$x = \sum_{i=1}^d x_i e_i, \qquad x_i = \langle x, e_i \rangle .$$

However, the restriction of orthogonal expansions is that coefficients x_i are independent in the sense that if we lost one of them, then we cannot recover it even approximately. Furthermore, each coefficient x_i may carry very different portion of the total information that vector x contains; some coefficients may carry more information than others and thus be more sensitive to compression.

Algorithm 1 Computing Kashin's representation

```
Input: orthogonal $d \times D$ matrix $U$ which satisfies RIP with parameters $\delta, \eta \in (0,1)$, a vector $x \in \mathbb{R}^d$ and a number of iterations $r$.
Initialize $a = 0 \in \mathbb{R}^D$, $M = \|x\|_2/\sqrt{\delta D}$.
repeat $r$ times
$b = U^\top x$
$\hat{b} = \mathrm{sign}(b) \cdot \min(|b|, M)$
$x = x - U\hat{b}$
$a = a + \hat{b}$
$M = \eta M$
return $a$
Output: Kashin's coefficients of $x$ with level $K = 1/(\sqrt{\delta}(1-\eta))$ and with accuracy $\eta^r \|x\|_2$, i.e.
```

$$\|x - Ua\|_2 \le \eta^r \|x\|_2, \qquad \max_{1 \le i \le D} |a_i| \le \frac{K}{\sqrt{D}} \|x\|_2.$$

For this reason, it is preferable to use *tight frames* and *frame representations* instead. Tight frames are generalizations of orthonormal bases, where the system of vectors are not required to be linearly independent. Formally, vectors $(u_i)_{i=1}^D$ in $\mathbb{R}^d$ form a tight frame if any vector $x \in \mathbb{R}^d$ admits a frame representation

$$x = \sum_{i=1}^D a_i u_i, \qquad a_i = \langle x, u_i \rangle . \tag{7}$$

Clearly, if $D > d$ (the case we are interested in), then the system $(u_i)_{i=1}^D$ is linearly dependent and hence the representation (7) with coefficients a_i is not unique. The idea is to exploit this redundancy and choose coefficients a_i in such a way to spread the information uniformly among these coefficients. However, the frame representation may not distribute the information well enough. Thus, we need a particular representation for which coefficients a_i have smallest possible dynamic range.

For a frame $(u_i)_{i=1}^D$ define the $d \times D$ *frame matrix* U by stacking frame vectors u_i as columns. It can be easily seen that being a tight frame is equivalent to frame matrix to be orthogonal, i.e. $UU^\top = I_d$, where I_d is the $d \times d$ identity matrix. Using the frame matrix U, frame representation (7) takes the form $x = Ua$.

Definition 3 (Kashin's representation) *Let $(u_i)_{i=1}^D$ be a tight frame in $\mathbb{R}^d$. Define Kashin's representation of $x \in \mathbb{R}^d$ with level K the following expansion*

$$x = \sum_{i=1}^D a_i u_i, \qquad \max_{1 \le i \le D} |a_i| \le \frac{K}{\sqrt{D}} \|x\|_2. \tag{8}$$

Optimality. As noted in (Lyubarskii & Vershynin, 2010), Kashin's representation has the smallest possible dynamic range $K/\sqrt{D}$, which is $\sqrt{d}$ times smaller then dynamic range of the frame representation (7).

Existence. It turns out that not every tight frame can guarantee Kashin's representation with constant level. The following existence result is based on Kashin's theorem (Kashin, 1977):

Theorem 3 *There exist tight frames in $\mathbb{R}^d$ with arbitrarily small redundancy $\lambda = D/d > 1$, and such that every vector $x \in \mathbb{R}^d$ admits Kashin's representation with level $K = K(\lambda)$ that depends on λ only (not on d or D).*

4.2 Computing Kashin's representation

To compute Kashin's representation we use the algorithm developed by Lyubarskii & Vershynin (2010), which transforms the frame representation (7) into Kashin's representation (8). The algorithm requires tight frame with frame matrix satisfying the restricted isometry property:

Definition 4 (Restricted Isometry Property (RIP)) *A given $d \times D$ matrix U satisfies the Restricted Isometry Property with parameters $\delta, \eta \in (0,1)$ if for any $x \in \mathbb{R}^d$*

$$|\mathrm{supp}(x)| \le \delta D \quad \Rightarrow \quad \|Ux\|_2 \le \eta \|x\|_2. \tag{9}$$

In general, for an orthogonal $d \times D$ matrix U we can only guarantee the inequality $\|Ux\|_2 \le \|x\|_2$ if $x \in \mathbb{R}^d$. The RIP requires U to be a contraction mapping for sparse x. With a frame matrix satisfying RIP, the analysis of Algorithm 1 from (Lyubarskii & Vershynin, 2010) yields a formula for the level of Kashin's representation:

Theorem 4 *Let $(u_i)_{i=1}^D$ be a tight frame in $\mathbb{R}^d$ which satisfies RIP with parameters δ, η. Then any vector $x \in \mathbb{R}^d$ admits a Kashin's representation with level*

$$K = \frac{1}{\sqrt{\delta}(1-\eta)}. \tag{10}$$

4.3 Quantizing Kashin's representation

We utilize Kashin's representation to design a compression method, which will enjoy dimension-free variance bound on the approximation error. Let $x \in \mathbb{R}^d$ be the vector that we want to communicate and $\lambda > 1$ be the redundancy factor so that $D = \lambda d$ is positive integer. First we find Kashin's representation of x, i.e. $x = Ua$ for some $a \in \mathbb{R}^D$, and then quantize coefficients a_i using any unbiased compression operator $\mathcal{C} \colon \mathbb{R}^D \to \mathbb{R}^D$ that preserves the sign and maximum magnitude:

$$0 \le \mathcal{C}(a)\,\mathrm{sign}(a) \le \|a\|_\infty, \quad a \in \mathbb{R}^D. \tag{11}$$

For example, ternary quantization or any dithering (standard random, natural) can be applied. The vector that we communicate is the quantized coefficients $\mathcal{C}(a) \in \mathbb{R}^D$ and KC is defined via

$$\mathcal{C}_\kappa(x) = U\mathcal{C}(a).$$

Due to unbiasedness of $\mathcal{C}$ and linearity of expectation, we preserve unbiasedness for $\mathcal{C}_\kappa$:

$$\mathbb{E}[\mathcal{C}_\kappa(x)] = \mathbb{E}\left[U\mathcal{C}(a)\right] = U\mathbb{E}\left[\mathcal{C}(a)\right] = Ua = x.$$

Then we bound the error of approximation uniformly (without the expectation) as follows

$$\begin{aligned} \|\mathcal{C}_\kappa(x) - x\|_2^2 = \|U\mathcal{C}(a) - Ua\|_2^2 &\le \|\mathcal{C}(a) - a\|_2^2 \\ &\le D \max_{1 \le i \le D} (\mathcal{C}(a)_i - a_i)^2 \le D\|a\|_\infty^2 \le D\left(\frac{K(\lambda)}{\sqrt{D}}\|x\|_2\right)^2 = K^2(\lambda)\|x\|_2^2. \end{aligned}$$

The obtained uniform upper bound $K(\lambda)^2$ does not depend on the dimension d. It depends only on the redundancy factor $\lambda > 1$ which should be chosen depending on how less we want to communicate. Thus, KC $\mathcal{C}_\kappa$ with any unbiased quantization (11) belongs to $\mathbb{U}\left(K^2(\lambda)\right)$. Note, that we are not restrained to use only unbiased compressions with Kashin's representation. For instance, instead of random sparsification (which is unbiased and satisfies (11)) one can use Top-k sparsification, which satisfies (11) and in practice works much better despite having similar theoretical properties.

5 Measure concentration and orthogonal matrices

The concentration of the measure is a remarkable high-dimensional phenomenon which roughly claims that a function defined on a high-dimensional space and having small oscillations takes values highly concentrated around the average (Ledoux, 2001; Giannopoulos & Milman, 2000). Here we present one example of such concentration for Lipschitz functions on the unit sphere, which will be the key to justify the restricted isometry property.

5.1 Concentration on the sphere for Lipschitz functions

Let $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \colon \|x\|_2 = 1\}$ be the unit sphere. We say that $f \colon \mathbb{S}^{d-1} \to \mathbb{R}$ is a Lipschitz function with constant L if

$$|f(x) - f(y)| \le L\|x - y\|_2$$

for any x, $y \in \mathbb{S}^{d-1}$.

Theorem 5 *Let $X \in \mathbb{S}^{d-1}$ be a random vector uniformly distributed on the unit Euclidean sphere. If $f \colon \mathbb{S}^{d-1} \to \mathbb{R}$ is $L-$Lipschitz function, then for any $t \ge 0$*

$$\operatorname{Prob}\left(|f(X) - \mathbb{E}f(X)| \ge t\right) \le 5\exp\left(-\frac{(d-2)t^2}{8L^2}\right).$$

Informally and rather surprisingly, Lipschitz functions on a high-dimensional unit sphere are almost constants. Particularly, it implies that deviations of function values from the average are at most ${}^{8L}/\sqrt{d}$ with confidence level more than 0.99. We will apply this concentration inequality for the function $x \to \|Ux\|_2$ which is $1-$Lipschitz if U is orthogonal.

5.2 Random orthogonal matrices

Up to this point we did not discuss how to choose the frame vectors u_i or the frame matrix U, which is used in the construction of Kashin's representation. We only know that it should be orthogonal and satisfy RIP for some parameters δ, η. We now describe how to construct frame matrix U and how to estimate parameters δ, η. Unluckily, there is no an explicit construction scheme for such matrices. There are random generation processes that provide probabilistic guarantees (Candès & Tao, 2005, 2006; Lyubarskii & Vershynin, 2010).

Consider random $d \times D$ matrices with orthonormal rows. Such matrices are obtained from selecting the first d rows of orthogonal $D \times D$ matrices. Let $O(D)$ be the space of all orthogonal $D \times D$ matrices with the unique translation invariance and normalized measure, which is called Haar measure for that space. Then the space of $d \times D$ orthogonal matrices is

$$O(d \times D) = \{U = P_d V \colon V \in O(D)\},$$

where $P_d \colon \mathbb{R}^D \to \mathbb{R}^d$ is the orthogonal projection on the first d coordinates. The probability measure on $O(d \times D)$ is induces by the Haar measure on $O(D)$. Next we show that, with respect to the normalized Haar measure, randomly generated orthogonal matrices satisfy RIP with high probability.

Theorem 6 *Let $\lambda > 1$ and $D = \lambda d$, then with probability at least*

$$1 - 5\exp\left[-d\left(\sqrt{\lambda} - 1\right)^2\left(\frac{1}{26} + \frac{1}{208}\log\left(1 - \frac{1}{\sqrt{\lambda}}\right)\right)\right],$$

a random orthogonal $d \times D$ matrix U satisfies RIP with parameters

$$\eta = \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{\sqrt{\lambda}}, \qquad \delta = \frac{1}{5^4}\left(1 - \frac{1}{\sqrt{\lambda}}\right)^2. \tag{12}$$

Note that the expression for the probability can be negative if λ is too close to 1. Specifically, the logarithmic term vanishes for $\lambda \approx 1.0005$ giving negative probability. However, the probability approaches to 1 quite rapidly for bigger λ's. To get a sense of how high that probability can be, note that for $d = 1000$ variables and $\lambda = 2$ inflation it is bigger than 0.98.

Now that we have explicit formulas for the parameters δ and η, we can combine it with the results of Section 4 and summarize with the following theorem.

Theorem 7 *Let $\lambda > 1$ be the redundancy factor and $\mathcal{C}$ be any unbiased compression operator satisfying (11). Then Kashin Compression $\mathcal{C}_\kappa \in \mathbb{U}(\omega_\lambda)$ is an unbiased compression with dimension independent variance*

$$\omega_\lambda = \left(\frac{10\sqrt{\lambda}}{\sqrt{\lambda} - 1}\right)^4. \tag{13}$$

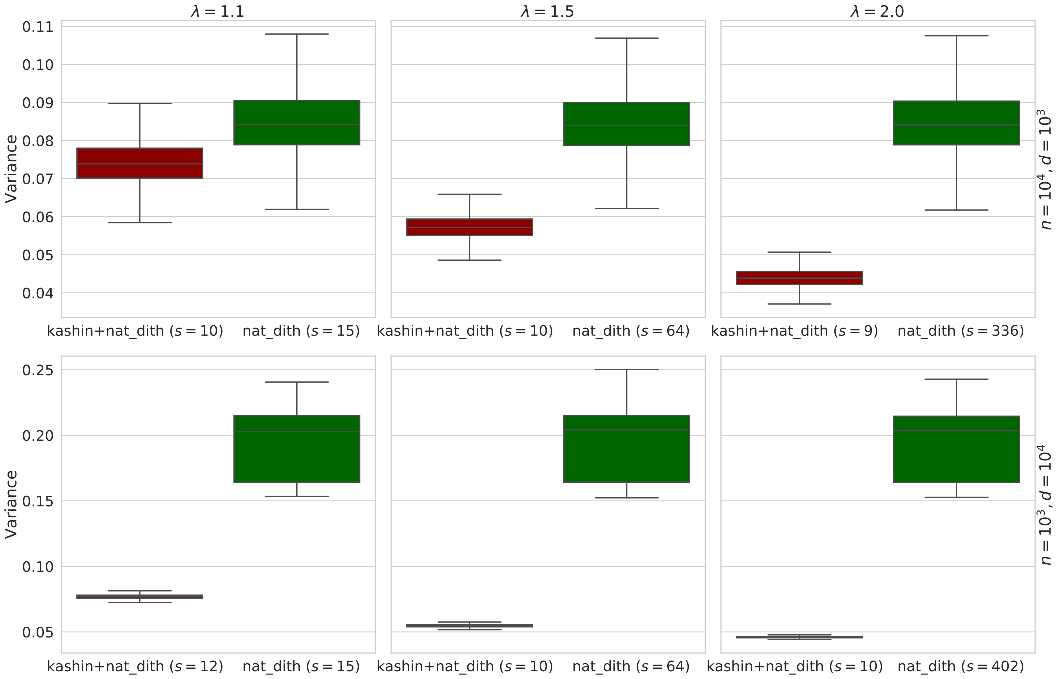


Figure 2: Comparison of empirical variances (14) of natural dithering and KC with natual dithering.

6 Experiments

In this section we describe the implementation details of KC and present our experiments of KC compared to other popular compression methods in the literature.

6.1 Implementation details of KC

To generate a random (fat) orthogonal frame matrix U, we first generate a random matrix with entries drown independently from Gaussian distribution. Then we extract an orthogonal matrix by applying QR decomposition. Note that, for big dimensions the generation process of frame matrix U becomes computationally expensive. However, after fixing the dimension of to-be-compressed vectors then the frame matrix needs to be generated only once and can be used throughout the learning process.

Afterwards, we turn to the estimation of the parameters δ and η of RIP, which are necessary to compute Kashin's representations. These parameters are estimated iteratively so to minimize the representation level K (10) subject to the constraint (9) of RIP. For fixed δ we first find the least η such 9 holds for unit vectors, which were obtained by normalizing Gaussian random vectors (we chose sample size of $10^4 - 10^5$, which provided a good estimate). Then we tune the parameter δ (initially chosen 0.9) to minimize the level K (10).

6.2 Empirical variance comparison

We empirically compare the variance produced by natural dithering against KC with natural dithering and observe that latter introduces much less variance. We generated n vectors with d independent entries from standard Gaussian distribution. Then we fix the minimum number of levels s that allows obtaining an acceptable variance for performing KC with natural dithering. Next, we adjust levels s for natural dithering to the almost same number of bits used for transmission of the compressed vector. For each of these vectors we compute normalized empirical variance via

$$\omega(x) := \frac{\|\mathcal{C}(x) - x\|^2}{\|x\|^2}. \tag{14}$$

In Figure 2 we provide boxplots for empirical variances, which show that the increase of parameter λ leads to smaller variance for KC. They also confirm that for natural dithering, the variance ω scales with the dimension d

while for KC that scaling is significantly reduced (see also Table 1 for variance bounds). This shows the positive effect of KC combined with other compression methods. For additional insights, we present also swarmplots provided by Seaborn Library. Figure 3 illustrates the strong robustness property of KC with respect to outliers.

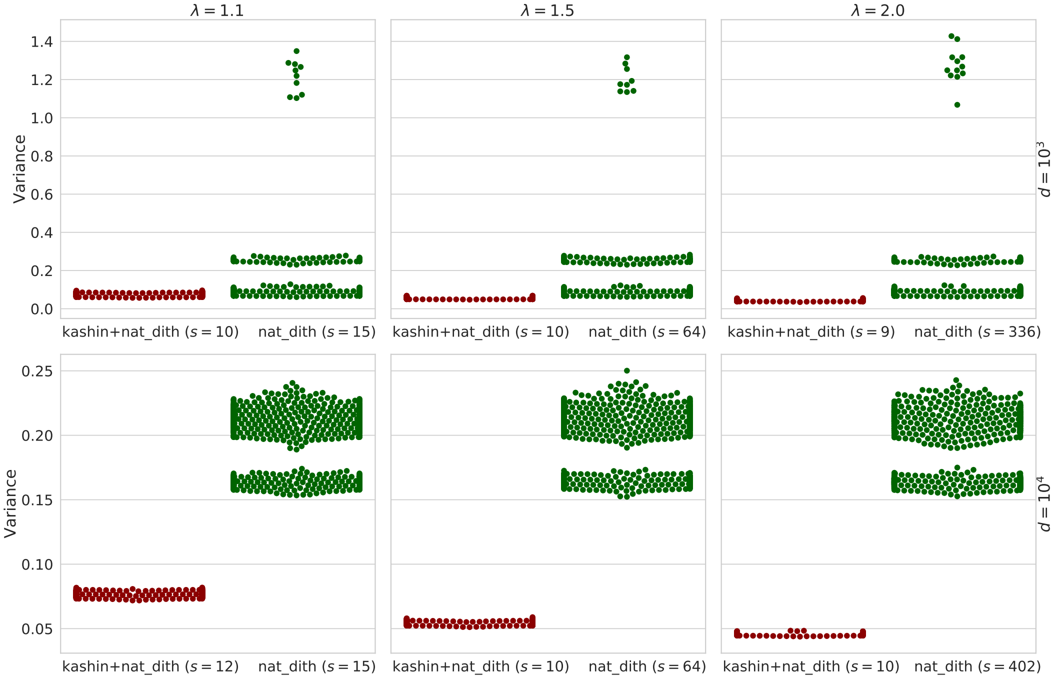


Figure 3: Swarmplots (with sub-sample size $n = 1000$) of empirical variances (14) for natural dithering and KC with natural dithering.

6.3 Minimizing quadratics with CGD

To illustrate the advantages of KC in optimization algorithms, we minimized randomly generated quadratic functions (15) for $d = 10^4$ using gradient descent with compressed gradients.

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \frac{1}{2} x^\top A x - b^\top x, \tag{15}$$

In Figure 4a we evaluate functional suboptimality

$$\frac{f(x_k) - f^*}{f(x_0) - f^*}$$

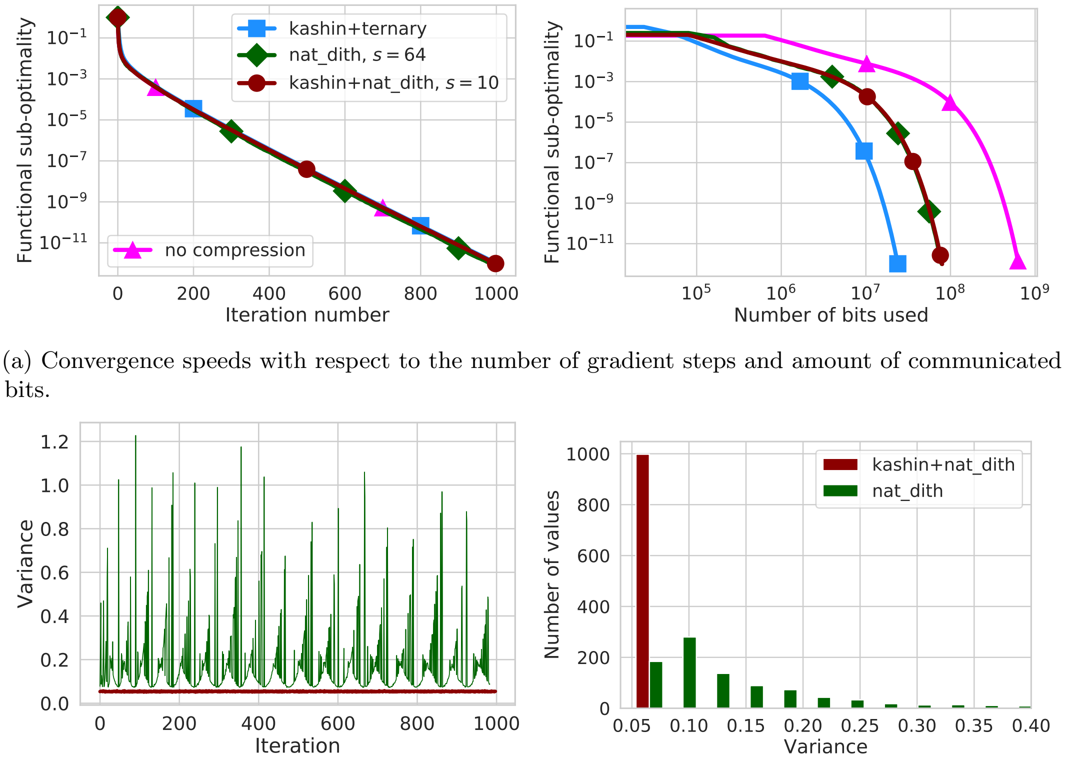
in log-scale for vertical axis. These plots illustrate the superiority of KC with ternary quantization, where it does not degrade the convergence at all and saves in communication compared to other compression methods and without any compression scheme.

To provide more insights into this setting, Figure 4b visualizes empirical variances of the compressed gradients throughout the optimization process, revealing both the low variance feature and the stabilization property of KC.

6.4 Minimizing quadratics with distributed CGD

Consider the minimization problem of the average of n quadratics

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x), \qquad \text{where} \qquad f_i(x) = \frac{1}{2} x^\top A_i x, \tag{16}$$



(a) Convergence speeds with respect to the number of gradient steps and amount of communicated bits.

(b) Empirical variances of compressed gradients throughout the optimization process.

Figure 4: Performance of different compression methods during the minimization of quadratics (15). Hyperparameters of compression operators (λ for KC and s for natural dithering) were chosen in such a way so to have either identical function suboptimalities (4a) or an identical number of compressed bits (4b).

with synthetically generated matrices A_i. We solve this problem with Distributed Compressed Gradient Descent (Algorithm 2) using a selection of compression operators.

Figures 5 and 6 show that KC combined with ternary quantization leads to faster convergence and uses less bits to communicate than ternary quantization alone. Note that in higher dimension the gap between KC with ternary quantization and no compression gets smaller in the iteration plot, while in the communication plot it gets bigger. So, in high dimensions KC convergences slightly worse than no compression scheme, but the savings in communication are huge.

7 Conclusion and future plans

We formalized, for the first time, the limitation of (randomized) compression operators in communication and mathematically proved an uncertainty principle for communication compression. We also presented a highly robust new—Kashin compressor (KC)—and showed that in combinations with some other compression methods gives almost optimal compression, thus closing the gap established by our uncertainty principle. As a future work, we plan to implement a sparse and efficient generation of large-size random orthogonal matrices using block structured small-size orthogonal matrices. This should reduce both the storage requirement and the computational effort to use KC in practical applications.

Algorithm 2 Distributed Compressed Gradient Descent (DCGD)

```
**Input:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, compression operator $\mathcal{C} \in \mathbb{B}(\alpha)$.
**for** $k = 0, 1, 2, \dots$ **do**
  **for all nodes** $i \in \{1, 2, \dots, n\}$ **in parallel do**
    Compute local gradient $\nabla f_i\left(x^k\right)$
    Compress local gradient $g_i^k = \mathcal{C}\left(\nabla f_i\left(x^k\right)\right)$
    Receive the aggregate $g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$
    $x^{k+1} = x^k - \gamma g^k$
```

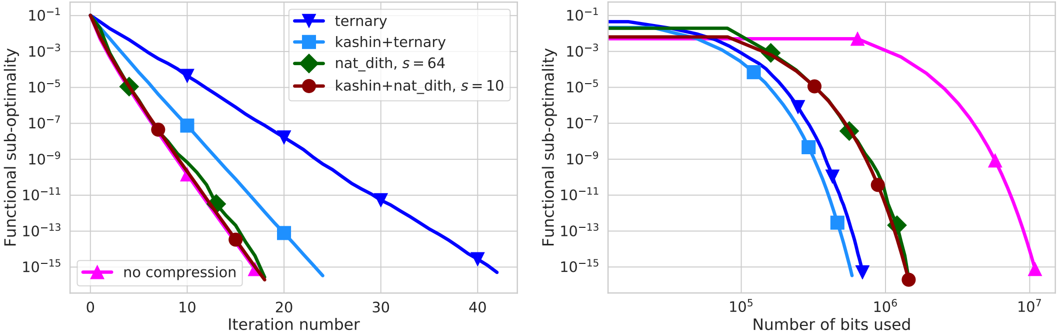


Figure 5: Performance of Distributed Compressed Gradient Descent (Algorithm 2 with different compression operators for problem (16) with $n = 10$ workers and $d = 10^3$.

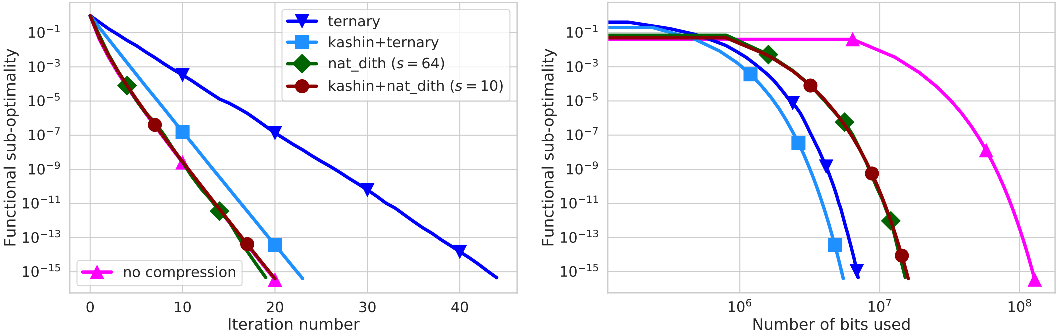


Figure 6: Performance of Distributed Compressed Gradient Descent (Algorithm 2) with different compression operators for problem (16) with $n = 10$ workers and $d = 10^4$.

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30*, pp. 1709–1720, 2017.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pp. 5977–5987, 2018.

Bekkerman, R., Bilenko, M., and Langford, J. *Scaling up machine learning: Parallel and distributed approaches.* Cambridge University Press, 2011.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 560–569. PMLR, 2018.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Caldas, S., Konečný, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. In *arXiv preprint arXiv:1812.07210v2*, 2019.

Candès, E. J. and Tao, T. Decoding by linear programming. In *IEEE Transactions on Information Theory*, volume 51, 2005.

Candès, E. J. and Tao, T. Near-optimal signal recovery from random projections and universal encoding strategies. In *IEEE Transactions on Information Theory*, volume 52, 2006.

Gabor, D. Theory of communication. *Journal of the Institute of Electrical Engineering*, 93:429–457, 1946.

Giannopoulos, A. A. and Milman, V. Concentration property on probability spaces. *Advances in Mathematics*, 156: 77–106, 2000.

Goodall, W. M. Television by pulse code modulation. *The Bell System Technical Journal*, 30(1):33–49, Jan 1951. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1951.tb01365.x.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: training ImageNet in 1 hour. *CoRR*, abs/1706.02677, 2017.

Havin, V. and Jöricke, B. *The Uncertainty Principle in Harmonic Analysis.* Springer-Verlag, 1994.

Heisenberg, W. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3–4):172–198, 1927.

Horváth, S., Ho, C.-Y., Horváth, L., Sahu, A. N., Canini, M., and Richtárik, P. Natural compression for distributed deep learning. *arXiv:1905.10988*, 2019a.

Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019b.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Scaffold: Stochastic controlled averaging for on-device federated learning. *ArXiv*, abs/1910.06378, 2019a.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3252–3261, 2019b.

Kashin, B. S. Diameters of some finite-dimensional sets and classes of smooth functions. *Jour. Izv. Akad. Nauk SSSR Ser. Mat.*, 41(2):334–351, 1977. URL `http://mi.mathnet.ru/izv1805`.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. Distributed learning with compressed gradients. In *arXiv preprint arXiv:1806.06573*, 2018.

Kochol, M. Constructive approximation of a ball by polytopes. *Mathematica Slovaca*, 44(1):99–105, 1994. ISSN 0139-9918. URL `https://eudml.org/doc/34376`.

Kochol, M. A note on approximation of a ball by polytopes. *Discrete Optimization*, 1(2):229 – 231, 2004. ISSN 1572-5286. doi: https://doi.org/10.1016/j.disopt.2004.07.003. URL `http://www.sciencedirect.com/science/article/pii/S1572528604000295`.

Konečný, J. and Richtárik, P. Randomized distributed mean estimation: accuracy vs communication. *Frontiers in Applied Mathematics and Statistics*, 4(62):1–11, 2018.

Konečný, J., McMahan, H. B., Yu, F., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

Ledoux, M. *The Concentration of Measure Phenomenon.* American Mathematical Society, 2001.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

Liu, S., Chen, P.-Y., Chen, X., and Hong, M. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

Lyubarskii, Y. and Vershynin, R. Uncertainty principles and vector quantization. *IEEE Trans. Inf. Theor.*, 56(7): 3491–3501, July 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2048458. URL `http://dx.doi.org/10.1109/TIT.2010.2048458`.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Roberts, L. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, February 1962. ISSN 0096-1000. doi: 10.1109/TIT.1962.1057702.

Schmidhuber, J. Deep learning in neural networks: An overview. In *Neural networks*, volume 61, pp. 85117, 2015.

Stich, S. U. Local SGD converges fast and communicates little. In *CoRR, abs/1805.09767*, 2018.

Tang, H., Yu, C., Lian, X., Zhang, T., and Liu, J. `DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6155–6165, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/tang19d.html`.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pp. 1195–1204, 2019.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Vogels, T., Karimireddy, S. P., and Jaggi, M. PowerSGD: Practical low-rank gradient compression for distributed optimization. *CoRR*, abs/1905.13727, 2019. URL `http://arxiv.org/abs/1905.13727`.

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1306–1316, 2018.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pp. 1509–1519, 2017.

Zhang, H., Li, J., Kara, K., Alistarh, D., Liu, J., and Zhang, C. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 40354043, 2017.

Appendix

A Proofs for Section 2

A.1 Proof of Theorem 1: UP for biased compressions $\mathbb{B}(\alpha)$

Fix $R > 0$ and let $B^d(R)$ be the d-dimensional Euclidean closed ball with center at the origin and with radius R. Denote by $m = 2^b$ the number of possible outcomes of compression operator $\mathcal{C}$ and by $\{v_1, \dots, v_m\} \subset \mathbb{R}^d$ the set of compressed vectors. We relax the α-contractive requirement and prove (3) in the case when the restricted compression operator $\mathcal{C}\colon B^d(R) \to \{v_1, \dots, v_m\}$ satisfies

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \le \alpha R^2, \quad x \in B^d(R). \tag{17}$$

Define probability functions p_k as follows

$$p_k(x) = \mathrm{Prob}\left(\mathcal{C}(x) = v_k\right), \quad x \in B^d(R), \quad k \in [m].$$

Then we stack functions p_k together and get a vector valued function $p\colon B^d(R) \to \Delta^m$, where Δ^m is the standard m-simplex

$$\Delta^m = \left\{ (p_1, p_2, \dots, p_m) \in \mathbb{R}^m \colon \sum_{k=1}^m p_k = 1,\, p_k \ge 0 \text{ for all } k \in [m] \right\}.$$

We can express the expectation in (17) as

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] = \sum_{k=1}^m p_k(x) \|v_k - x\|^2 \tag{18}$$

and taking into account the inequality (17) itself, we conclude

$$\max_{x \in B^d(R)} \sum_{k=1}^m p_k(x) \|v_k - x\|^2 \le \alpha R^2.$$

The above inequality holds for the particular probability function p defined from the compression $\mathcal{C}$. Therefore the inequality will remain valid if we take the minimum of left hand side over all possible probability functions $\hat{p}\colon B^d(R) \to \Delta^m$:

$$\min_{\hat{p}\colon B^d(R) \to \Delta^m} \max_{x \in B^d(R)} \sum_{k=1}^m \hat{p}_k(x) \|v_k - x\|^2 \le \alpha R^2. \tag{19}$$

We then swap the order of min-max by adjusting domains properly:

$$\min_{\hat{p}\colon B^d(R) \to \Delta^m} \max_{x \in B^d(R)} \sum_{k=1}^m \hat{p}_k(x) \|v_k - x\|^2 = \max_{x \in B^d(R)} \min_{\hat{p} \in \Delta^m} \sum_{k=1}^m \hat{p}_k \|v_k - x\|^2,$$

where the second minimum is over all probability vectors $\hat{p} \in \Delta^m$ (not over vector valued functions as in the first minimum). Next, notice that

$$\min_{\hat{p} \in \Delta^m} \sum_{k=1}^m \hat{p}_k \|v_k - x\|^2 = \|v_x - x\|^2,$$

where $v_x \in \arg\min_{v\in\{v_1,\dots,v_m\}} \|v-x\|^2$ is the closest v_k to x. Therefore, we have transformed (19) into

$$\max_{x\in B^d(R)} \|v_x - x\| \le R\sqrt{\alpha} =: \hat{R}.$$

The last inequality means that the set $\{v_1,\dots,v_m\}$ is an $\hat{R}$-net for the ball $B^d(R)$. Using the following result on covering numbers and volume (see Proposition 4.2.12, (Vershynin, 2018)) we conclude

$$m = \#\{v_1,\dots,v_m\} \ge \frac{\mathrm{vol}(B^d(R))}{\mathrm{vol}(B^d(\hat{R}))} = \frac{R^d}{\hat{R}^d} = \alpha^{-d/2},$$

which completes the proof since

$$\alpha \cdot 4^{b/d} = \alpha \cdot m^{2/d} \ge 1.$$

A.2 Proof of Lemma 1

Let $\mathcal{C} \in \mathbb{U}(\omega)$. Using relations $\mathbb{E}\left[\mathcal{C}(x)\right] = x$ and $\mathbb{E}\left[\|\mathcal{C}(x)\|^2\right] \le (\omega+1)\|x\|^2$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\omega+1}\mathcal{C}(x) - x\right\|^2\right] &= \frac{1}{(\omega+1)^2}\mathbb{E}\left[\|\mathcal{C}(x)\|^2\right] - \frac{2}{\omega+1}\mathbb{E}\left[\langle\mathcal{C}(x), x\rangle\right] + \|x\|^2 \\
&= \frac{1}{(\omega+1)^2}\mathbb{E}\left[\|\mathcal{C}(x)\|^2\right] + \left(-\frac{2}{\omega+1} + 1\right)\|x\|^2 \\
&\le \left(\frac{1}{\omega+1} - \frac{2}{\omega+1} + 1\right)\|x\|^2 = \frac{\omega}{\omega+1}\|x\|^2,
\end{aligned}$$

which concludes the lemma.

B Proofs for Section 5

B.1 Proof of Theorem 5: Concentration on the sphere for Lipschitz functions

Let $\mathbb{S}^{d-1}$ be the unit sphere with the normalized Lebesgue measure μ and the geodesic metric $\mathrm{dist}(x,y) = \arccos\langle x,y\rangle$ representing the angle between x and y. Using this metric, we define the spherical caps as the balls in $\mathbb{S}^{d-1}$:

$$B_a(r) = \{x \in \mathbb{S}^{n-1} \colon\ \mathrm{dist}(x,a) \le r\}, \quad a \in \mathbb{S}^{d-1},\ r > 0.$$

For a set $A \subset \mathbb{S}^{d-1}$ and non-negative number $t \ge 0$ denote by $A(t)$ the t-neighborhood of A with respect to geodesic metric:

$$A(t) = \left\{x \in \mathbb{S}^{d-1} \colon\ \mathrm{dist}(x,A) \le t\right\}.$$

The famous result of P. Levy on isoperimetric inequality for the sphere states that among all subsets $A \subset \mathbb{S}^{d-1}$ of a given measure, the spherical cap has the smallest measure for the neighborhood (see e.g. (Ledoux, 2001)).

Theorem 8 (Levy's isoperimetric inequality) *Let $A \subset \mathbb{S}^{d-1}$ be a closed set and let $t \ge 0$. If $B = B_a(r)$ is a spherical cap with $\mu(A) = \mu(B)$, then*

$$\mu\left(A(t)\right) \ge \mu\left(B(t)\right) \equiv \mu(B_a(r+t)).$$

We also need the following upper bound on the measure of spherical caps[2].

Lemma 2 *Let $t \ge 0$. If $B \subset \mathbb{S}^{d-1}$ is a spherical cap with radius $\pi/2 - t$, then*

$$\mu(B) \le \sqrt{\frac{\pi}{8}}\exp\left(-\frac{(d-2)t^2}{2}\right). \tag{20}$$

[2] https://en.wikipedia.org/wiki/Spherical_measure

These two results yield a concentration inequality on the unit sphere around median of the Lipschitz function.

Theorem 9 *Let $f\colon \mathbb{S}^{d-1} \to \mathbb{R}$ be a L-Lipschitz function (w.r.t. geodesic metric*[3]*) and let $M = M_f$ be its median, i.e.*

$$\mu\{x\colon f(x) \geq M\} \geq \frac{1}{2} \quad \text{and} \quad \mu\{x\colon f(x) \leq M\} \geq \frac{1}{2}.$$

Then, for any $t \geq 0$

$$\mu\{x\colon |f(x) - M| \geq t\} \leq \sqrt{\frac{\pi}{2}} \exp\left(-\frac{(d-2)t^2}{2L^2}\right). \tag{21}$$

B.1.1 Proof of Theorem 9: Concentration around the median

Without loss of generality we can assume that $L = 1$. Denote

$$A_+ = \{x\colon f(x) \geq M\} \quad \text{and} \quad A_- = \{x\colon f(x) \leq M\},$$

so that $\mu(A_\pm) \geq 1/2 = \mu(B_a(\pi/2))$ for some $a \in \mathbb{S}^{d-1}$. Then the isoperimetric inequality (21) and the upper bound (20) imply

$$\begin{aligned}\mu(A^c_\pm(t)) = \mu\{x\colon \operatorname{dist}(x, A_\pm) > t\} &\leq \mu\{x\colon \operatorname{dist}(x, B_a(\pi/2)) > t\} \\ &= \mu(B_a(\pi/2 - t)) \\ &\leq \sqrt{\frac{\pi}{8}} \exp\left(-\frac{(d-2)t^2}{2}\right).\end{aligned}$$

Note that $x \in A_-(t)$ implies that $\operatorname{dist}(x, y) \leq t$, $f(y) \leq M$ for some $y \in A_-$. Using the Lipschitzness of f we get $f(x) \leq f(y) + \operatorname{dist}(x, y) \leq M + t$. Analogously, $x \in A_+(t)$ implies that $\operatorname{dist}(x, y) \leq t$, $f(y) \geq M$ for some $y \in A_+$. Again, the Lipschitzness of f gives $-f(x) \leq -f(y) + \operatorname{dist}(x, y) \leq -M + t$. Thus

$$|f(x) - M| \leq t \quad \text{for any} \quad x \in A_+(t) \cap A_-(t).$$

To complete the proof, it remains to combine this with inequalities for measures of complements

$$\begin{aligned}\mu\left(\{x\colon |f(x) - M| > t\}\right) &= 1 - \mu\left(\{x\colon |f(x) - M| \leq t\}\right) \\ &\leq 1 - \mu(A_+(t) \cap A_-(t)) \\ &\leq \mu(A^c_+(t)) + \mu(A^c_-(t)) \leq \sqrt{\frac{\pi}{2}} \exp\left(-\frac{(d-2)t^2}{2}\right).\end{aligned}$$

Continuity of μ and f give the result with the relaxed inequality.

B.1.2 Proof of Theorem 5: Concentration around the mean

Now, from (21) we derive a concentration inequality around the mean rather than median, where mean is defined via

$$\mathbb{E}f = \int_{\mathbb{S}^{d-1}} f(x)\, d\mu(x).$$

Again, without loss of generality we assume that $L = 1$ and $d \geq 3$. Fix $\epsilon \in [0, 1]$ and decompose the set $\{x\colon |f(x) - \mathbb{E}f| \geq t\}$ into two parts:

$$\mu\left(\{x\colon |f(x) - \mathbb{E}f| \geq t\}\right) \leq \mu\left(\{x\colon |f(x) - M| \geq \epsilon t\}\right) + \mu\left(\{x\colon |\mathbb{E}f - M| \geq (1-\epsilon)t\}\right) =: A_1 + A_2,$$

where M is a median of f. From the concentration (21) around the median, we get an estimate for A_1

$$A_1 \leq \sqrt{\frac{\pi}{2}} \exp\left(-\frac{(d-2)t^2\epsilon^2}{2}\right).$$

[3] notice that Lipschitzness w.r.t. geodesic metric is weaker than w.r.t. Euclidean metric. This implies that the obtained concentration holds for L-Lipschitz function w.r.t. standard Euclidean distance.

Now we want to estimate the second term A_2 with a similar upper bound so to combine them. Obviously, the condition in A_2 does not depend on x, and it is a piecewise constant function of t. Therefore

$$\begin{aligned}A_2 &\le \mu\left(\{x\colon \mathbb{E}|f-M| \ge (1-\epsilon)t\}\right) = \mu\left(\{x\colon \|f-M\|_1 \ge (1-\epsilon)t\}\right)\\ &= \begin{cases}1 & \text{if} \quad t \le {}^1\!/\!_{(1-\epsilon)}\|f-M\|_1\\ 0 & \text{otherwise}\end{cases} \le \begin{cases}1 & \text{if} \quad t \le \frac{\pi}{2(1-\epsilon)\sqrt{d-2}}\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

where we bounded $\|f-M\|_1$ as follows

$$\begin{aligned}\|f-M\|_1 &= \int_0^\infty \mu\left(\{x\colon |f(x)-M| \ge u\}\right)\, du\\ &\le \sqrt{\frac{\pi}{2}}\int_0^\infty \exp\left(-\frac{(d-2)u^2}{2}\right)\, du\\ &= \sqrt{\frac{\pi}{d-2}}\int_0^\infty \exp(-u^2)\, du\\ &= \sqrt{\frac{\pi}{d-2}}\frac{\sqrt{\pi}}{2} = \frac{\pi}{2\sqrt{d-2}}.\end{aligned}$$

We further upper bound A_2 to get the same exponential term as for A_1:

$$A_2 \le \begin{cases}1 & \text{if} \quad t \le \frac{\pi}{2(1-\epsilon)\sqrt{d-2}}\\ 0 & \text{otherwise}\end{cases} \le \exp\left[\frac{\pi^2}{8}\frac{\epsilon^2}{(1-\epsilon)^2}\right]\exp\left(-\frac{(d-2)t^2\epsilon^2}{2}\right). \tag{22}$$

To check the validity of the latter upper bound, first notice that for $t = \frac{\pi}{2(1-\epsilon)\sqrt{d-2}}$ both are equal to 1. Then, the monotonicity and positiveness of the exponential function imply (22) for $0 \le t < \frac{\pi}{2(1-\epsilon)\sqrt{d-2}}$ and $t > \frac{\pi}{2(1-\epsilon)\sqrt{d-2}}$. Combining these two upper bounds for A_1 and A_2, we get

$$A_1 + A_2 \le \left(\exp\left[\frac{\pi^2}{8}\frac{\epsilon^2}{(1-\epsilon)^2}\right] + \sqrt{\frac{\pi}{2}}\right)\exp\left(-\frac{(d-2)t^2\epsilon^2}{2}\right) \le 5\exp\left(-\frac{(d-2)t^2}{8}\right)$$

if we set $\epsilon = {}^1\!/\!_2$. To conclude the theorem, note that normalized uniform measure μ on the unit sphere can be seen as a probability measure on $\mathbb{S}^{d-1}$.

B.2 Proof of Theorem 6: Random orthogonal matrices with RIP

The proof follows the steps of the proof of Theorem 4.1 of Lyubarskii & Vershynin (2010). First, we relax the inequality in Theorem 5 to

$$\text{Prob}\left(|f(X) - \mathbb{E}f(X)| \ge t\right) \le 5\exp\left(-\frac{d\,t^2}{9L^2}\right), \quad t \ge 0,\, d \ge 20. \tag{23}$$

Let $x \in \mathbb{S}^{D-1}$ be fixed. Any orthogonal $d \times D$ matrix $U \in O(d \times D)$ can be represented as the projection $U = P_d V$ of $D \times D$ orthogonal matrix $V \in O(D)$. The uniform probability measure (or Haar measure) on $O(D)$ ensures that if $V \in O(D)$ is random then the vector $z = Vx$ is uniformly distributed on $\mathbb{S}^{D-1}$. Therefore, if $U \in O(d \times D)$ is random with respect to the induced Haar measure on $O(d \times D)$, then random vectors Ux and $P_d z$ have identical distributions. Denote $f(z) = \|P_d z\|_2$ and notice that f is 1-Lipschitz on the sphere $\mathbb{S}^{D-1}$. To apply the concentration inequality (23), we compute the expected norm of these random vectors:

$$\mathbb{E}f(z) = \int_{\mathbb{S}^{D-1}} \|P_d z\|_2\, d\mu(z) \quad \le \quad \left(\int_{\mathbb{S}^{D-1}} \|P_d z\|_2^2\, d\mu(z)\right)^{1/2} = \left(\sum_{i=1}^d \int_{\mathbb{S}^{D-1}} z_i^2\, d\mu(z)\right)^{1/2} = \left(\sum_{i=1}^d \frac{1}{D}\right)^{1/2} = \sqrt{\frac{d}{D}},$$

where we used the fact that coordinates z_i^2 are distributed identically and therefore they have the same ${}^1\!/\!_D$ mean. Applying inequality (23) yields, for any $t \ge 0$

$$\begin{aligned}\text{Prob}\left(U \in O(d \times D)\colon \|Ux\|_2 > \sqrt{{}^d\!/\!_D} + t\right) \quad &\le \quad \text{Prob}\left(z \in \mathbb{S}^{D-1}\colon |f(z) - \mathbb{E}f(z)| > t\right)\\ &\le \quad 5\exp\left(-\frac{Dt^2}{9}\right).\end{aligned} \tag{24}$$

Let S^δ be the set of vectors $x \in \mathbb{S}^{D-1}$ with at most δD non-zero elements

$$S^\delta := \left\{x \in \mathbb{S}^{D-1} \colon |\mathrm{supp}(x)| \le \delta D\right\} = \bigcup_{|I| \le \delta D} \left\{x \in \mathbb{S}^{D-1} \colon \mathrm{supp}(x) \subseteq I\right\} = \bigcup_{|I| \le \delta D} S_I^\delta,$$

where S_I^δ denotes the subset of vectors S^δ having a given support $I \subseteq [D]$ of indices. Fix $\varepsilon > 0$. For each I, we can find an ε-net for S_I^δ in the Euclidean norm with cardinality at most $({}^3\!/\varepsilon)^{\delta D}$ (see Proposition 4.2.12 and Corollary 4.2.13 in (Vershynin, 2018)). Taking the union over all sets I with $|I| = \lceil \delta D \rceil$, we conclude by the Stirling's approximation that there exists an ε-net $\mathcal{N}_\varepsilon$ of S^δ with cardinality

$$|\mathcal{N}_\varepsilon| \le \binom{D}{\lceil \delta D \rceil} \left(\frac{3}{\varepsilon}\right)^{\delta D} \le \left(\frac{3e}{\varepsilon\delta}\right)^{\delta D}. \tag{25}$$

Applying inequality (24), we have

$$\mathrm{Prob}\left(U \in O(d \times D) \colon \|Uy\|_2 > \sqrt{d/D} + t, \text{ for some } y \in \mathcal{N}_\varepsilon\right) \le |\mathcal{N}_\varepsilon| \cdot 5 \exp\left(-\frac{Dt^2}{9}\right). \tag{26}$$

Since $\mathcal{N}_\varepsilon$ is an ε-net for S^δ, then for any $x \in S^\delta$ there exists such $y \in \mathcal{N}_\varepsilon$ that $\|x - y\|_2 \le \varepsilon$. Furthermore, from the orthogonality of matrix U we conclude

$$\|Ux\|_2 \le \|Uy\|_2 + \|U(x-y)\|_2 \le \|Uy\|_2 + \varepsilon.$$

Hence, by relaxing the condition of probability in (26) and using the upper bound (25), we get

$$\begin{aligned}\mathrm{Prob}\left(U \in O(d \times D) \colon \|Ux\|_2 > \sqrt{d/D} + t + \varepsilon, \text{ for some } x \in S^\delta\right) &\le \left(\frac{3e}{\varepsilon\delta}\right)^{\delta D} \cdot 5 \exp\left(-\frac{Dt^2}{9}\right) \\ &= 5 \exp\left[-D\left(\frac{t^2}{9} - \delta \log \frac{3e}{\varepsilon\delta}\right)\right].\end{aligned}$$

The above inequality can be reformulated in terms of RIP condition for a random matrix $U \in O(d \times D)$

$$\mathrm{Prob}\left(U \in \mathrm{RIP}\left(\delta, \frac{1}{\sqrt{\lambda}} + t + \varepsilon\right)\right) \ge 1 - 5 \exp\left[-D\left(\frac{t^2}{9} - \delta \log \frac{3e}{\varepsilon\delta}\right)\right]. \tag{27}$$

Thus, recalling the formula (10) for the level K, we aim to choose such ε, t, δ (depending on λ) that to maximize both ${}^1\!/K$ and the probability in (27), i.e. the following two expressions

$$\sqrt{\delta}\left(1 - \frac{1}{\sqrt{\lambda}} - \varepsilon - t\right) \quad \text{and} \quad \frac{t^2}{9} - \delta \log \frac{3e}{\varepsilon\delta}. \tag{28}$$

Note that choosing parameters ε, t, δ is not trivial in this case as we want to maximize both terms and there is a trade-off between them. We choose the parameters as follows[4]

$$\varepsilon = \frac{1}{100}\left(1 - \frac{1}{\sqrt{\lambda}}\right), \quad t = 74\,\varepsilon, \quad \delta = 16\varepsilon^2. \tag{29}$$

With these choice of parameters we establish (12).

$$\begin{aligned}\eta &= \frac{1}{\sqrt{\lambda}} + t + \varepsilon = 1 - 25\,\varepsilon = 1 - \frac{1}{4}\left(1 - \frac{1}{\sqrt{\lambda}}\right) = \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{\sqrt{\lambda}}, \\ \delta &= 16\varepsilon^2 = \frac{1}{5^4}\left(1 - \frac{1}{\sqrt{\lambda}}\right)^2.\end{aligned} \tag{30}$$

[4] these expressions were constructed using two techniques: solving optimality conditions for the Lagrangian and numerical simulations.

To complete the theorem we need to bound the second expression of (28) for the probability. Letting $\nu = \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \in (0,1)$ and plugging the expressions (29) in (28) we get

$$\begin{aligned}
\frac{t^2}{9} - \delta \log \frac{3e}{\varepsilon\delta} &= \frac{74^2}{9}\varepsilon^2 - 16\varepsilon^2 \log \frac{3e/16}{\varepsilon^3} \\
&= \frac{74^2}{9 \cdot 10^4}\nu - \frac{16}{10^4}\frac{3}{2}\nu \log \frac{(3e/16)^{2/3} \cdot 10^4}{\nu} \\
&= A\nu - B\nu \log \frac{C}{\nu} \\
&= (A - B\log C)\nu + B\nu \log \nu \\
&= \nu\left(A - B\log C + B \log \nu\right) \\
&= \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \left((A - B\log C) + 2B \log\left(1 - \frac{1}{\sqrt{\lambda}}\right)\right) \\
&\geq \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \left(\frac{1}{26} + \frac{1}{208} \log\left(1 - \frac{1}{\sqrt{\lambda}}\right)\right),
\end{aligned}$$

where we defined absolute constants A, B, C as

$$A = \frac{74^2}{9 \cdot 10^4}, \quad B = \frac{24}{10^4}, \quad C = \left(\frac{3e}{16}\right)^{2/3} \cdot 10^4.$$

and used the following estimates

$$A - B\log C \geq \frac{1}{26}, \quad 2B = \frac{3}{5^4} \leq \frac{1}{208}.$$

This concludes the theorem as

$$\begin{aligned}
\mathrm{Prob}\left(U \in \mathrm{RIP}\left(\delta, \eta\right)\right) &\geq 1 - 5\exp\left[-D\left(\frac{t^2}{9} - \delta \log \frac{3e}{\varepsilon\delta}\right)\right] \\
&\geq 1 - 5\exp\left[-D\left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \left(\frac{1}{26} + \frac{1}{208}\log\left(1 - \frac{1}{\sqrt{\lambda}}\right)\right)\right] \\
&\geq 1 - 5\exp\left[-d\left(\sqrt{\lambda} - 1\right)^2 \left(\frac{1}{26} + \frac{1}{208}\log\left(1 - \frac{1}{\sqrt{\lambda}}\right)\right)\right].
\end{aligned}$$

B.3 Proof of Theorem 7: Kashin Compression

The unbiasedness of $\mathcal{C}_\kappa$ has been shown in part 4.3 with uniform upper bound $K(\lambda)^2$ for the variance. To prove the formula (13) we use expressions (30)

$$\omega_\lambda = K(\lambda)^2 = \left(\frac{1}{\sqrt{\delta}(1-\eta)}\right)^2 = \left(\frac{1}{4\varepsilon \cdot 25\varepsilon}\right)^2 = \left(\frac{1}{10\varepsilon}\right)^4 = \left(\frac{10\sqrt{\lambda}}{\sqrt{\lambda} - 1}\right)^4.$$