

On Biased Compression for Distributed Learning

Aleksandr Beznosikov*

ANBEZDOSIKOV@GMAIL.COM

*Computer, Electrical and Math. Sciences and Engineering Division
King Abdullah University of Science and Technology, 23955, Thuwal, KSA
Skolkovo Institute of Science and Technology, 121205, Moscow, Russia
School of Applied Mathematics and Informatics
Moscow Institute of Physics and Technology, 141701, Moscow, Russia*

Samuel Horváth†

SAMUEL.HORVATH@KAUST.EDU.SA

*Computer, Electrical and Math. Sciences and Engineering Division
King Abdullah University of Science and Technology, 23955, Thuwal, KSA*

Peter Richtárik

PETER.RICHTARIK@KAUST.EDU.SA

*Computer, Electrical and Math. Sciences and Engineering Division
King Abdullah University of Science and Technology, 23955, Thuwal, KSA*

Mher Safaryan‡

MHER.SAFARYAN.1@KAUST.EDU.SA

*Computer, Electrical and Math. Sciences and Engineering Division
King Abdullah University of Science and Technology, 23955, Thuwal, KSA*

Editor: Silvia Villa

Abstract

In the last few years, various communication compression techniques have emerged as an indispensable tool helping to alleviate the communication bottleneck in distributed learning. However, despite the fact *biased* compressors often show superior performance in practice when compared to the much more studied and understood *unbiased* compressors, very little is known about them. In this work we study three classes of biased compression operators, two of which are new, and their performance when applied to (stochastic) gradient descent and distributed (stochastic) gradient descent. We show for the first time that biased compressors can lead to linear convergence rates both in the single node and distributed settings. We prove that distributed compressed SGD method, employed with error feedback mechanism, enjoys the ergodic rate $\mathcal{O}\left(\delta L \exp\left[-\frac{\mu K}{\delta L}\right] + \frac{(C+\delta D)}{K\mu}\right)$, where $\delta \geq 1$ is a compression parameter which grows when more compression is applied, L and μ are the smoothness and strong convexity constants, C captures stochastic gradient noise ($C = 0$ if full gradients are

*The work in Sections 1-5 was conducted while A. Beznosikov was a research intern in the Optimization and Machine Learning Lab of Peter Richtárik at KAUST; this visit was funded by the KAUST Baseline Research Funding Scheme. The work of A. Beznosikov in Section 6 was conducted in Skoltech and was supported by Ministry of Science and Higher Education grant No. 075-10-2021-068. Alphabetical author order.

†Samuel Horváth is currently affiliated with Mohamed bin Zayed University of Artificial Intelligence (MBZUAI).

‡Mher Safaryan is currently affiliated with Institute of Science and Technology Austria (ISTA).

computed on each node) and D captures the variance of the gradients at the optimum ($D = 0$ for over-parameterized models). Further, via a theoretical study of several synthetic and empirical distributions of communicated gradients, we shed light on why and by how much biased compressors outperform their unbiased variants. Finally, we propose several new biased compressors with promising theoretical guarantees and practical performance.

Keywords: Compression operators, biased compressors, distributed learning, linear convergence, error feedback

1. Introduction

In order to achieve state-of-the-art performance, modern machine learning models need to be trained using large corpora of training data, and often feature an even larger number of trainable parameters (Vaswani et al., 2019; Brown et al., 2020). The data is typically collected in a distributed manner and stored across a network of edge devices, as is the case in federated learning (Konečný et al., 2016; McMahan et al., 2017; Li et al., 2019; Kairouz and et al, 2019), or collected centrally in a data warehouse composed of a large collection of commodity clusters. In either scenario, communication among the workers is typically the bottleneck.

Motivated by the need for more efficient training methods in traditional distributed and emerging federated environments, we consider optimization problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}, \quad (1)$$

where $x \in \mathbb{R}^d$ collects the parameters of a statistical model to be trained, n is the number of workers/devices, and $f_i(x)$ is the loss incurred by model x on data stored on worker i . The loss function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ often has the form

$$f_i(x) := \mathbb{E}_{\xi \sim \mathcal{P}_i} [f_\xi(x)],$$

with \mathcal{P}_i being the distribution of training data owned by worker i . In federated learning applications, these local distributions can be very different and we do not impose any similarity assumption for them.

1.1 Distributed optimization

A fundamental baseline for solving problem (1) is (distributed) gradient descent (GD), performing updates of the form

$$x^{k+1} = x^k - \frac{\eta^k}{n} \sum_{i=1}^n \nabla f_i(x^k),$$

where $\eta^k > 0$ is a stepsize. Due to the communication issues inherent to distributed systems, several enhancements to this baseline have been proposed that can better deal with the communication cost challenges of distributed environments, including acceleration (Nesterov, 2013; Beck and Teboulle, 2009; Allen-Zhu, 2017), reducing the number of iterations via momentum, local methods (McMahan et al., 2017; Khaled et al., 2020a; Karimireddy et al.,

2019a), reducing the number of communication rounds via performing multiple local updates before each communication round, and communication compression (Seide et al., 2014; Alistarh et al., 2017; Zhang et al., 2017; Lim et al., 2018; Alistarh et al., 2018; Lin et al., 2018; Safaryan and Richtárik, 2021), reducing the size of communicated messages via compression operators.

1.2 Contributions

In this paper we contribute to a better understanding of the latter approach to alleviating the communication bottleneck: *communication compression*. In particular, we study the theoretical properties of gradient-type methods which employ *biased* gradient compression operators, such as Top- k sparsification (Alistarh et al., 2018), or deterministic rounding (Sapio et al., 2019). Surprisingly, current * theoretical understanding of such methods is very limited. For instance, there is no general theory of such methods even in the $n = 1$ case, only a handful of biased compression techniques have been proposed in the literature, we do not have any theoretical understanding of why biased compression operators could outperform their unbiased counterparts and when. More importantly, there is no good convergence theory for any gradient-type method with a biased compression in the crucial $n > 1$ setting.

In this work we address all of the above problems. In particular, our main contributions are:

- (a) We define and study three parametric classes of biased compression operators (see Section 2), which we denote $\mathbb{B}^1(\alpha, \beta)$, $\mathbb{B}^2(\gamma, \beta)$ and $\mathbb{B}^3(\delta)$, the first two of which are new. We prove that, despite they are alternative parameterization of the same collection of operators, the last two are more favorable than the first one, thus highlighting the importance of parametrization and providing further reductions. We show how is the commonly used class of unbiased compression operators, which we denote $\mathbb{U}(\zeta)$, relates to these biased classes. We also study scaling and compositions of such compressors.
- (b) We then proceed to give a long list of new and known biased (and some unbiased) compression operators which belong to the above classes in Section 2.2. A summary of all compressors considered can be found in Table 3.
- (c) In Section 3 we analyze compressed GD in the $n = 1$ case for compressors belonging to all three classes under smoothness and strong convexity assumption. Our theorems generalize existing results which hold for unbiased operators in a tight manner, and also recover the rate of GD in this regime. Our linear convergence results are summarized in Table 1.
- (d) We ask the question: do biased compressors outperform their unbiased counterparts in theory, and by how much? We answer this question by studying the performance of compressors under various synthetic and empirical statistical assumptions on the distribution of the entries of gradient vectors which need to be compressed. Particularly, we quantify the gains of the Top- k sparsifier when compared against the unbiased Rand- k sparsifier in Section 4.

*Here we refer to the initial online appearance of our work on February of 2020, after which several enhancements were developed. See Section 1.3 for more details.

Compressor	$\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$	$\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$	$\mathcal{C} \in \mathbb{B}^3(\delta)$
Theorem	Theorem 17	Theorem 18	Theorem 19
Complexity	$\mathcal{O}\left(\frac{\beta^2 L}{\alpha \mu} \log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\beta L}{\gamma \mu} \log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\delta \mu} \log \frac{1}{\epsilon}\right)$

Table 1: Complexity results for GD with biased compression. The identity compressor $\mathcal{C}(x) \equiv x$ belongs to all classes with $\alpha = \beta = \gamma = \delta = 1$; all three results recover standard rate of GD.

Stepsize	Weights	Rate
$\mathcal{O}(\frac{1}{k})$	$\mathcal{O}(k)$	$\mathcal{O}\left(\frac{A_1}{K^2} + \frac{A_2}{K}\right)$
$\mathcal{O}(1)$	$\mathcal{O}(e^{-k})$	$\tilde{\mathcal{O}}\left(A_3 \exp\left[-\frac{K}{A_4}\right] + \frac{A_2}{K}\right)$
$\mathcal{O}(1)$	1	$\mathcal{O}\left(\frac{A_3}{K} + \frac{A_5}{\sqrt{K}}\right)$

Table 2: Ergodic convergence of distributed SGD with biased compression and error-feedback (Algorithm 1) for L -smooth and μ -strongly convex functions (K communications). Details are given in Theorem 21.

- (e) Finally, we study the important $n > 1$ setting in Section 5 and argue by giving a counterexample that a naive application of biased compression to distributed GD might diverge. We then show that distributed SGD method equipped with an error-feedback mechanism (Stich and Karimireddy, 2019) provably handles biased compressors. In our main result (Theorem 21; also see Table 2) we consider three learning schedules and iterate averaging schemes to provide three distinct convergence rates. Our analysis provides the first convergence guarantee for distributed gradient-type method which provably converges for biased compressors, and we thus solve a major open problem in the literature.

1.3 Related work

There has been extensive work related to communication compression, mostly focusing on unbiased compressions (Alistarh et al., 2017) as these are much easier to analyze. In particular, it was shown (Gorbunov et al., 2020a) that both the classical method with unbiased compression (Alistarh et al., 2017) and more advanced modifications (Mishchenko et al., 2019; Horváth et al., 2019b) can be considered as special versions of SGD. Subsequently, the results of (Gorbunov et al., 2020a) for strongly convex problems were transferred to general convex (Khaled et al., 2020b) and non-convex (Li and Richtárik, 2020) target functions. In the meantime, works concerning biased compressions show stronger empirical results but with limited or no analysis (Vogels et al., 2019; Lin et al., 2017a; Sun et al., 2019). There have been several attempts trying to address this issue, e.g., Wu et al. (2018) provided analysis for quadratics in distributed setting, Zhao et al. (2019) gave analysis for momentum

SGD with a specific biased compression, but under unreasonable assumptions, i.e., bounded gradient norm and memory. The first result that obtained linear rate of convergence for biased compression was done by Karimireddy et al. (2019b), but only for one node and under bounded gradient norm assumption, which was later overcome by Stich and Karimireddy (2019).

After the initial online appearance of our work, there has been several enhancements in the literature. In particular, Ajalloeian and Stich (2021) developed theory for non-convex objectives in the single node setup, Gorbunov et al. (2020b) designed a novel error compensated SGD algorithm converging linearly in a more relaxed setting with the help of additional unbiased compressor, Horváth and Richtárik (2021) proposed a simple trick to convert any biased compressor to corresponding induced (unbiased) compressor leading to improved theoretical guarantees. Recently, a new variant of error feedback mechanism was introduced in (Richtárik et al., 2021; Fatkhullin et al., 2021) showing an improved rates for distributed non-convex problems.

1.4 Basic notation and definitions

We use $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$ to denote standard inner product of $x, y \in \mathbb{R}^d$, where x_i corresponds to the i -th component of x in the standard basis in \mathbb{R}^d . This induces the ℓ_2 -norm in \mathbb{R}^d in the following way $\|x\|_2 := \sqrt{\langle x, x \rangle}$. We denote ℓ_p -norms as $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{1/p}$ for $p \in (1, \infty)$. By $\mathbb{E}[\cdot]$ we denote mathematical expectation. For a given differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we say that it is L -smooth if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

We say that it is μ -strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

2. Biased Compressors

By compression operator we mean a (possibly random) mapping $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ with some constraints. Typically, literature considers *unbiased* compression operators \mathcal{C} with a bounded second moment, i.e.

Definition 1 Let $\zeta \geq 1$. We say that $\mathcal{C} \in \mathbb{U}(\zeta)$ if \mathcal{C} is unbiased (i.e., $\mathbb{E}[\mathcal{C}(x)] = x$ for all x) and if the second moment is bounded as[†]

$$\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \zeta \|x\|_2^2, \quad \forall x \in \mathbb{R}^d. \quad (2)$$

2.1 Three classes of biased compressors

We instead focus on understanding *biased* compression operators, or “compressors” in short. We now introduce three classes of biased compressors, the first two are new, which can be seen as natural extensions of unbiased compressors.

[†](2) can be also written as $\mathbb{E} \left[\|\mathcal{C}(x) - x\|_2^2 \right] \leq (\zeta - 1) \|x\|_2^2$.

Definition 2 We say that $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$ for some $\alpha, \beta > 0$ if

$$\alpha \|x\|_2^2 \leq \mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \beta \langle \mathbb{E}[\mathcal{C}(x)], x \rangle, \quad \forall x \in \mathbb{R}^d. \quad (3)$$

As we shall show next, the second inequality in (3) implies $\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \beta^2 \|x\|_2^2$.

Lemma 3 For any $x \in \mathbb{R}^d$, if $\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \beta \langle \mathbb{E}[\mathcal{C}(x)], x \rangle$, then

$$\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \beta^2 \|x\|_2^2. \quad (4)$$

Proof Fix any $x \in \mathbb{R}^d$. Applying Jensen's inequality, the second inequality in (3) and Cauchy-Schwarz, we get

$$\|\mathbb{E}[\mathcal{C}(x)]\|_2^2 \leq \mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \stackrel{(3)}{\leq} \beta \langle \mathbb{E}[\mathcal{C}(x)], x \rangle \leq \beta \|\mathbb{E}[\mathcal{C}(x)]\|_2 \|x\|_2. \quad (5)$$

If $\mathbb{E}[\mathcal{C}(x)] \neq 0$, this implies $\|\mathbb{E}[\mathcal{C}(x)]\|_2 \leq \beta \|x\|_2$. Plugging this back into (5), we get (4). If $\mathbb{E}[\mathcal{C}(x)] = 0$, then from (3) we see that $\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] = 0$, and (4) holds trivially. \blacksquare

In the second class, we require the inner product between uncompressed x and compressed $\mathcal{C}(x)$ vectors to dominate the squared norms of both vectors in expectation.

Definition 4 We say that $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$ for some $\gamma, \beta > 0$ if

$$\max \left\{ \gamma \|x\|_2^2, \frac{1}{\beta} \mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \right\} \leq \langle \mathbb{E}[\mathcal{C}(x)], x \rangle \quad \forall x \in \mathbb{R}^d. \quad (6)$$

Finally, in the third class, we require the compression error $\|\mathcal{C}(x) - x\|_2^2$ to be strictly smaller than the squared norm $\|x\|_2^2$ of the input vector x in expectation.

Definition 5 We say $\mathcal{C} \in \mathbb{B}^3(\delta)$ for some $\delta > 0$ if

$$\mathbb{E} \left[\|\mathcal{C}(x) - x\|_2^2 \right] \leq \left(1 - \frac{1}{\delta} \right) \|x\|_2^2, \quad \forall x \in \mathbb{R}^d. \quad (7)$$

This last definition was also considered by Stich et al. (2018); Cordonnier (2018). All three definitions require the compressed vector $\mathcal{C}(x)$ to be in the neighborhood of the uncompressed vector x so that initial information is preserved with some accuracy. We now establish several basic properties and connections between the classes. We first show that the three classes of biased compressors defined above are equivalent in the following sense: a compressor from any of those three classes can be shown to belong to all three classes with different parameters and after possible scaling.

Theorem 6 (Equivalence between biased compressors) Let $\lambda > 0$ be a free scaling parameter.

1. If $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$, then

- (i) $\beta^2 \geq \alpha$ and $\lambda\mathcal{C} \in \mathbb{B}^1(\lambda^2\alpha, \lambda\beta)$, and
 - (ii) $\mathcal{C} \in \mathbb{B}^2(\alpha/\beta, \beta)$ and $\frac{1}{\beta}\mathcal{C} \in \mathbb{B}^3(\beta^2/\alpha)$.
2. If $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$, then
- (i) $\beta \geq \gamma$ and $\lambda\mathcal{C} \in \mathbb{B}^2(\lambda\gamma, \lambda\beta)$, and
 - (ii) $\mathcal{C} \in \mathbb{B}^1(\gamma^2, \beta)$ and $\frac{1}{\beta}\mathcal{C} \in \mathbb{B}^3(\beta/\gamma)$.
3. If $\mathcal{C} \in \mathbb{B}^3(\delta)$, then
- (i) $\delta \geq 1$, and
 - (ii) $\mathcal{C} \in \mathbb{B}^2(\frac{1}{2\delta}, 2) \subseteq \mathbb{B}^1(\frac{1}{4\delta^2}, 2)$.

Proof Let us prove this implications for each class separately.

1. Case $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$:

- (i) Let us choose any $x \neq 0$ and observe that (3) implies that $\mathbf{E}[\mathcal{C}(x)] \neq 0$. Further, from (3) we get the bounds

$$\frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\langle \mathbf{E}[\mathcal{C}(x)], x \rangle} \leq \beta, \quad \alpha \leq \frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\|x\|_2^2}.$$

Finally,

$$\beta^2 \geq \left(\frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\langle \mathbf{E}[\mathcal{C}(x)], x \rangle} \right)^2 \geq \frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2] \mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\|\mathbf{E}[\mathcal{C}(x)]\|_2^2 \|x\|_2^2} \geq \alpha \frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\|\mathbf{E}[\mathcal{C}(x)]\|_2^2} \geq \alpha,$$

where the second inequality is due to Cauchy-Schwarz, and the last inequality follows by applying Jensen inequality.

The scaling property $\lambda\mathcal{C} \in \mathbb{B}^1(\alpha\lambda^2, \beta\lambda)$ follows directly from (3).

- (ii) The inclusion $\mathcal{C} \in \mathbb{B}^2(\alpha/\beta, \beta)$ follows directly from inequalities (3) and (6). In view of (i), $\lambda\mathcal{C} \in \mathbb{B}^1(\lambda^2\alpha, \lambda\beta)$. If we choose $\lambda \leq \frac{2}{\beta}$, then

$$\begin{aligned} \mathbf{E}[\|\lambda\mathcal{C}(x) - x\|_2^2] &= \mathbf{E}[\|\lambda\mathcal{C}(x)\|_2^2] - 2\langle \mathbf{E}[\lambda\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &\stackrel{(3)}{\leq} (\beta\lambda - 2)\langle \mathbf{E}[\lambda\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &\stackrel{(3)}{\leq} (\beta\lambda - 2)\frac{\alpha\lambda^2}{\beta\lambda} \|x\|_2^2 + \|x\|_2^2 \\ &\stackrel{(3)}{\leq} \left(\alpha\lambda^2 - 2\frac{\alpha}{\beta}\lambda + 1 \right) \|x\|_2^2. \end{aligned}$$

Minimizing the above expression in λ , we get $\lambda = \frac{1}{\beta}$, and the result follows.

2. Case $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$.

(i) Using (6) we get

$$\gamma \leq \frac{\langle \mathbf{E}[\mathcal{C}(x)], x \rangle}{\|x\|_2^2} \leq \frac{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]}{\sqrt{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]} \|x\|_2} \leq \beta \frac{\langle \mathbf{E}[\mathcal{C}(x)], x \rangle}{\sqrt{\mathbf{E}[\|\mathcal{C}(x)\|_2^2]} \|x\|_2} \leq \beta,$$

where the first and third inequalities follow from (6) and the third and the last from Cauchy-Schwarz inequality with Jensen inequality.

The scaling property $\lambda\mathcal{C} \in \mathbb{B}^2(\lambda\gamma, \lambda\beta)$ follows directly from (6).

(ii) If $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$, then $\mathbf{E}[\|\mathcal{C}(x)\|_2^2] \leq \beta \langle \mathbf{E}[\mathcal{C}(x)], x \rangle$ and

$$\gamma^2 \|x\|_2^4 \stackrel{(6)}{\leq} \langle \mathbf{E}[\mathcal{C}(x)], x \rangle^2 \leq \|\mathbf{E}[\mathcal{C}(x)]\|_2^2 \|x\|_2^2 \leq \mathbf{E}[\|\mathcal{C}(x)\|_2^2] \|x\|_2^2,$$

where the second inequality is Cauchy-Schwarz, and the third is Jensen. Therefore, $\mathcal{C} \in \mathbb{B}^1(\gamma^2, \beta)$.

Further, for any $\lambda > 0$, we get

$$\begin{aligned} \mathbf{E}[\|\lambda\mathcal{C}(x) - x\|_2^2] &= \mathbf{E}[\|\lambda\mathcal{C}(x)\|_2^2] - 2\langle \mathbf{E}[\lambda\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &= \lambda^2 \mathbf{E}[\|\mathcal{C}(x)\|_2^2] - 2\lambda \langle \mathbf{E}[\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &\stackrel{(6)}{\leq} (\lambda\beta - 2)\lambda \langle \mathbf{E}[\mathcal{C}(x)], x \rangle + \|x\|_2^2. \end{aligned}$$

If we choose $\lambda = \frac{1}{\beta}$, then we can continue as follows:

$$\begin{aligned} \mathbf{E}[\|\lambda\mathcal{C}(x) - x\|_2^2] &\leq -\frac{1}{\beta} \langle \mathbf{E}[\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &\stackrel{(6)}{\leq} \left(1 - \frac{\gamma}{\beta}\right) \|x\|_2^2, \end{aligned}$$

whence $\frac{1}{\beta}\mathcal{C} \in \mathbb{B}^3(\beta/\gamma)$.

3. Case $\mathcal{C} \in \mathbb{B}^3(\delta)$.

(i) Pick $x \neq 0$. Since $0 \leq \mathbf{E}[\|\mathcal{C}(x) - x\|_2^2] \leq (1 - \frac{1}{\delta}) \|x\|_2^2$ and we assume $\delta > 0$, we must necessarily have $\delta \geq 1$.

(ii) If $\mathcal{C} \in \mathbb{B}^3(\delta)$ then

$$\mathbf{E}[\|\mathcal{C}(x)\|_2^2] - 2\langle \mathbf{E}[\mathcal{C}(x)], x \rangle + \frac{1}{\delta} \|x\|_2^2 \leq 0,$$

which implies that

$$\frac{1}{2\delta} \|x\|_2^2 \leq \langle \mathbf{E}[\mathcal{C}(x)], x \rangle \quad \text{and} \quad \mathbf{E}[\|\mathcal{C}(x)\|_2^2] \leq 2\langle \mathbf{E}[\mathcal{C}(x)], x \rangle.$$

Therefore, $\mathcal{C} \in \mathbb{B}^2(\frac{1}{2\delta}, 2) \subseteq \mathbb{B}^1(\frac{1}{4\delta^2}, 2)$.

■

Next, we show that, with a proper scaling, any unbiased compressor also belongs to all the three classes of biased compressors.

Theorem 7 (From unbiased to biased with scaling) *If $\mathcal{C} \in \mathbb{U}(\zeta)$, then for the scaled operator $\lambda\mathcal{C}$ we have*

- (i) $\lambda\mathcal{C} \in \mathbb{B}^1(\lambda^2, \lambda\zeta)$ for $\lambda > 0$,
- (ii) $\lambda\mathcal{C} \in \mathbb{B}^2(\lambda, \lambda\zeta)$ for $\lambda > 0$,
- (iii) $\lambda\mathcal{C} \in \mathbb{B}^3\left(\frac{1}{\lambda(2-\zeta\lambda)}\right)$ for $\zeta\lambda < 2$.

Proof Let $\mathcal{C} \in \mathbb{U}(\zeta)$.

- (i) Given any $\lambda > 0$, consider the scaled operator $\lambda\mathcal{C}$. We have

$$\lambda^2 \|x\|_2^2 = \|\mathbb{E}[\lambda\mathcal{C}(x)]\|_2^2 \leq \mathbb{E}\left[\|\lambda\mathcal{C}(x)\|_2^2\right] \leq \lambda^2\zeta \|x\|_2^2 = \lambda\zeta \langle \mathbb{E}[\lambda\mathcal{C}(x)], x \rangle,$$

whence $\mathcal{C} \in \mathbb{B}^1(\lambda^2, \lambda\zeta)$.

- (ii) Given any $\lambda > 0$, consider the scaled operator $\lambda\mathcal{C}$. We have

$$\begin{aligned} \lambda \|x\|_2^2 &= \langle \mathbb{E}[\lambda\mathcal{C}(x)], x \rangle, \\ \mathbb{E}\left[\|\lambda\mathcal{C}(x)\|_2^2\right] &\leq \lambda^2\zeta \|x\|_2^2 = \lambda\zeta \langle \mathbb{E}[\lambda\mathcal{C}(x)], x \rangle, \end{aligned}$$

whence $\lambda\mathcal{C} \in \mathbb{B}^2(\lambda, \lambda\zeta)$.

- (iii) Given $\lambda > 0$ such that $\zeta\lambda < 2$, consider the scaled operator $\lambda\mathcal{C}$. We have

$$\begin{aligned} \mathbb{E}\left[\|\lambda\mathcal{C}(x) - x\|_2^2\right] &= \mathbb{E}\left[\|\lambda\mathcal{C}(x)\|_2^2\right] - 2\langle \mathbb{E}[\lambda\mathcal{C}(x)], x \rangle + \|x\|_2^2 \\ &\leq (\zeta\lambda^2 - 2\lambda + 1) \|x\|_2^2 \end{aligned}$$

whence $\lambda\mathcal{C} \in \mathbb{B}^3\left(\frac{1}{\lambda(2-\zeta\lambda)}\right)$.

■

2.2 Examples of biased compressors: old and new

We now give some examples of compression operators belonging to the classes \mathbb{B}^1 , \mathbb{B}^2 , \mathbb{B}^3 and \mathbb{U} . Several of them are new. For a summary, refer to Table 3.

Compression Operator \mathcal{C}	Unbiased?	α	β	γ	δ	ζ
Unbiased random sparsification	✓					d/k
Biased random sparsification [NEW]	✗	q	1	q	$1/q$	
Adaptive random sparsification [NEW]	✗	$1/d$	1	$1/d$	d	
Top- k sparsification (Alistarh et al., 2018)	✗	k/d	1	k/d	d/k	
General unbiased rounding [NEW]	✓					$\frac{1}{4} \sup \left(\frac{a_k}{a_{k+1}} + \frac{a_{k+1}}{a_k} + 2 \right)$
Unbiased exponential rounding [NEW]	✓					$\frac{1}{4} \left(b + \frac{1}{b} + 2 \right)$
Biased exponential rounding [NEW]	✗	$\left(\frac{2}{b+1} \right)^2$	$\frac{2b}{b+1}$	$\frac{2}{b+1}$	$\frac{(b+1)^2}{4b}$	
Natural compression (Horváth et al., 2019a)	✓					$9/s$
General exponential dithering [NEW]	✓					ζ_b
Natural dithering (Horváth et al., 2019a)	✓					ζ_2
Top- k + exponential dithering [NEW]	✗	k/d	ζ_b	k/d	$\zeta_b d/k$	

Table 3: Compressors \mathcal{C} described in Section 2.2 and their membership in $\mathbb{B}^1(\alpha, \beta)$, $\mathbb{B}^2(\gamma, \beta)$, $\mathbb{B}^3(\delta)$ and $\mathbb{U}(\zeta)$.

- (a) For $k \in [d] := \{1, \dots, d\}$, the **unbiased random (aka Rand- k) sparsification** operator is defined via

$$\mathcal{C}(x) := \frac{d}{k} \sum_{i \in S} x_i e_i, \quad (8)$$

where $S \subseteq [d]$ is the k -nice sampling; i.e., a subset of $[d]$ of cardinality k chosen uniformly at random, and e_1, \dots, e_d are the standard unit basis vectors in \mathbb{R}^d .

Lemma 8 *The Rand- k sparsifier (8) belongs to $\mathbb{U}(\frac{d}{k})$.*

- (b) Let $S \subseteq [d]$ be a random set, with probability vector $p := (p_1, \dots, p_d)$, where $p_i := \text{Prob}(i \in S) > 0$ for all i (such a set is called a proper sampling (Richtárik and Takáč, 2016)). Define **biased random sparsification** operator via

$$\mathcal{C}(x) := \sum_{i \in S} x_i e_i. \quad (9)$$

Lemma 9 *Letting $q := \min_i p_i$, the biased random sparsification operator (9) belongs to $\mathbb{B}^1(q, 1)$, $\mathbb{B}^2(q, 1)$, $\mathbb{B}^3(1/q)$.*

- (c) **Adaptive random sparsification** is defined via

$$\mathcal{C}(x) := x_i e_i \quad \text{with probability} \quad \frac{|x_i|}{\|x\|_1}. \quad (10)$$

Lemma 10 *Adaptive random sparsification operator (10) belongs to $\mathbb{B}^1(\frac{1}{d}, 1)$, $\mathbb{B}^2(\frac{1}{d}, 1)$, $\mathbb{B}^3(d)$.*

- (d) **Greedy (aka Top- k) sparsification** operator is defined via

$$\mathcal{C}(x) := \sum_{i=d-k+1}^d x_{(i)} e_{(i)}, \quad (11)$$

where coordinates are ordered by their magnitudes so that $|x_{(1)}| \leq |x_{(2)}| \leq \dots \leq |x_{(d)}|$.

Lemma 11 *Top- k sparsification operator (11) belongs to $\mathbb{B}^1(\frac{k}{d}, 1)$, $\mathbb{B}^2(\frac{k}{d}, 1)$, and $\mathbb{B}^3(\frac{d}{k})$.*

- (e) Let $(a_k)_{k \in \mathbb{Z}}$ be an arbitrary increasing sequence of positive numbers such that $\inf a_k = 0$ and $\sup a_k = \infty$. Then **general unbiased rounding** \mathcal{C} is defined as follows: if $a_k \leq |x_i| \leq a_{k+1}$ for some coordinate $i \in [d]$, then

$$\mathcal{C}(x)_i = \begin{cases} \text{sign}(x_i)a_k & \text{with probability } \frac{a_{k+1}-|x_i|}{a_{k+1}-a_k} \\ \text{sign}(x_i)a_{k+1} & \text{with probability } \frac{|x_i|-a_k}{a_{k+1}-a_k} \end{cases} \quad (12)$$

Lemma 12 *General unbiased rounding operator (12) belongs to $\mathbb{U}(\zeta)$, where*

$$\zeta = \frac{1}{4} \sup_{k \in \mathbb{Z}} \left(\frac{a_k}{a_{k+1}} + \frac{a_{k+1}}{a_k} + 2 \right).$$

Notice that ζ is minimizing for exponential roundings $a_k = b^k$ with some basis $b > 1$, in which case $\zeta = \frac{1}{4}(b + 1/b + 2)$.

- (f) Let $(a_k)_{k \in \mathbb{Z}}$ be an arbitrary increasing sequence of positive numbers such that $\inf a_k = 0$ and $\sup a_k = \infty$. Then **general biased rounding** is defined via

$$\mathcal{C}(x)_i := \text{sign}(x_i) \arg \min_{t \in (a_k)} |t - |x_i||, \quad i \in [d]. \quad (13)$$

Lemma 13 *General biased rounding operator (13) belongs to $\mathbb{B}^1(\alpha, \beta)$, $\mathbb{B}^2(\gamma, \beta)$, and $\mathbb{B}^3(\delta)$, where*

$$\beta = \sup_{k \in \mathbb{Z}} \frac{2a_{k+1}}{a_k + a_{k+1}}, \quad \gamma = \inf_{k \in \mathbb{Z}} \frac{2a_k}{a_k + a_{k+1}}, \quad \alpha = \gamma^2, \quad \delta = \sup_{k \in \mathbb{Z}} \frac{(a_k + a_{k+1})^2}{4a_k a_{k+1}}.$$

In the special case of exponential rounding $a_k = b^k$ with some base $b > 1$, we get

$$\alpha = \left(\frac{2}{b+1} \right)^2, \quad \beta = \frac{2b}{b+1}, \quad \gamma = \frac{2}{b+1}, \quad \delta = \frac{(b+1)^2}{4b}.$$

Remark 14 *Plugging these parameters into the iteration complexities of Table 1, we find that the class \mathbb{B}^3 gives the best iteration complexity as $\frac{\beta^2}{\alpha} = b^2 > \frac{\beta}{\gamma} = b > \delta = \frac{(b+1)^2}{4b}$.*

- (g) **Natural compression** operator \mathcal{C}_{nat} of Horváth et al. (2019a) is the special case of general unbiased rounding operator (12) when $b = 2$. So,

$$\mathcal{C}_{nat} \in \mathbb{U} \left(\frac{9}{8} \right).$$

- (h) For $b > 1$, define **general exponential dithering** operator with respect to l_p -norm and with s exponential levels $0 < b^{1-s} < b^{2-s} < \dots < b^{-1} < 1$ via

$$\mathcal{C}(x) := \|x\|_p \times \text{sign}(x) \times \xi \left(\frac{|x_i|}{\|x\|_p} \right), \quad (14)$$

where the random variable $\xi(t)$ for $t \in [b^{-u-1}, b^{-u}]$ is set to either b^{-u-1} or b^{-u} with probabilities proportional to $b^{-u} - t$ and $t - b^{-u-1}$, respectively.

Lemma 15 *General exponential dithering operator (14) belongs to $\mathbb{U}(\zeta_b)$, where, letting $r = \min(p, 2)$,*

$$\zeta_b = \frac{1}{4} \left(b + \frac{1}{b} + 2 \right) + d^{\frac{1}{r}} b^{1-s} \min(1, d^{\frac{1}{r}} b^{1-s}). \quad (15)$$

- (i) **Natural dithering** introduced by Horváth et al. (2019a) without norm compression is the spacial case of general exponential dithering (14) when $b = 2$.
- (j) **Ternary quantization** of Wen et al. (2017) is the extreme case of general exponential dithering (14) with $s = 1$ levels and $b = 1$.
- (k) **Top- k combined with exponential dithering.** Let \mathcal{C}_{top} be the Top- k sparsification operator (11) and \mathcal{C}_{dith} be general exponential dithering operator (14) with some base $b > 1$ and parameter ζ_b from (15). Define a new compression operator as the composition of these two:

$$\mathcal{C}(x) := \mathcal{C}_{dith}(\mathcal{C}_{top}(x)). \quad (16)$$

Lemma 16 *The composition operator (16) of Top- k sparsification and exponential dithering with base b belongs to $\mathbb{B}^1(\frac{k}{d}, \zeta_b)$, $\mathbb{B}^2(\frac{k}{d}, \zeta_b)$, $\mathbb{B}^3(\frac{d}{k}\zeta_b)$, where ζ_b is as in (15).*

3. Gradient Descent with Biased Compression

As we discussed in previous section, compression operators can have different equivalent parametrizations. Next, we aim to investigate the influence of those parametrizations on the theoretical convergence rate of an algorithm employing compression operators. To achieve clearer understanding of the interaction of compressor parametrization and convergence rate, we first consider the single node, unconstrained optimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth and μ -strongly convex. We study the method

$$\boxed{x^{k+1} = x^k - \eta \mathcal{C}^k(\nabla f(x^k))}, \quad (\text{CGD})$$

where $\mathcal{C}^k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ are (potentially biased) compression operators belonging to one of the classes \mathbb{B}^1 , \mathbb{B}^2 and \mathbb{B}^3 studied in the previous sections, and $\eta > 0$ is a stepsize. We refer to this method as CGD: Compressed Gradient Descent.

3.1 Complexity theory

We now establish three theorems, performing complexity analysis for each of the three classes \mathbb{B}^1 , \mathbb{B}^2 and \mathbb{B}^3 individually. Let $\mathcal{E}_k := \mathbb{E} [f(x^k)] - f(x^*)$, with $\mathcal{E}_0 = f(x^0) - f(x^*)$.

Theorem 17 *Let $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$. Then as long as $0 \leq \eta \leq \frac{2}{\beta L}$, we have $\mathcal{E}_k \leq \left(1 - \frac{\alpha}{\beta} \eta \mu (2 - \eta \beta L)\right) \mathcal{E}_{k-1}$. If we choose $\eta = \frac{1}{\beta L}$, then*

$$\mathcal{E}_k \leq \left(1 - \frac{\alpha \mu}{\beta^2 L}\right)^k \mathcal{E}_0.$$

Theorem 18 *Let $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$. Then as long as $0 \leq \eta \leq \frac{2}{\beta L}$, we have $\mathcal{E}_k \leq (1 - \gamma \eta (2 - \eta \beta L)) \mathcal{E}_{k-1}$. If we choose $\eta = \frac{1}{\beta L}$, then*

$$\mathcal{E}_k \leq \left(1 - \frac{\gamma \mu}{\beta L}\right)^k \mathcal{E}_0.$$

Theorem 19 *Let $\mathcal{C} \in \mathbb{B}^3(\delta)$. Then as long as $0 \leq \eta \leq \frac{1}{L}$, we have $\mathcal{E}_k \leq (1 - \frac{1}{8} \eta \mu)^k \mathcal{E}_{k-1}$. If we choose $\eta = \frac{1}{L}$, then*

$$\mathcal{E}_k \leq \left(1 - \frac{1}{8} \frac{\mu}{L}\right)^k \mathcal{E}_0.$$

The iteration complexity for these results can be found in Table 1. Note that the identity compressor $\mathcal{C}(x) \equiv x$ belongs to $\mathbb{B}^1(1, 1)$, $\mathbb{B}^2(1, 1)$, and $\mathbb{B}^3(1)$, hence all these result exactly recover the rate of GD. In the first two theorems, scaling the compressor by a positive scalar $\lambda > 0$ does not influence the rate (see Theorem 6).

One can note that it is possible to adapt the results of Theorems 17-19 to the case with the time-varying parameters of compression operators (Lin et al., 2017b; Agarwal et al., 2020). In the case of Theorem 19 and \mathbb{B}^3 compressors, it is sufficient to introduce a sequence $\{\delta^i\}$ responsible for the changes of the parameter δ and obtain the following convergence: $\mathcal{E}_k \leq \left[\prod_{i=0}^{k-1} \left(1 - \frac{1}{\delta^i} \frac{\mu}{L}\right)\right] \mathcal{E}_0$. In the case of Theorems 17 and 18, the results are similar, but it is necessary to add in the algorithm (CGD) the possibility of using time-varying steps $\eta^k = \frac{1}{\beta^k L}$.

3.2 \mathbb{B}^3 and \mathbb{B}^2 are better than \mathbb{B}^1

In light of the results above, we make the following observation. If $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$, then by Theorem 6, $\frac{1}{\beta} \mathcal{C} \in \mathbb{B}^3(\frac{\beta^2}{\alpha})$. Applying Theorem 19, we get the bound $\mathcal{O}\left(\frac{\beta^2}{\alpha} \frac{L}{\mu} \log \frac{1}{\epsilon}\right)$. This is the same result as that obtained by Theorem 17. On the other hand, if $\mathcal{C} \in \mathbb{B}^3(\delta)$, then by Theorem 6, $\mathcal{C} \in \mathbb{B}^1(\frac{1}{4\delta^2}, 2)$. Applying Theorem 17, we get the bound $\mathcal{O}\left(16\delta^2 \frac{L}{\mu} \log \frac{1}{\epsilon}\right)$. This is a worse result than what Theorem 19 offers by a factor of 16δ .

Similarly, if $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$, then by Theorem 6, $\mathcal{C} \in \mathbb{B}^2(\frac{\alpha}{\beta}, \beta)$. Applying Theorem 18, we get the bound $\mathcal{O}\left(\frac{\beta^2}{\alpha} \frac{L}{\mu} \log \frac{1}{\epsilon}\right)$. This is the same result as that obtained by Theorem 17. On the other hand, if $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$, then by Theorem 6, $\mathcal{C} \in \mathbb{B}^1(\gamma^2, \beta)$. Applying Theorem 17,

we get the bound $\mathcal{O}\left(\frac{\beta^2}{\gamma^2} \frac{L}{\mu} \log \frac{1}{\varepsilon}\right)$. This is a worse result than what Theorem 18 offers by a factor of $\frac{\beta}{\gamma} \geq 1$.

Hence, while $\mathbb{B}^1, \mathbb{B}^2$ and \mathbb{B}^3 describe the same classes of compressors, for the purposes of CGD it is better to parameterize them as members of \mathbb{B}^2 or \mathbb{B}^3 .

4. Superiority of Biased Compressors Under Statistical Assumptions

Here we highlight some advantages of biased compressors by comparing them with their unbiased cousins. We evaluate compressors by their average capacity of preserving the gradient information or, in other words, by expected approximation error they produce. In the sequel, we assume that gradients have i.i.d. coordinates drawn from some distribution.

4.1 Top- k vs Rand- k

We now compare two sparsification operators: Rand- k (8) which is unbiased and which we denote as \mathcal{C}_{rnd}^k , and Top- k (11) which is biased and which we denote as \mathcal{C}_{top}^k . We define variance of the approximation error of x via

$$\omega_{rnd}^k(x) := \mathbb{E} \left[\left\| \frac{k}{d} \mathcal{C}_{rnd}^k(x) - x \right\|_2^2 \right] = \left(1 - \frac{k}{d}\right) \|x\|_2^2$$

and

$$\omega_{top}^k(x) := \left\| \mathcal{C}_{top}^k(x) - x \right\|_2^2 = \sum_{i=1}^{d-k} x_{(i)}^2$$

and the energy “saving” via

$$s_{rnd}^k(x) := \|x\|_2^2 - \omega_{rnd}^k(x) = \mathbb{E} \left[\left\| \frac{k}{d} \mathcal{C}_{rnd}^k(x) \right\|_2^2 \right] = \frac{k}{d} \|x\|_2^2$$

and

$$s_{top}^k(x) := \|x\|_2^2 - \omega_{top}^k(x) = \left\| \mathcal{C}_{top}^k(x) \right\|_2^2 = \sum_{i=d-k+1}^d x_{(i)}^2.$$

Expectations in these expressions are taken with respect to the randomization of the compression operator rather than input vector x . Clearly, there exists x for which these two operators incur identical variance, e.g. $x_1 = \dots = x_d$. However, in practice we apply compression to gradients x which evolve in time, and which may have heterogeneous components. In such situations, $\omega_{top}^k(x)$ could be much smaller than $\omega_{rnd}^k(x)$. This motivates a *quantitative study* of the *average case* behavior in which we make an *assumption* on the distribution of the coordinates of the compressed vector.

Uniform and exponential distribution. We first consider the case of uniform and exponentially distributed entries, and quantify the difference.

Lemma 20 Assume the coordinates of $x \in \mathbb{R}^d$ are i.i.d.

(a) If they follow uniform distribution over $[0, 1]$, then

$$\frac{\mathbb{E}[\omega_{top}^k]}{\mathbb{E}[\omega_{rnd}^k]} = \left(1 - \frac{k}{d+1}\right) \left(1 - \frac{k}{d+2}\right), \quad \frac{\mathbb{E}[s_{top}^1]}{\mathbb{E}[s_{rnd}^1]} = \frac{3d}{d+2}.$$

(b) If they follow standard exponential distribution, then

$$\frac{\mathbb{E}[s_{top}^1]}{\mathbb{E}[s_{rnd}^1]} = \frac{1}{2} \sum_{i=1}^d \frac{1}{i^2} + \frac{1}{2} \left(\sum_{i=1}^d \frac{1}{i} \right)^2 = \mathcal{O}(\log^2 d).$$

	Top-3				Top-5			
d	10^2	10^3	10^4	10^5	10^2	10^3	10^4	10^5
$\mathcal{N}(0; 1)$	$3 \cdot (\sigma^2 + \mu^2) = 3$				$5 \cdot (\sigma^2 + \mu^2) = 5$			
$\mathbb{E}[s_{top}^k(x)]$	18.65	31.10	43.98	57.08	27.14	47.70	69.07	90.85
$\mathcal{N}(2; 1)$	$3 \cdot (\sigma^2 + \mu^2) = 15$				$5 \cdot (\sigma^2 + \mu^2) = 25$			
$\mathbb{E}[s_{top}^k(x)]$	53.45	75.27	95.81	115.53	81.60	118.56	153.13	186.22

Table 4: Information savings of greedy and random sparsifiers for $k = 3$ and $k = 5$.

Empirical comparison. Now we compare these two sparsification methods on an empirical bases and show the significant advantage of greedy sparsifier against random sparsifier. We assume that coordinates of to-be-compressed vector are i.i.d. Gaussian random variables.

First, we compare the savings s_{top}^k and s_{rnd}^k of these compressions. For random sparsification, we have

$$\mathbb{E}[s_{rnd}^k(x)] = k \cdot (\sigma^2 + \mu^2),$$

where μ and σ^2 are the mean and variance of the Gaussian distribution. For computing $\mathbb{E}[s_{top}^k(x)]$, we use the probability density function of k -th order statistics (see (31) or (2.2.2) of (Arnold et al., 1992)). Table 4 shows that Top-3 and Top-5 sparsifiers “save” 3×–40× more information in expectation and the factor grows with the dimension.

Next we compare normalized variances $\frac{\omega_{top}^k(x)}{\|x\|_2^2}$ and $\frac{\omega_{rnd}^k(x)}{\|x\|_2^2}$ for randomly generated Gaussian vectors. In an attempt to give a dimension independent comparison, we compare them against the average number of encoding bits per coordinate, which is quite stable with respect to the dimension. Figure 1 reveals the superiority of greedy sparsifier against the random one.

Practical distribution. We obtained various gradient distributions via logistic regression (*mushrooms* LIBSVM dataset) and least squares. We used the sklearn package and built Gaussian smoothing of the practical gradient density. The second moments, i.e. energy

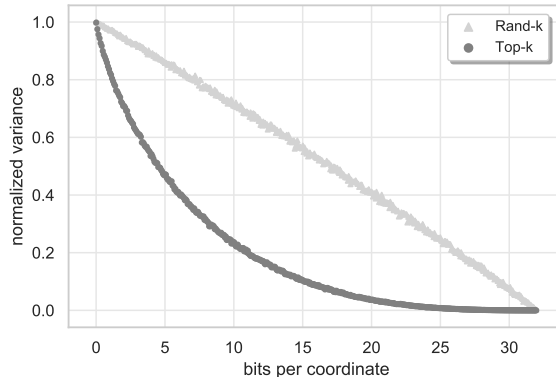


Figure 1: The comparison of Top- k and Rand- k sparsifiers with respect to normalized variance and the number of encoding bits used for each coordinate on average. Each point/marker represents a single $d = 10^4$ dimensional vector drawn from Gaussian distribution and then compressed by the specified operator. Each curve was obtained by varying the free parameter $k \in \{1, 2, \dots, d\}$. Plots for different d look very similar. Notice that, for random sparsification the normalized variance is perfectly linear with respect to the number of bit per coordinate. Letting b be the total number of bits to encode the compressed vector (say in *binary32* system), the normalized variance produced by random sparsifier is almost $1 - \frac{b/d}{32}$. However, greedy sparsifier achieves exponentially lower variance $\approx 0.86^{b/d}$ utilizing the same amount of bits.

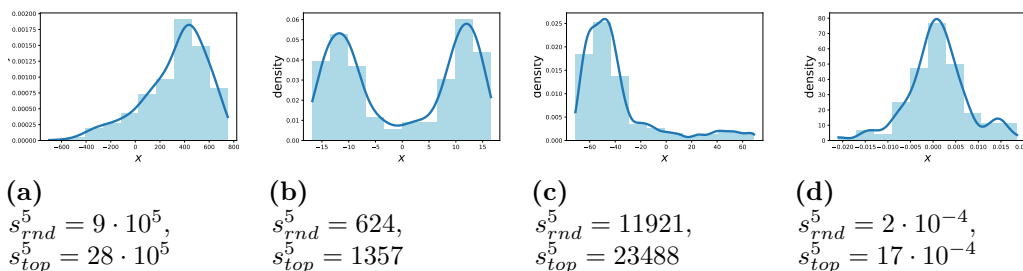


Figure 2: Calculations of the Rand-5 and Top-5 energy “saving” for practical gradient distributions ((a),(b),(c): quadratic problem, (d): logistic regression). The results of Top-5 are 3–5 \times better.

“saving”, were already calculated from it by formula for density function of k -order statistics, see Appendix A.4 or (Arnold et al., 1992). We conclude experiments for Top-5 and Rand-5, see Figure 2 for details.

4.2 New compressor: Top- k combined with dithering

In Section 2.2 we gave a new biased compression operator (see (16)), where we combined Top- k sparsification operator (see (11)) with the general exponential dithering (see (14)).

Consider the composition operator with natural dithering, i.e., with base $b = 2$. We showed that it belongs to $\mathbb{B}^1(\frac{k}{d}, \frac{9}{8})$, $\mathbb{B}^2(\frac{k}{d}, \frac{9}{8})$ and $\mathbb{B}^3(\frac{9d}{8k})$. Figure 3 empirically confirms that it attains the lowest compression parameter $\delta \geq 1$ among all other known compressors (see (5)). Furthermore, the iteration complexity $\mathcal{O}\left(\delta \frac{L}{\mu} \log \frac{1}{\varepsilon}\right)$ of CGD for $\mathcal{C} \in \mathbb{B}^3(\delta)$ implies that it enjoys fastest convergence.

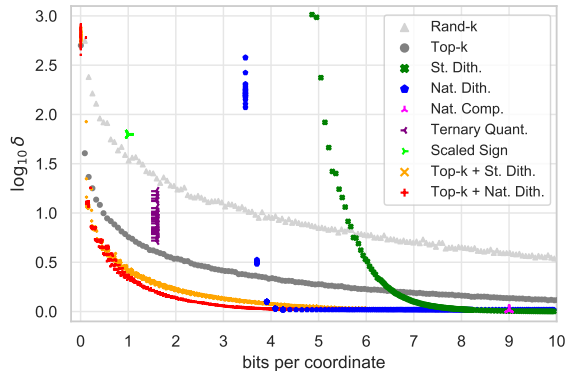


Figure 3: Comparison of various compressors (with and without free design parameters) with respect to the parameter $\delta \geq 1$ in \log_{10} –scale and the number of encoding bits used for each coordinate on average. Each point/marker represents a single $d = 10^4$ dimensional vector x drawn from Gaussian distribution and then compressed by the specified operator. Each curve was obtained by varying free parameters ($k \in \{1, 2, \dots, d\}$ for sparsifiers, $s \geq 1$ for ditherings) of the specified compressor. Parameter-free compressors, such as ternary quantization and natural compression, have fixed communication budgets which explains the vertical arrangements of the points.

5. Distributed Setting

We now focus attention on a distributed setup with n machines, each of which owns *non-iid* data defining one loss function f_i . Our goal is to minimize the average loss:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}. \quad (17)$$

5.1 Distributed CGD with unbiased compressors

Perhaps the most straightforward extension of CGD to the distributed setting is to consider the method

$$x^{k+1} = x^k - \eta \frac{1}{n} \sum_{i=1}^n \mathcal{C}_i^k(\nabla f_i(x^k)). \quad (\text{DCGD})$$

Indeed, for $n = 1$ this method reduces to CGD. For unbiased compressors belonging to $\mathbb{U}(\zeta)$, this method converges under suitable assumptions on the functions. For instance, if f_i are L -smooth and f is μ -strongly convex, then as long as the stepsize is chosen appropriately,

the method converges to a $\mathcal{O}\left(\frac{\eta D(\zeta-1)}{\mu n}\right)$ neighborhood of the (necessarily unique) solution x^* with the linear rate

$$\mathcal{O}\left(\left(\frac{L}{\mu} + \frac{L(\zeta-1)}{\mu n}\right) \log \frac{1}{\epsilon}\right),$$

where $D := \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|_2^2$ (Gorbunov et al., 2020a). In particular, in the overparameterized setting when $D = 0$, the method converges to the exact solution, and does so at the same rate as GD as long as $\zeta = \mathcal{O}(n)$. These results hold even if a regularizer is considered, and a proximal step is added to DCGD. Moreover, as shown by Mishchenko et al. (2019) and Horváth et al. (2019b), a variance reduction technique can be devised to remove the neighborhood convergence and replace it by convergence to x^* , at the negligible additional cost of $\mathcal{O}((\zeta-1) \log \frac{1}{\epsilon})$.

5.2 Failure of DCGD with biased compressors

However, as we now demonstrate by giving some counter-examples, DCGD may fail if the compression operators are allowed to be biased. In the first example below, DCGD used with the Top-1 compressor diverges at an exponential rate.

Example 1 Consider $n = d = 3$ and define

$$f_1(x) = \langle a, x \rangle^2 + \frac{1}{4} \|x\|_2^2, \quad f_2(x) = \langle b, x \rangle^2 + \frac{1}{4} \|x\|_2^2, \quad f_3(x) = \langle c, x \rangle^2 + \frac{1}{4} \|x\|_2^2,$$

where $a = (-3, 2, 2)$, $b = (2, -3, 2)$ and $c = (2, 2, -3)$. Let the starting iterate be $x^0 = (t, t, t)$, where $t > 0$. Then

$$\nabla f_1(x^0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x^0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x^0) = \frac{t}{2}(9, 9, -11).$$

Using the Top-1 compressor, we get $\mathcal{C}(\nabla f_1(x^0)) = \frac{t}{2}(-11, 0, 0)$, $\mathcal{C}(\nabla f_2(x^0)) = \frac{t}{2}(0, -11, 0)$ and $\mathcal{C}(\nabla f_3(x^0)) = \frac{t}{2}(0, 0, -11)$. The next iterate of DCGD is

$$x^1 = x^0 - \eta \frac{1}{3} \sum_{i=1}^3 \mathcal{C}(\nabla f_i(x^0)) = \left(1 + \frac{11\eta}{6}\right) x^0.$$

Repeated application gives

$$x^k = \left(1 + \frac{11\eta}{6}\right)^k x^0.$$

Since $\eta > 0$, the entries of x^k diverge exponentially fast to $+\infty$.

The above counter-example can be extended to the case of Top- d_1 when $d_1 < \lceil \frac{d}{2} \rceil$.

Example 2 Fix the dimension $d \geq 3$ and let $n = \binom{d}{d_1}$ be the number of nodes, where $d_1 < \lceil \frac{d}{2} \rceil$ and $d_2 = d - d_1 > d_1$. Choose positive numbers $b, c > 0$ such that

$$-bd_1 + cd_2 = 1, \quad b > c + 1.$$

One possible choice could be $b = d_2 + \frac{d_2}{d_1}$, $c = d_1 + \frac{1}{d_2} + 1$. Define vectors $a_j \in \mathbb{R}^d$, $j \in [n]$ via

$$a_j = \sum_{i \in I_j} (-b)e_i + \sum_{i \in [d] \setminus I_j} ce_i,$$

where sets $I_j \subset [d]$, $j \in [n]$ are all possible d_1 -subsets of $[d]$ enumerated in some way. Define

$$f_j(x) = \langle a_j, x \rangle^2 + \frac{1}{2} \|x\|_2^2, \quad j \in [n]$$

and let the initial point be $x^0 = te$, $t > 0$, where $e = \sum_{i=1}^d e_i$ is the vector of all 1s. Then

$$\nabla f_j(x^0) = 2\langle a_j, x^0 \rangle \cdot a_j + x^0 = 2t(-bd_1 + cd_2) \cdot a_j + te = t(2a_j + e).$$

Since $|2(-b) + 1| > |2c + 1|$, then using the Top- d_1 compressor, we get

$$\mathcal{C}(\nabla f_j(x^0)) = -t(2b - 1) \sum_{i \in I_j} e_i.$$

Therefore, the next iterate of DCGD is

$$\begin{aligned} x^1 &= x^0 - \eta \frac{1}{n} \sum_{j=1}^n \mathcal{C}(\nabla f_j(x^0)) = x^0 + \frac{\eta t(2b - 1)}{n} \sum_{j=1}^n \sum_{i \in I_j} e_i \\ &= x^0 + \frac{\eta(2b - 1)}{n} \begin{pmatrix} d \\ d_1 - 1 \end{pmatrix} x^0 = \left(1 + \frac{\eta(2b - 1)d_1}{d_2 + 1} \right) x^0, \end{aligned}$$

which implies

$$x^k = \left(1 + \frac{\eta(2b - 1)d_1}{d_2 + 1} \right)^k x^0.$$

Since $\eta > 0$ and $b > 1$, the entries of x^k diverge exponentially fast to $+\infty$.

Finally, we present more general counter-example with different type of failure for DCGD when non-randomized compressors are used.

Example 3 Let $\mathcal{C}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ be any deterministic mapping for which there exist vectors $v_1, v_2, \dots, v_m \in \mathbb{R}^d$ such that

$$\sum_{i=1}^m v_i \neq 0, \quad \text{and} \quad \sum_{i=1}^m \mathcal{C}(v_i) = 0. \quad (18)$$

Consider the distributed optimization problem (17) with $n = m$ devices and with the following strongly convex loss functions

$$f_i(x) = \langle v_i, x \rangle + \frac{1}{2} \|x\|^2, \quad i \in [m].$$

Then $\nabla f_i(x) = v_i + x$ and $\nabla f(x) = \frac{1}{n} \sum_{i=1}^n v_i + x$. Hence, the optimal point $x^* = -\frac{1}{n} \sum_{i=1}^n v_i \neq 0$. However, it can be easily checked that, with initialization $x_0 = 0$, we have

$$x^1 = x^0 - \gamma \frac{1}{n} \sum_{i=1}^n \mathcal{C}(\nabla f_i(x^0)) = x^0 - \gamma \frac{1}{n} \sum_{i=1}^n \mathcal{C}(v_i + x^0) = x^0.$$

Thus, when initialized at $x_0 = 0$, not only DCGD does not converge to the solution $x^* \neq 0$, it remains stuck at the same initial point for all iterates, namely $x^k = x^0 = 0$ for all $k \geq 1$.

Condition (18) can be easily satisfied for specific biased compressors. For instance, Top-1 satisfies (18) with $v_1 = [\frac{1}{4}]$, $v_2 = [-\frac{1}{2}]$, $v_3 = [-\frac{1}{2}]$.

The above examples suggests that one needs to devise a different approach to solving the distributed problem (17) with biased compressors. We resolve this problem by employing a memory feedback mechanism.

5.3 Error Feedback

We show that distributed version of Distributed SGD with Error-Feedback (Karimireddy et al., 2019b), displayed in Algorithm 1, is able to resolve the issue. Moreover, this algorithm allows for the computation of stochastic gradients. Each step starts with all machines i in parallel computing a stochastic gradient g_i^k of the form

$$g_i^k = \nabla f_i(x^k) + \xi_i^k, \quad (19)$$

where $\nabla f_i(x^k)$ is the true gradient, and ξ_i^k is a stochastic error. Then, this is multiplied by a stepsize η^k and added to the memory/error-feedback term e_i^k , and subsequently compressed. The compressed messages are communicated and aggregated. The difference of message we wanted to send and its compressed version becomes stored as e_i^{k+1} for further correction in the next communication round. The output \bar{x}^K is an ergodic average of the form

$$\bar{x}^K := \frac{1}{W^K} \sum_{k=0}^K w^k x^k, \quad W^K := \sum_{k=0}^K w^k. \quad (20)$$

5.4 Complexity theory

We assume the stochastic error ξ_i^k in (19) satisfies the following condition.

Assumption 1 *Stochastic error ξ_i^k is unbiased, i.e. $\mathbb{E}[\xi_i^k] = 0$, and for some constants $B, C \geq 0$*

$$\mathbb{E} \left[\left\| \xi_i^k \right\|_2^2 \right] \leq B \left\| \nabla f_i(x^k) \right\|_2^2 + C, \quad \forall i \in [n], k \geq 0. \quad (24)$$

Note that this assumption is much weaker than the bounded variance assumption (i.e., $\mathbb{E} \left[\left\| \xi_i^k \right\|_2^2 \right] \leq C$) and bounded gradient assumption (i.e., $\mathbb{E} \left[\left\| g_i^k \right\|_2^2 \right] \leq C$). We can now state the main result of this section. To the best of our knowledge, *this was an open problem*: we are not aware of any convergence results for distributed optimization that tolerate general classes of *biased* compression operators and have reasonable assumptions on the stochastic gradient.

Algorithm 1 Distributed SGD with Biased Compression and Error Feedback

Parameters: Compressors $\mathcal{C}_i^k \in \mathbb{B}^3(\delta)$; Stepsizes $\{\eta^k\}_{k \geq 0}$; Iteration count K

Initialization: Choose $x^0 \in \mathbb{R}^d$ and $e_i^0 = 0$ for all i

for $k = 0, 1, 2, \dots, K$ **do**

Server sends x^k to all n machines

All machines in parallel perform these updates:

$$\tilde{g}_i^k = \mathcal{C}_i^k(e_i^k + \eta^k g_i^k) \quad (21)$$

$$e_i^{k+1} = e_i^k + \eta^k g_i^k - \tilde{g}_i^k \quad (22)$$

Each machine i sends \tilde{g}_i^k to the server

Server performs aggregation:

$$x^{k+1} = x^k - \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^k \quad (23)$$

end for

Output: Weighted average of the iterates: \bar{x}^K (20)

Theorem 21 (Convergence guarantees for Algorithm 1) *Let $\{x^k\}_{k \geq 0}$ denote the iterates of Algorithm 1 for solving problem (1), where each f_i is L -smooth and μ -strongly convex. Let x^* be the minimizer of f and let $f^* := f(x^*)$ and*

$$D := \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|_2^2.$$

Assume the compression operator used by all nodes is in $\mathbb{B}^3(\delta)$. Then we have the following convergence rates under three different stepsize and iterate weighting regimes:

(i) $\mathcal{O}(\frac{1}{k})$ **stepsizes & $\mathcal{O}(k)$ weights.** *Let, for all $k \geq 0$, the stepsizes and weights be set as $\eta^k = \frac{4}{\mu(\kappa+k)}$ and $w^k = \kappa + k$, respectively, where $\kappa = \frac{56(2\delta+B)L}{\mu}$. Then*

$$\mathbb{E} [f(\bar{x}^K)] - f^* = \mathcal{O} \left(\frac{A_1}{K^2} + \frac{A_2}{K} \right),$$

where $A_1 := \frac{L^2(2\delta+B)^2}{\mu} \|x^0 - x^\|_2^2$ and $A_2 := \frac{C(1+1/n)+D(2B/n+3\delta)}{\mu}$.*

(ii) $\mathcal{O}(1)$ **stepsizes & $\mathcal{O}(e^{-k})$ weights.** *Let, for all $k \geq 0$, the stepsizes and weights be set as $\eta^k = \eta$ and $w^k = (1 - \mu\eta/2)^{-(k+1)}$, respectively, where $\eta \leq \frac{1}{14(2\delta+B)L}$. Then*

$$\mathbb{E} [f(\bar{x}^K)] - f^* = \tilde{\mathcal{O}} \left(A_3 \exp \left[-\frac{K}{A_4} \right] + \frac{A_2}{K} \right),$$

where $A_3 := L(2\delta + B) \|x^0 - x^\|_2^2$ and $A_4 := \frac{28L(2\delta+B)}{\mu}$.*

(iii) $\mathcal{O}(1)$ **stepsizes & equal weights.** Let, for all $k \geq 0$, the stepsizes and weights be set as $\eta^k = \eta$ and $w^k = 1$, respectively, where $\eta \leq \frac{1}{14(2\delta+B)L}$. Then

$$\mathbb{E} [f(\bar{x}^K)] - f^* = \mathcal{O} \left(\frac{A_3}{K} + \frac{A_5}{\sqrt{K}} \right),$$

where $A_5 := \sqrt{C(1+1/n) + D(2B/n + 3\delta)} \|x^0 - x^*\|_2$.

Let us make a few observations on these results. First, Algorithm 1 employing general biased compressors and error feedback mechanism indeed *resolves* convergence issues of DCGD method by *converging the optimal solution* x^* . Second, note that the choice of stepsizes η^k and weights w^k leading to convergence is not unique and *several schedules are feasible*. Third, all the rates are sublinear and based on the second rate (ii) above, *linear convergence* is guaranteed if $C = D = 0$. Based on (24), one setup when the condition $C = 0$ holds is when all devices compute full local gradients (i.e., $g_i^k = \nabla f_i(x^k)$). Furthermore, the condition $D = 0$ is equivalent to $\nabla f_i(x^*) = 0$ for all $i \in [n]$, which is typically satisfied for over-parameterized models. Lastly, under these two assumptions (i.e., devices can compute full local gradients and the model is over-parameterized), we show that distributed SGD method with error feedback converges with the same $\mathcal{O} \left(\delta \frac{L}{\mu} \log \frac{1}{\epsilon} \right)$ linear rate as single node CGD algorithm. *To the best of our knowledge, this was the first regime where distributed first order method with biased compression is guaranteed to converge linearly.*

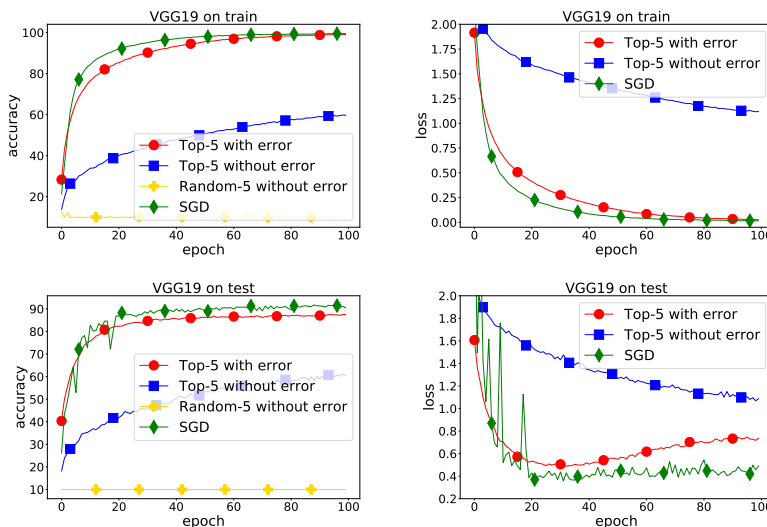


Figure 4: Training/Test loss and accuracy for VGG19 on CIFAR10 distributed among 4 nodes for 4 different compression operators.

6. Experiments

In Sections 6.1–6.4, we present our experiments, which are primarily focused on supporting our theoretical findings. Therefore, we simulate these experiments on one machine which

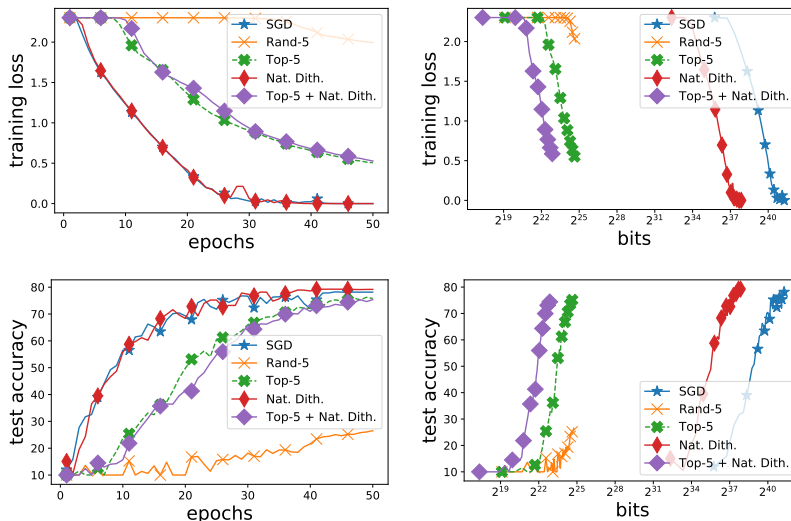


Figure 5: Training loss and test accuracy for VGG11 on CIFAR10 distributed among 4 nodes for 5 different compression operators.

enable us to do rapid direct comparisons against the prior methods. In more details, we use the machine with 24 Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz cores and GPU GeForce GTX 1080 Ti. Section 6.5 is devoted to real experiments with a large model and big data. For these experiments, we use a computational cluster with 10 GPUs Tesla T4. We implement all methods in Python 3.7 using Pytorch Paszke et al. (2019).

6.1 Lower empirical variance induced by biased compressors during deep network training

Motivated by our theoretical results in Section 4, we show that similar behaviour can be seen in the empirical variance of gradients. We run 2 sets of experiments with Resnet18 on CIFAR10 dataset. In Figure 6, we display empirical variance, which is obtained by running a training procedure with specific compression. We compare unbiased and biased compressions with the same communication complexities—deterministic with classic/unbiased \mathcal{C}_{nat} and Top- k with Rand- k with k to be $1/5$ of coordinates. One can clearly see, that there is a gap in empirical variance between biased and unbiased methods, similar to what we have shown in theory, see Section 4.

6.2 Error-feedback is needed in distributed training with biased compression

The next experiment shows the need of error-feedback for methods with biased compression operators. Based on Example 1, error feedback is necessary to prevent divergence from the optimal solution. Figure 4 displays training/test loss and accuracy for VGG19 on CIFAR10 with data equally distributed among 4 nodes. We use plain SGD with a default step size equal to 0.01 for all methods, i.e. Top-5 with and without error feedback, Rand-5 and no compression. As suggested by the counterexample, not using error feedback can really hurt

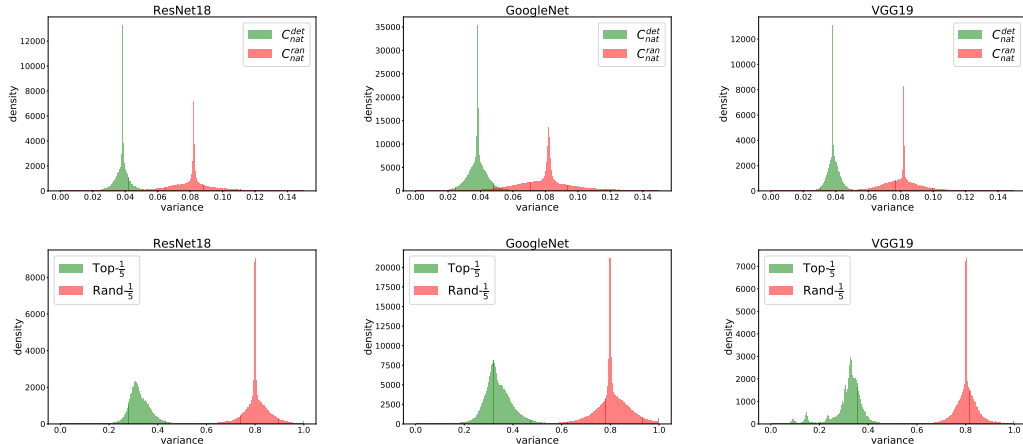


Figure 6: Comparison of empirical variance $\|\mathcal{C}(x) - x\|_2^2 / \|x\|_2^2$ during training procedure for two pairs of method– deterministic with classic/unbiased C_{nat} and Top- k with Rand- k with Top- $1/5$ of coordinates. Both of the plots were produced using ResNet18, GoogleNet, and VGG19 on CIFAR10 dataset.

the performance when biased compressions are used. Also note, that performance of Rand-5 is significantly worse than Top-5.

6.3 Top- k mixed with natural dithering saves in communication significantly

Next, we experimentally show the superiority of our newly proposed compressor–Top- k combined with natural dithering. We compare this against current state-of-the-art for low bandwidth approach Top- k for some small k . In Figure 5, we plot comparison of 5 methods–Top- k , Rand- k , natural dithering, Top- k combined with natural dithering and plain SGD. We use 2 levels with infinity norm for natural dithering and $k = 5$ for sparsification methods. For all the compression operators, we train VGG11 on CIFAR10 with plain SGD as an optimizer and default step size equal to 0.01. We can see that adding natural dithering after Top- k has the same effect as the natural dithering comparing to no compression, which is a significant reduction in communications without almost no effect on convergence or generalization. Using this intuition, one can come to the conclusion that Top- k with natural dithering is the best compression operator for any bandwidth, where we adjust to given bandwidth by adjusting k . This exactly matches with our previous theoretical variance estimates displayed in Figure 3.

6.4 Theoretical behavior predicts the actual performance in practice

In the next experiment, we provide numerical results to further show that our predicted theoretical behavior matches the actual performance observed in practice. We run two regression experiments optimized by gradient descent with step-size $\eta = \frac{1}{L}$. We use a slightly

adjusted version of Theorem 19 with adaptive step-sizes, namely

$$\frac{f(x^k) - f(x^*)}{f(x^0) - f(x^*)} \leq \prod_{i=1}^k \left(1 - \frac{\mu}{L\delta_i}\right),$$

where

$$1 - \frac{1}{\delta_i} = \frac{\|\mathcal{C}(\nabla f(x^i)) - \nabla f(x^i)\|_2^2}{\|\nabla f(x^i)\|_2^2}.$$

Note that this is the direct consequence of our analysis. We apply this property to display the theoretical convergence. In the first experiment depicted in Figure 7, we randomly generate random square matrix A of dimension 100 where it is constructed in the following way: we sample random diagonal matrix D , which elements are independently sampled from the uniform distribution $(1, 10)$, $(1, 100)$, and $(1, 1000)$, respectively. A is then constructed using $Q^\top D Q$, where $P = QR$ is a random matrix and QR is obtained using QR-decomposition. The label y is generated the same way from the uniform distribution $(0, 1)$. The optimization objective is then

$$\min_{x \in \mathbb{R}^d} x^\top A x - y^\top x.$$

For the second experiment shown in Figure 8, we run standard linear regression on two scikit-learn datasets—*Boston* and *Diabetes*—and applied data normalization as the preprocessing step.

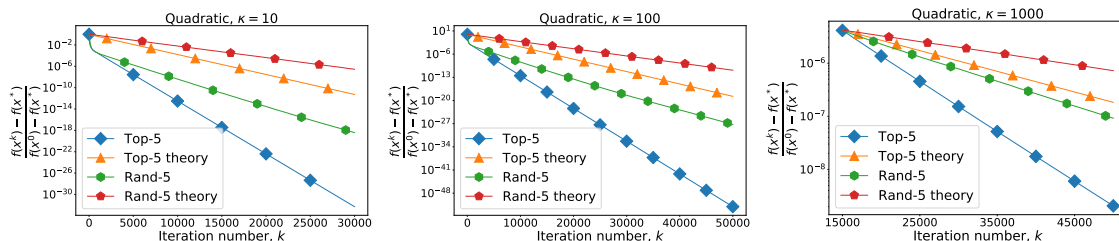


Figure 7: Theoretical vs. Practical Convergence of Compressed Gradient Descent on Quadratics problem with different condition number κ for Top-5 and Rand-5 compression operators.

Looking into Figures 7 and 8, one can clearly see that as predicted by our theory, biased compression with less empirical variance leads to better convergence in practice and the gap almost matches the improvement.

6.5 Transformer training

In the last experiment, we work with a real big model. In particular, we train ALBERT-large (Lan et al., 2020) (18M parameters) with layer sharing on a combination of Book-corpora (Zhu et al., 2015) and Wikipedia (Devlin et al., 2018) datasets. We use the same optimizer (LAMB) and the same tuning for it as in the original paper (Lan et al., 2020). Our goal is to find an unbiased and biased operators such that we maximize the improvement in terms of communication cost without losing much in terms of training quality. In this case,

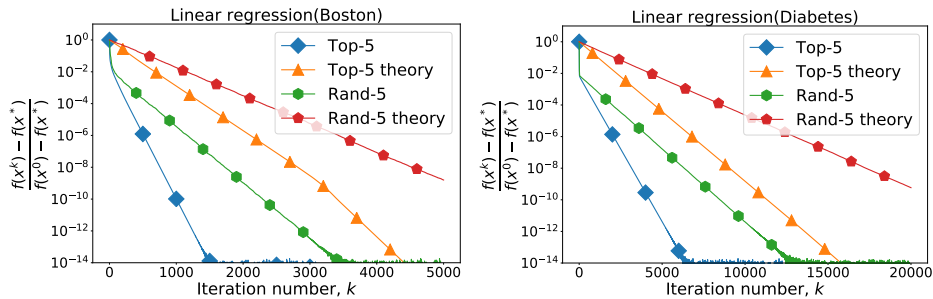


Figure 8: Theoretical vs. Practical Convergence of Compressed Gradient Descent on Linear Regression problem for *Boston* and *Diabetes* datasets with Top-5 and Rand-5 compression operators.

we include in the communication cost both the time to perform the communication round and the time to perform the compression and decompression operations. It is important to note that we do not compress packages with gradients corresponding to LayerNorm scales, but this is less than 1 percent of the whole package. Among unbiased compressors, we try natural compression (Section 2.2 (g)) and random sparsification (Section 2.2 (a)). The best result is shown by natural compression, which compresses the packages by a factor of 4. The Rand-25 operator (which also compresses the information by a factor of 4) performs much worse even with the use of the error feedback technique. For communications with natural compression, we use the classical allreduce procedure. Among unbiased compressors, we try Top- k sparsification (Section 2.2 (d)) and Power compression (Vogels et al., 2019). The best result is shown by Power compression with the rank parameter $r=8$ and the error feedback. The organization of communications (allreduce procedures) occurs as in the original paper (Vogels et al., 2019). We measure how the training loss changes (Figure 9) as well as at the end of training we evaluate the final performance for each model on several popular tasks from (Wang et al., 2018) (Table 5). The results show that the use of biased compression can significantly reduce the communication time cost compared to uncompressed and even unbiased compression setups. At the same time, the quality of the training does not drop much.

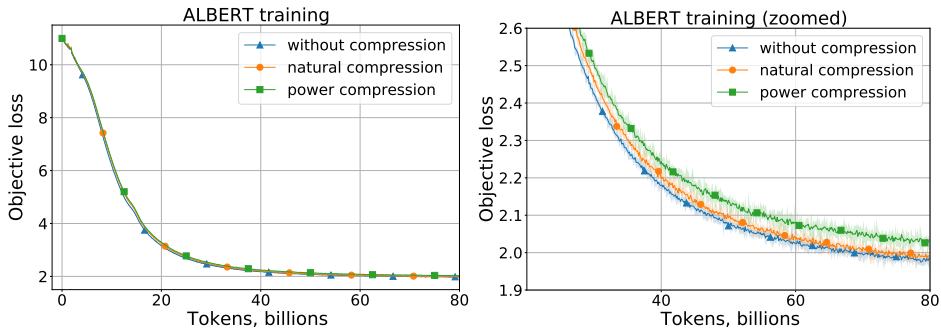


Figure 9: Training objective value for ALBERT-large with different compression operators.

Setup	Avg time	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STS-B
Without compression	9.62 ± 0.03	46.2	81.1	82.5	87.9	88.0	66.3	85.1	88.0
Natural compression	4.05 ± 0.05	48.3	81.0	87.5	87.8	84.4	63.2	87.8	86.9
Power compression	1.04 ± 0.04	42.4	80.3	85.1	88.2	85.3	46.3	88.0	87.4

Table 5: Comparison of average time per one communication round and downstream evaluation scores on GLUE benchmark tasks.

Appendix

Appendix A. Basic Facts and Inequalities

A.1 Strong convexity

Function f is strongly convex on \mathbb{R}^d when it is continuously differentiable and there is a constant $\mu > 0$ such that the following inequality holds:

$$\frac{\mu}{2} \|x - y\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d. \quad (25)$$

A.2 Smoothness

Function f is called L -smooth in \mathbb{R}^d with $L > 0$ when it is differentiable and its gradient is L -Lipschitz continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2, \quad \forall x, y \in \mathbb{R}^d.$$

If convexity is assumed as well, then the following inequalities hold:

$$\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d \quad (26)$$

By plugging $y = x^*$ to (26), we get

$$\|\nabla f(x)\|_2^2 \leq 2L(f(x) - f(x^*)), \quad \forall x \in \mathbb{R}^d. \quad (27)$$

A.3 Useful inequalities

For all $a, b, x_1, \dots, x_n \in \mathbb{R}^d$ and $\xi > 0$ the following inequalities holds:

$$2\langle a, b \rangle \leq \frac{\|a\|_2^2}{\xi} + \xi \|b\|_2^2, \quad (28)$$

$$\|a + b\|_2^2 \leq \left(1 + \frac{1}{\xi}\right) \|a\|_2^2 + (1 + \xi) \|b\|_2^2, \quad (29)$$

$$\left\| \sum_{i=1}^n x_i \right\|_2^2 \leq n \cdot \sum_{i=1}^n \|x_i\|_2^2. \quad (30)$$

A.4 Facts from order statistics

For i.i.d. samples x_1, x_2, \dots, x_d from an absolutely continuous distribution with probability density function ϕ and cumulative distribution function Φ let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(d)}$ be the order statistics obtained by arranging samples in increasing order of magnitude. Then the following expressions give the density function of $x_{(i)}$ ($1 \leq i \leq d$)

$$\phi_{x_{(i)}}(u) = \frac{d!}{(i-1)!(d-i)!} \{\Phi(u)\}^{i-1} \{1 - \Phi(u)\}^{n-i} \phi(u), \quad -\infty < u < \infty, \quad (31)$$

and the joint density function of all d order statistics

$$\phi_{x_{(1)}, \dots, x_{(d)}}(u_1, \dots, u_d) = d! \prod_{i=1}^d \phi(u_i), \quad -\infty < u_1 \leq \dots \leq u_d < \infty. \quad (32)$$

Appendix B. Proofs for Section 2.2

B.1 Proof of Lemma 8: Unbiased Random Sparsification

From the definition of k -nice sampling we have $p_i := \text{Prob}(i \in S) = \frac{k}{d}$. Hence

$$\begin{aligned} \mathbb{E}[\mathcal{C}(x)] &= \frac{d}{k} \mathbb{E} \left[\sum_{i \in S} x_i e_i \right] = \frac{d}{k} \sum_{i=1}^d p_i x_i e_i = \sum_{i=1}^d x_i e_i = x, \\ \mathbb{E}[\|\mathcal{C}(x)\|_2^2] &= \frac{d^2}{k^2} \mathbb{E} \left[\sum_{i \in S} x_i^2 \right] = \frac{d^2}{k^2} \sum_{i=1}^d p_i x_i^2 = \frac{d}{k} \sum_{i=1}^d x_i^2 = \frac{d}{k} \|x\|_2^2, \end{aligned}$$

which implies $\mathcal{C} \in \mathbb{U}(\frac{d}{k})$.

B.2 Proof of Lemma 9: Biased Random Sparsification

Let $S \subseteq [d]$ be a proper sampling with probability vector $p = (p_1, \dots, p_d)$, where $p_i := \text{Prob}(i \in S) > 0$ for all i . Then

$$\mathbb{E}[\mathcal{C}(x)] = \mathbf{Diag}(p) x = \sum_{i=1}^d p_i x_i e_i \quad \text{and} \quad \mathbb{E}[\|\mathcal{C}(x)\|_2^2] = \sum_{i=1}^d p_i x_i^2.$$

Letting $q := \min_i p_i$, we get

$$q \|x\|_2^2 \leq \sum_{i=1}^d p_i x_i^2 = \mathbb{E}[\|\mathcal{C}(x)\|_2^2] = \langle \mathbb{E}[\mathcal{C}(x)], x \rangle.$$

So, $\mathcal{C} \in \mathbb{B}^1(q, 1)$ and $\mathcal{C} \in \mathbb{B}^2(q, 1)$. For the third class, note that

$$\mathbb{E}[\|\mathcal{C}(x) - x\|_2^2] = \sum_{i=1}^d (1 - p_i) x_i^2 \leq (1 - q) \|x\|_2^2.$$

Hence, $\mathcal{C} \in \mathbb{B}^3(\frac{1}{q})$.

B.3 Proof of Lemma 10: Adaptive Random Sparsification

From the definition of the compression operator, we have

$$\begin{aligned} \mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] &= \mathbb{E} [x_i^2] = \sum_{i=1}^d \frac{|x_i|}{\|x\|_1} x_i^2 = \frac{\|x\|_3^3}{\|x\|_1}, \\ \mathbb{E} [\langle \mathcal{C}(x), x \rangle] &= \mathbb{E} [x_i^2] = \frac{\|x\|_3^3}{\|x\|_1}, \end{aligned}$$

whence $\beta = 1$. Furthermore, by Chebychev's sum inequality, we have

$$\frac{1}{d^2} \|x\|_1 \|x\|_2^2 = \left(\sum_{i=1}^d \frac{1}{d} |x_i| \right) \left(\sum_{i=1}^d \frac{1}{d} x_i^2 \right) \leq \sum_{i=1}^d \frac{1}{d} |x_i| x_i^2 = \frac{1}{d} \|x\|_3^3,$$

which implies that $\alpha = \frac{1}{d}$, $\delta = d$. So, $\mathcal{C} \in \mathbb{B}^1(\frac{1}{d}, 1)$, $\mathcal{C} \in \mathbb{B}^2(\frac{1}{d}, 1)$, and $\mathcal{C} \in \mathbb{B}^3(d)$.

B.4 Proof of Lemma 11: Top- k sparsification

Clearly, $\|\mathcal{C}(x)\|_2^2 = \sum_{i=d-k+1}^d x_{(i)}^2$ and $\|\mathcal{C}(x) - x\|_2^2 = \sum_{i=1}^{d-k} x_{(i)}^2$. Hence

$$\frac{k}{d} \|x\|_2^2 \leq \|\mathcal{C}(x)\|_2^2 = \langle \mathcal{C}(x), x \rangle \leq \|x\|_2^2, \quad \|\mathcal{C}(x) - x\|_2^2 \leq \left(1 - \frac{k}{d}\right) \|x\|_2^2.$$

So, $\mathcal{C} \in \mathbb{B}^1(\frac{k}{d}, 1)$, $\mathcal{C} \in \mathbb{B}^2(\frac{k}{d}, 1)$, and $\mathcal{C} \in \mathbb{B}^3(\frac{d}{k})$.

B.5 Proof of Lemma 12: General Unbiased Rounding

The unbiasedness follows immediately from the definition (12)

$$\mathbb{E} [\mathcal{C}(x)] = \sum_{i=1}^d \mathbb{E} [\mathcal{C}(x)_i] e_i = \sum_{i=1}^d \text{sign}(x_i) \left(a_k \frac{a_{k+1} - |x_i|}{a_{k+1} - a_k} + a_{k+1} \frac{|x_i| - a_k}{a_{k+1} - a_k} \right) e_i = \sum_{i=1}^d x_i e_i = x. \quad (33)$$

Since the rounding compression operator \mathcal{C} applies to each coordinate independently, without loss of generality we can consider the compression of scalar values $x = t > 0$ and show that $\mathbb{E} [\mathcal{C}(t)^2] \leq \zeta \cdot t^2$. From the definition we compute the second moment as follows

$$\mathbb{E} [\mathcal{C}(t)^2] = a_k^2 \frac{a_{k+1} - t}{a_{k+1} - a_k} + a_{k+1}^2 \frac{t - a_k}{a_{k+1} - a_k} = (a_k + a_{k+1})t - a_k a_{k+1} = t^2 + (t - a_k)(a_{k+1} - t), \quad (34)$$

from which

$$\frac{\mathbb{E} [\mathcal{C}(t)^2]}{t^2} = 1 + \left(1 - \frac{a_k}{t}\right) \left(\frac{a_{k+1}}{t} - 1\right), \quad a_k \leq t \leq a_{k+1}. \quad (35)$$

Checking the optimality condition, one can show that the maximum is achieved at

$$t_* = \frac{2a_k a_{k+1}}{a_k + a_{k+1}} = \frac{2}{\frac{1}{a_k} + \frac{1}{a_{k+1}}},$$

which being the harmonic mean of a_k and a_{k+1} , is in the range $[a_k, a_{k+1}]$. Plugging it to the expression for variance we get

$$\frac{\mathbb{E}[\mathcal{C}(t_*)^2]}{t_*^2} = 1 + \frac{1}{4} \left(1 - \frac{a_k}{a_{k+1}}\right) \left(\frac{a_{k+1}}{a_k} - 1\right) = \frac{1}{4} \left(\frac{a_k}{a_{k+1}} + \frac{a_{k+1}}{a_k} + 2\right).$$

Thus, the parameter ζ for general unbiased rounding would be

$$\zeta = \sup_{t>0} \frac{\mathbb{E}[\mathcal{C}(t)^2]}{t^2} = \sup_{k \in \mathbb{Z}} \sup_{a_k \leq t \leq a_{k+1}} \frac{\mathbb{E}[\mathcal{C}(t)^2]}{t^2} = \frac{1}{4} \sup_{k \in \mathbb{Z}} \left(\frac{a_k}{a_{k+1}} + \frac{a_{k+1}}{a_k} + 2\right) \geq 1.$$

B.6 Proof of Lemma 13: General Biased Rounding

From the definition (13) of compression operator \mathcal{C} we derive the following inequalities

$$\begin{aligned} \inf_{k \in \mathbb{Z}} \left(\frac{2a_k}{a_k + a_{k+1}}\right)^2 \|x\|_2^2 &\leq \|\mathcal{C}(x)\|_2^2, \\ \|\mathcal{C}(x)\|_2^2 &\leq \sup_{k \in \mathbb{Z}} \frac{2a_{k+1}}{a_k + a_{k+1}} \langle \mathcal{C}(x), x \rangle, \\ \inf_{k \in \mathbb{Z}} \frac{2a_k}{a_k + a_{k+1}} \|x\|_2^2 &\leq \langle \mathcal{C}(x), x \rangle, \end{aligned}$$

which imply that $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$ and $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$, with

$$\beta = \sup_{k \in \mathbb{Z}} \frac{2a_{k+1}}{a_k + a_{k+1}}, \quad \gamma = \inf_{k \in \mathbb{Z}} \frac{2a_k}{a_k + a_{k+1}}, \quad \alpha = \gamma^2.$$

For the third class $\mathbb{B}^3(\delta)$, we need to upper bound the ratio $\|\mathcal{C}(x) - x\|_2^2 / \|x\|_2^2$. Again, as \mathcal{C} applies to each coordinate independently, without loss of generality we consider the case when $x = t > 0$ is a scalar. From definition (13), we get

$$\frac{(\mathcal{C}(t) - t)^2}{t^2} = \min \left[\left(1 - \frac{a_k}{t}\right)^2, \left(1 - \frac{a_{k+1}}{t}\right)^2 \right], \quad a_k \leq t \leq a_{k+1}. \quad (36)$$

It can be easily checked that $\left(1 - \frac{a_k}{t}\right)^2$ is an increasing function and $\left(1 - \frac{a_{k+1}}{t}\right)^2$ is a decreasing function of $t \in [a_k, a_{k+1}]$. Thus, the maximum is achieved when they are equal. In contrast to unbiased general rounding, it happens at the middle of the interval,

$$t_* = \frac{a_k + a_{k+1}}{2} \in [a_k, a_{k+1}].$$

Plugging t_* into (36), we get

$$\frac{(\mathcal{C}(t_*) - t_*)^2}{t_*^2} = \left(\frac{a_{k+1} - a_k}{a_{k+1} + a_k}\right)^2.$$

Given this, the parameter δ can be computed from

$$1 - \frac{1}{\delta} = \sup_{k \in \mathbb{Z}} \sup_{a_k \leq t \leq a_{k+1}} \frac{(\mathcal{C}(t) - t)^2}{t^2} = \sup_{k \in \mathbb{Z}} \left(\frac{a_{k+1} - a_k}{a_{k+1} + a_k}\right)^2,$$

which gives

$$\delta = \sup_{k \in \mathbb{Z}} \frac{(a_k + a_{k+1})^2}{4a_k a_{k+1}} \geq 1,$$

and $\mathcal{C} \in \mathbb{B}^3(\delta)$.

B.7 Proof of Lemma 15: General Exponential Dithering

The proof goes with the same steps as in Theorem 4 of Horváth et al. (2019a). To show the unbiasedness of \mathcal{C} , first we show the unbiasedness of $\xi(t)$ for $t \in [0, 1]$ in the same way as (33) was done. Then we note that

$$\mathbb{E}[\mathcal{C}(x)] = \text{sign}(x) \times \|x\|_p \times \mathbb{E}\left[\xi\left(\frac{|x|}{\|x\|_p}\right)\right] = \text{sign}(x) \times \|x\|_p \times \left(\frac{|x|}{\|x\|_p}\right) = x.$$

To compute the parameter ζ , we first estimate the second moment of ξ as follows:

$$\begin{aligned} &\leq \mathbb{1}\left(\frac{|x_i|}{\|x\|_p} \geq b^{1-s}\right) \cdot \frac{1}{4} \left(b + \frac{1}{b} + 2\right) \frac{x_i^2}{\|x\|_p^2} + \mathbb{1}\left(\frac{|x_i|}{\|x\|_p} < b^{1-s}\right) \cdot \frac{|x_i|}{\|x\|_p} b^{1-s} \\ &\leq \frac{1}{4} \left(b + \frac{1}{b} + 2\right) \frac{x_i^2}{\|x\|_p^2} + \mathbb{1}\left(\frac{|x_i|}{\|x\|_p} < b^{1-s}\right) \cdot \frac{|x_i|}{\|x\|_p} b^{1-s}. \end{aligned}$$

Then we use this bound to estimate the second moment of compressor \mathcal{C} :

$$\begin{aligned} \mathbb{E}\left[\|\mathcal{C}(x)\|_2^2\right] &= \|x\|_p^2 \sum_{i=1}^d \mathbb{E}\left[\xi\left(\frac{|x_i|}{\|x\|_p}\right)^2\right] \\ &\leq \|x\|_p^2 \sum_{i=1}^d \left(\frac{1}{4} \left(b + \frac{1}{b} + 2\right) \frac{x_i^2}{\|x\|_p^2} + \mathbb{1}\left(\frac{|x_i|}{\|x\|_p} < b^{1-s}\right) \cdot \frac{|x_i|}{\|x\|_p} b^{1-s}\right) \\ &= \frac{1}{4} \left(b + \frac{1}{b} + 2\right) \|x\|_2^2 + \sum_{i=1}^d \mathbb{1}\left(\frac{|x_i|}{\|x\|_p} < b^{1-s}\right) \cdot |x_i| \|x\|_p b^{1-s} \\ &\leq \frac{1}{4} \left(b + \frac{1}{b} + 2\right) \|x\|_2^2 + \min(\|x\|_1 \|x\|_p b^{1-s}, d \|x\|_p^2 b^{2-2s}) \\ &\leq \frac{1}{4} \left(b + \frac{1}{b} + 2\right) \|x\|_2^2 + \min(d^{1/2} \|x\|_2 \|x\|_p b^{1-s}, d \|x\|_p^2 b^{2-2s}) \\ &\leq \left[\frac{1}{4} \left(b + \frac{1}{b} + 2\right) + d^{1/r} b^{1-s} \min(1, d^{1/r} b^{1-s})\right] \|x\|_2^2 \\ &= \zeta_b \|x\|_2^2, \end{aligned}$$

where $r = \min(p, 2)$ and Hölder's inequality is used to bound $\|x\|_p \leq d^{1/p-1/2} \|x\|_2$ in case of $0 \leq p \leq 2$ and $\|x\|_p \leq \|x\|_2$ in the case $p \geq 2$.

B.8 Proof of Lemma 16: Top- k Combined with Exponential Dithering

From the unbiasedness of general dithering operator \mathcal{C}_{dith} we have

$$\mathbb{E}[\mathcal{C}(x)] = \mathbb{E}[\mathcal{C}_{dith}(\mathcal{C}_{top}(x))] = \mathcal{C}_{top}(x),$$

from which we conclude $\langle \mathbb{E}[\mathcal{C}(x)], x \rangle = \langle \mathcal{C}_{top}(x), x \rangle = \|\mathcal{C}_{top}(x)\|_2^2$. Next, using Lemma 15 on exponential dithering we get

$$\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \leq \zeta_b \cdot \|\mathcal{C}_{top}(x)\|_2^2 = \zeta_b \cdot \langle \mathbb{E}[\mathcal{C}(x)], x \rangle,$$

which implies $\beta = \zeta_b$. Using Lemma 11 we show $\gamma = \frac{k}{d}$ as $\langle \mathbb{E}[\mathcal{C}(x)], x \rangle = \|\mathcal{C}_{top}(x)\|_2^2 \geq \frac{k}{d} \|x\|_2^2$. Utilizing the derivations (34) and (35) it can be shown that $\mathbb{E} \left[\|\mathcal{C}_{dith}(x)\|_2^2 \right] \geq \|x\|_2^2$ and therefore

$$\mathbb{E} \left[\|\mathcal{C}(x)\|_2^2 \right] \geq \|\mathcal{C}_{top}(x)\|_2^2 \geq \frac{k}{d} \|x\|_2^2.$$

Hence, $\alpha = \frac{k}{d}$. To compute the parameter δ we use Theorem 6, which yields $\delta = \frac{\beta}{\gamma} = \frac{d}{k} \zeta_b$.

Appendix C. Proofs for Section 3

We now perform analysis of CGD for compression operators in \mathbb{B}^1 , \mathbb{B}^2 and \mathbb{B}^3 , establishing Theorems 17, 18 and 19, respectively.

C.1 Analysis for $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$

Lemma 22 *Assume f is L -smooth. Let $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$. Then as long as $0 \leq \eta \leq \frac{2}{\beta L}$, for each $x \in \mathbb{R}^d$ we have*

$$\mathbb{E} [f(x - \eta \mathcal{C}(\nabla f(x)))] \leq f(x) - \alpha \eta \left(1 - \frac{\eta \beta L}{2} \right) \|\nabla f(x)\|_2^2.$$

Proof Letting $g = \nabla f(x)$, we have[‡]

$$\begin{aligned} \mathbb{E} [f(x - \eta \mathcal{C}(g))] &\leq \mathbb{E} \left[f(x) + \langle g, -\eta \mathcal{C}(g) \rangle + \frac{L}{2} \|\eta \mathcal{C}(g)\|_2^2 \right] \\ &= f(x) - \eta \langle \mathbb{E}[\mathcal{C}(g)], g \rangle + \frac{\eta^2 L}{2} \mathbb{E} [\|\mathcal{C}(g)\|_2^2] \\ &\stackrel{(3)}{\leq} f(x) - \eta \langle \mathbb{E}[\mathcal{C}(g)], g \rangle + \frac{\eta^2 \beta L}{2} \langle \mathbb{E}[\mathcal{C}(g)], g \rangle \\ &= f(x) - \eta \left(1 - \frac{\eta \beta L}{2} \right) \langle \mathbb{E}[\mathcal{C}(g)], g \rangle \\ &\stackrel{(3)}{\leq} f(x) - \frac{\alpha}{\beta} \eta \left(1 - \frac{\eta \beta L}{2} \right) \|g\|_2^2. \end{aligned}$$

[‡]Alternatively, we can write

$$\begin{aligned} \mathbb{E} [f(x - \eta \mathcal{C}(g))] &\leq f(x) - \eta \langle \mathbb{E}[\mathcal{C}(g)], g \rangle + \frac{\eta^2 L}{2} \mathbb{E} [\|\mathcal{C}(g)\|_2^2] \\ &\stackrel{(3)}{\leq} f(x) - \frac{\eta}{\beta} \mathbb{E} [\|\mathcal{C}(g)\|_2^2] + \frac{\eta^2 L}{2} \mathbb{E} [\|\mathcal{C}(g)\|_2^2] \\ &= f(x) - \frac{\eta}{\beta} \left(1 - \frac{\eta \beta L}{2} \right) \mathbb{E} [\|\mathcal{C}(g)\|_2^2] \\ &\stackrel{(3)}{\leq} f(x) - \frac{\alpha}{\beta} \eta \left(1 - \frac{\eta \beta L}{2} \right) \|g\|_2^2. \end{aligned}$$

Both approaches lead to the same bound.

■

Proof of Theorem 17 Proof Since f is μ -strongly convex, $\|\nabla f(x^k)\|_2^2 \geq 2\mu(f(x^k) - f(x^*))$. Combining this with Lemma 22 applied to $x = x^k$ and $g = \nabla f(x^k)$, we get

$$\begin{aligned} \mathbb{E} \left[f \left(x^k - \eta \mathcal{C}(\nabla f(x^k)) \right) \right] - f(x^*) &\leq f(x^k) - f(x^*) - \frac{\alpha}{\beta} \eta \mu (2 - \eta \beta L) (f(x^k) - f(x^*)) \\ &= \left(1 - \frac{\alpha}{\beta} \eta \mu (2 - \eta \beta L) \right) (f(x^k) - f(x^*)). \end{aligned}$$

■

C.2 Analysis for $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$

Lemma 23 Assume f is L -smooth. Let $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$. Then as long as $0 \leq \eta \leq \frac{2}{\beta L}$, for each $x \in \mathbb{R}^d$ we have

$$\mathbb{E} [f(x - \eta \mathcal{C}(\nabla f(x)))] \leq f(x) - \gamma \eta \left(1 - \frac{\eta \beta L}{2} \right) \|\nabla f(x)\|_2^2.$$

Proof Letting $g = \nabla f(x)$, we have

$$\begin{aligned} \mathbb{E} [f(x - \eta \mathcal{C}(g))] &\leq \mathbb{E} \left[f(x) + \langle g, -\eta \mathcal{C}(g) \rangle + \frac{L}{2} \|\eta \mathcal{C}(g)\|_2^2 \right] \\ &= f(x) - \eta \langle \mathbb{E} [\mathcal{C}(g)], g \rangle + \frac{\eta^2 L}{2} \mathbb{E} [\|\mathcal{C}(g)\|_2^2] \\ &\stackrel{(6)}{\leq} f(x) - \eta \left(1 - \frac{\eta \beta L}{2} \right) \langle \mathbb{E} [\mathcal{C}(g)], g \rangle \\ &\stackrel{(6)}{\leq} f(x) - \gamma \eta \left(1 - \frac{\eta \beta L}{2} \right) \|g\|_2^2. \end{aligned}$$

■

Proof of Theorem 18 Proof Since f is μ -strongly convex, $\|\nabla f(x^k)\|_2^2 \geq 2\mu(f(x^k) - f(x^*))$. Combining this with Lemma 23 applied to $x = x^k$ and $g = \nabla f(x^k)$, we get

$$\begin{aligned} \mathbb{E} \left[f \left(x^k - \eta \mathcal{C}(\nabla f(x^k)) \right) \right] - f(x^*) &\leq f(x^k) - f(x^*) - \mu \gamma \eta (2 - \eta \beta L) (f(x^k) - f(x^*)) \\ &= (1 - \mu \gamma \eta (2 - \eta \beta L)) (f(x^k) - f(x^*)). \end{aligned}$$

■

C.3 Analysis for $\mathcal{C} \in \mathbb{B}^3(\delta)$

Lemma 24 *Assume f is L -smooth. Let $\mathcal{C} \in \mathbb{B}^3(\delta)$. Then as long as $0 \leq \eta \leq \frac{1}{L}$, for each $x \in \mathbb{R}^d$ we have*

$$\mathbb{E}[f(x - \eta\mathcal{C}(\nabla f(x)))] \leq f(x) - \frac{\eta}{2\delta} \|\nabla f(x)\|_2^2.$$

Proof Letting $g = \nabla f(x)$, note that for any stepsize $\eta \in \mathbb{R}$ we have

$$\begin{aligned} \mathbb{E}[f(x - \eta\mathcal{C}(g))] &\leq \mathbb{E}\left[f(x) + \langle g, -\eta\mathcal{C}(g) \rangle + \frac{L}{2} \|\eta\mathcal{C}(g)\|_2^2\right] \\ &= f(x) - \eta\langle \mathbb{E}[\mathcal{C}(g)], g \rangle + \frac{\eta^2 L}{2} \mathbb{E}[\|\mathcal{C}(g)\|_2^2]. \end{aligned} \quad (37)$$

Since $\mathcal{C} \in \mathbb{B}^3(\delta)$, we have $\mathbb{E}[\|\mathcal{C}(g) - g\|_2^2] \leq (1 - \frac{1}{\delta}) \|g\|_2^2$. Expanding the square, we get

$$\|g\|_2^2 - 2\mathbb{E}[\langle \mathcal{C}(g), g \rangle] + \mathbb{E}[\|\mathcal{C}(g)\|_2^2] \leq \left(1 - \frac{1}{\delta}\right) \|g\|_2^2.$$

Subtracting $\|g\|_2^2$ from both sides, and multiplying both sides by $\frac{\eta}{2}$ (now we assume that $\eta > 0$), we get

$$-\eta\langle \mathbb{E}[\mathcal{C}(g)], g \rangle + \frac{\eta}{2} \mathbb{E}[\|\mathcal{C}(g)\|_2^2] \leq -\frac{\eta}{2\delta} \|g\|_2^2.$$

Assuming that $\eta L \leq 1$, we can combine this with (37) and the lemma is proved. ■

Proof of Theorem 19 **Proof** Since f is μ -strongly convex, $\|\nabla f(x^k)\|_2^2 \geq 2\mu(f(x^k) - f(x^*))$. Combining this with Lemma 24 applied to $x = x^k$ and $g = \nabla f(x^k)$, we get

$$\begin{aligned} \mathbb{E}\left[f\left(x^k - \eta\mathcal{C}(\nabla f(x^k))\right)\right] - f(x^*) &\leq f(x^k) - f(x^*) - \frac{\eta\mu}{\delta} (f(x^k) - f(x^*)) \\ &= \left(1 - \frac{\eta\mu}{\delta}\right) (f(x^k) - f(x^*)). \end{aligned}$$
■

Appendix D. Proofs for Section 4

Proof [Proof of Lemma 20] **(a)** As it was already mentioned, we have the following expressions for ω_{rnd}^k and ω_{top}^k :

$$\omega_{rnd}^k(x) = \left(1 - \frac{k}{d}\right) \sum_{i=1}^d x_i^2, \quad \omega_{top}^k(x) = \sum_{i=1}^{d-k} x_i^2.$$

The expected variance $\mathbb{E}[\omega_{rnd}^k]$ for Rand- k is easy to compute as all coordinates are independent and uniformly distributed on $[0, 1]$:

$$\mathbb{E}[x_i^2] \equiv \int_{[0,1]^d} x_i^2 dx = \int_0^1 x_i^2 dx_i = \frac{1}{3}, \quad (38)$$

which implies

$$\mathbb{E} \left[\omega_{rnd}^k(x) \right] = \left(1 - \frac{k}{d} \right) \sum_{i=1}^d \mathbb{E} \left[x_i^2 \right] = \left(1 - \frac{k}{d} \right) \frac{d}{3} = \frac{d-k}{3}. \quad (39)$$

In order to compute the expected variance $\mathbb{E} \left[\omega_{top}^k \right]$ for Top- k , we use the following formula from order statistics (see (31), (32) or (2.2.2), (2.2.3) of Arnold et al. (1992)) [§]

$$\mathbb{E} \left[x_{(i)}^2 \right] \equiv \int_{[0,1]^d} x_{(i)}^2 dx = \frac{\Gamma(i+2)\Gamma(d+1)}{\Gamma(i)\Gamma(d+3)} = \frac{i(i+1)}{(d+1)(d+2)}, \quad (40)$$

from which we derive

$$\begin{aligned} \mathbb{E} \left[\omega_{top}^k \right] &= \sum_{i=1}^{d-k} \mathbb{E} \left[x_{(i)}^2 \right] = \frac{1}{(d+1)(d+2)} \sum_{i=1}^{d-k} i(i+1) \\ &= \frac{1}{(d+1)(d+2)} \cdot \frac{(d-k)(d-k+1)(d-k+2)}{3} \\ &= \frac{d-k}{3} \left(1 - \frac{k}{d+1} \right) \left(1 - \frac{k}{d+2} \right). \end{aligned} \quad (41)$$

Combining (39) and (41) completes the first relation. Thus, on average (w.r.t. uniform distribution) Top- k has roughly $(1 - k/d)^2$ times less variance than Rand- k .

For the second relation, we use (38) and (40) for $i = d$ and get

$$\frac{\mathbb{E} \left[s_{top}^1(x) \right]}{\mathbb{E} \left[s_{rnd}^1(x) \right]} = \frac{\mathbb{E} \left[x_{(d)}^2 \right]}{\mathbb{E} \left[x_d^2 \right]} = \frac{\frac{d(d+1)}{(d+1)(d+2)}}{\frac{1}{3}} = \frac{3d}{d+2}.$$

Clearly, one can extend this for any $k \in [d]$.

(b) Recall that for the standard exponential distribution (with $\lambda = 1$) probability density function (PDF) is given as follows:

$$\phi(t) = e^{-t}, \quad t \in [0, \infty).$$

Both mean and variance can be shown to be equal to 1. The expected saving $\mathbb{E} \left[s_{rnd}^1 \right]$ can be computed directly:

$$\mathbb{E} \left[s_{rnd}^1(x) \right] = \mathbb{E} \left[x_d^2 \right] = \text{Var} \left[x_d \right] + \mathbb{E} \left[x_d \right]^2 = 2.$$

To compute the expected saving $\mathbb{E} \left[s_{top}^1(x) \right] = \mathbb{E} \left[x_{(d)}^2 \right]$ we prove the following lemma:

Lemma 25 *Let x_1, x_2, \dots, x_d be an i.i.d. sample from the standard exponential distribution and*

$$y_i := (d - i + 1)(x_{(i)} - x_{(i-1)}), \quad 1 \leq i \leq d,$$

where $x_{(0)} := 0$. Then y_1, y_2, \dots, y_d is an i.i.d. sample from the standard exponential distribution.

[§]see also https://en.wikipedia.org/wiki/Order_statistic, <https://www.sciencedirect.com/science/article/pii/S0167715212001940>

Proof The joint density function of $x_{(1)}, \dots, x_{(d)}$ is given by (see (32))

$$\phi_{x_{(1)}, \dots, x_{(d)}}(u_1, \dots, u_d) = d! \prod_{i=1}^d \phi(u_i) = d! \exp\left(-\sum_{i=1}^d u_i\right), \quad 0 \leq u_1 \leq \dots \leq u_d < \infty.$$

Next we express variables $x_{(i)}$ using new variables y_i

$$x_{(1)} = \frac{y_1}{d}, \quad x_{(2)} = \frac{y_1}{d} + \frac{y_2}{d-1}, \quad \dots, \quad x_{(d)} = \frac{y_1}{d} + \frac{y_2}{d-1} + \dots + y_d,$$

with the transformation matrix

$$A = \begin{pmatrix} \frac{1}{d} & 0 & \dots & 0 \\ \frac{1}{d} & \frac{1}{d-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{d} & \frac{1}{d-1} & \dots & 1 \end{pmatrix}$$

Then the joint density $\psi_{y_1, \dots, y_d}(u) = \psi_{y_1, \dots, y_d}(u_1, \dots, u_d)$ of new variables y_1, \dots, y_d is given as follows

$$\psi_{y_1, \dots, y_d}(u) = \frac{\phi_{x_{(1)}, \dots, x_{(d)}}(Au)}{|\det A^{-1}|} = |\det A| \cdot \phi_{x_{(1)}, \dots, x_{(d)}}(Au)$$

Notice that $\sum_{i=1}^d u_i = \sum_{i=1}^d (Au)_i$ and $|\det A| = 1/d!$. Hence

$$\psi_{y_1, \dots, y_d}(u) = \exp\left(-\sum_{i=1}^d u_i\right), \quad 0 \leq u_1 \leq \dots \leq u_d \leq \infty,$$

which means that variables y_1, \dots, y_d are independent and have standard exponential distribution. ■

Using this lemma we can compute the mean and the second moment of $x_{(d)} = \sum_{i=1}^d \frac{y_i}{d-i+1}$ as follows

$$\begin{aligned} \mathbb{E}[x_{(d)}] &= \sum_{i=1}^d \mathbb{E}\left[\frac{y_i}{d-i+1}\right] = \sum_{i=1}^d \frac{\mathbb{E}[y_i]}{d-i+1} = \sum_{i=1}^d \frac{1}{i}, \\ \text{Var}[x_{(d)}] &= \sum_{i=1}^d \text{Var}\left[\frac{y_i}{d-i+1}\right] = \sum_{i=1}^d \frac{\text{Var}[y_i]}{(d-i+1)^2} = \sum_{i=1}^d \frac{1}{i^2}, \end{aligned}$$

from which we conclude the lemma as

$$\mathbb{E}[s_{top}^1(x)] = \mathbb{E}[x_{(d)}^2] = \text{Var}[x_{(d)}] + \mathbb{E}[x_{(d)}]^2 = \sum_{i=1}^d \frac{1}{i^2} + \left(\sum_{i=1}^d \frac{1}{i}\right)^2 \approx \mathcal{O}(\log^2 d). \quad \blacksquare$$

D.1 Proof of Theorem 21 (Convergence guarantees for Algorithm 1)

In this section, we include our analysis for the Distributed SGD with biased compression. Our analysis is closely related to the analysis of Stich and Karimireddy (2019).

We start with the definition of some auxiliary objects:

Definition 26 *The sequence $\{a^k\}_{k \geq 0}$ of positive values is τ -slow decreasing for parameter τ :*

$$a^{k+1} \leq a^k, \quad a^{k+1} \left(1 + \frac{1}{2\tau}\right) \geq a^k, \quad \forall k \geq 0 \quad (42)$$

The sequence $\{a^k\}_{k \geq 0}$ of positive values is τ -slow increasing for parameter τ :

$$a^{k+1} \geq a^k, \quad a^{k+1} \leq a^k \left(1 + \frac{1}{2\tau}\right), \quad \forall k \geq 0 \quad (43)$$

And let:

$$\tilde{x}^k = x^k - \frac{1}{n} \sum_{i=1}^n e_i^k, \quad \forall k \geq 0 \quad (44)$$

$$g^k = \frac{1}{n} \sum_{i=1}^n g_i^k \quad (45)$$

It is easy to see:

$$\begin{aligned} \tilde{x}^{k+1} &= x^{k+1} - \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \\ &\stackrel{(22),(23)}{=} \left(x^k - \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^k \right) - \left(\frac{1}{n} \sum_{i=1}^n [e_i^k + \eta^k g_i^k - \tilde{g}_i^k] \right) \\ &= \tilde{x}^k - \frac{\eta^k}{n} \sum_{i=1}^n g_i^k \end{aligned} \quad (46)$$

Lemma 27 *If $\eta^k \leq \frac{1}{4L(1+2B/n)}$, $\forall k \geq 0$, then for $\{\tilde{x}^k\}_{k \geq 0}$ defined as in (44),*

$$\begin{aligned} \mathbb{E} \left[\left\| \tilde{x}^{k+1} - x^* \right\|_2^2 \right] &\leq \left(1 - \frac{\mu\eta^k}{2} \right) \mathbb{E} \left[\left\| \tilde{x}^k - x^* \right\|_2^2 \right] - \frac{\eta^k}{2} \mathbb{E} \left[f(x^k) - f^* \right] \\ &\quad + 3L\eta^k \mathbb{E} \left[\left\| x^k - \tilde{x}^k \right\|_2^2 \right] + (\eta^k)^2 \frac{C + 2BD}{n} \end{aligned} \quad (47)$$

Proof We consider the following equalities, using the relationship between \tilde{x}_{k+1} and \tilde{x}_k :

$$\begin{aligned} \left\| \tilde{x}^{k+1} - x^* \right\|_2^2 &\stackrel{(45),(46)}{=} \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle g^k, \tilde{x}^k - x^* \rangle + (\eta^k)^2 \left\| g^k \right\|_2^2 \\ &= \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle g^k, x^k - x^* \rangle + (\eta^k)^2 \left\| g^k \right\|_2^2 + 2\eta^k \langle g^k, x^k - \tilde{x}^k \rangle. \end{aligned}$$

Taking the conditional expectation conditioned on previous iterates, we get

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \tilde{x}^{k+1} - x^* \right\|_2^2 \right] \\
 = & \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \mathbb{E} [g^k], x^k - x^* \rangle + (\eta^k)^2 \cdot \mathbb{E} \left[\left\| g^k \right\|_2^2 \right] + 2\eta^k \langle \mathbb{E} [g^k], x^k - \tilde{x}^k \rangle \\
 \stackrel{(19),(45)}{=} & \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \mathbb{E} [g^k], x^k - x^* \rangle \\
 & + (\eta^k)^2 \cdot \mathbb{E} \left[\left\| \nabla f(x^k) + \frac{1}{n} \sum_{i=1}^n \xi_i^k \right\|_2^2 \right] + 2\eta^k \langle \mathbb{E} [g^k], x^k - \tilde{x}^k \rangle \\
 = & \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \mathbb{E} [g^k], x^k - x^* \rangle \\
 & + (\eta^k)^2 \cdot \mathbb{E} \left[\left\| \nabla f(x^k) \right\|_2^2 + 2\langle \nabla f(x^k), \frac{1}{n} \sum_{i=1}^n \xi_i^k \rangle + \left\| \frac{1}{n} \sum_{i=1}^n \xi_i^k \right\|_2^2 \right] + 2\eta^k \langle \mathbb{E} [g^k], x^k - \tilde{x}^k \rangle.
 \end{aligned}$$

Given the unbiased stochastic gradient ($\mathbb{E} [\xi_i^k] = 0$):

$$\begin{aligned}
 \mathbb{E} \left[\left\| \tilde{x}^{k+1} - x^* \right\|_2^2 \right] &= \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
 &\quad + (\eta^k)^2 \left\| \nabla f(x^k) \right\|_2^2 + (\eta^k)^2 \cdot \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \xi_i^k \right\|_2^2 \right] + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle
 \end{aligned}$$

Using that ξ_i^k mutually independent and $\mathbb{E} [\xi_i^k] = 0$ we have:

$$\begin{aligned}
 & \stackrel{(30)}{\leq} \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
 & \quad + (\eta^k)^2 \cdot \left\| \nabla f(x^k) \right\|_2^2 + (\eta^k)^2 \cdot \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[\left\| \xi_i^k \right\|_2^2 \right] + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \\
 & \stackrel{(24)}{\leq} \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
 & \quad + (\eta^k)^2 \cdot \left\| \nabla f(x^k) \right\|_2^2 + \frac{(\eta^k)^2}{n^2} \sum_{i=1}^n \left[B \left\| \nabla f_i(x^k) \right\|_2^2 \right] + \frac{(\eta^k)^2}{n} C \\
 & \quad + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \\
 & \stackrel{(27)}{\leq} \left\| \tilde{x}^k - x^* \right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
 & \quad + (\eta^k)^2 \cdot 2L(f(x^k) - f(x^*)) + \frac{(\eta^k)^2}{n^2} \sum_{i=1}^n \left[B \left\| \nabla f_i(x^k) \right\|_2^2 \right] + \frac{(\eta^k)^2}{n} C \\
 & \quad + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle. \tag{48}
 \end{aligned}$$

All f_i are L -smooth and μ -strongly convex, thus f is L -smooth and μ -strongly convex. We can rewrite $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k)\|_2^2$:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k)\|_2^2 &= \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*) + \nabla f_i(x^*)\|_2^2 \\ &\stackrel{(30)}{\leq} \frac{2}{n} \sum_{i=1}^n \left(\|\nabla f_i(x^k) - \nabla f_i(x^*)\|_2^2 + \|\nabla f_i(x^*)\|_2^2 \right) \\ &\stackrel{(26)}{\leq} \frac{2}{n} \sum_{i=1}^n \left[2L \left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^* \rangle \right) + \|\nabla f_i(x^*)\|_2^2 \right]. \end{aligned}$$

Using definition of $D = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|_2^2$:

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k)\|_2^2 \leq 4L \left(f(x^k) - \nabla f(x^*) \right) + 2D \quad (49)$$

Substituting (49) to (48):

$$\begin{aligned} \mathbb{E} \left[\|\tilde{x}^{k+1} - x^*\|_2^2 \right] &= \|\tilde{x}^k - x^*\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle + (\eta^k)^2 \cdot 2L \left(1 + \frac{2B}{n} \right) (f(x^k) - f(x^*)) \\ &\quad + (\eta^k)^2 \frac{C + 2BD}{n} + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \end{aligned} \quad (50)$$

By (25) we have for f :

$$-2\langle \nabla f(x^k), x^k - x^* \rangle \leq -\mu \|x^k - x^*\|_2^2 - 2(f(x^k) - f^*). \quad (51)$$

Using (28) with $\xi = 1/2L$ and L -smoothness of f (27):

$$2\langle \nabla f(x^k), \tilde{x}^k - x^k \rangle \leq \frac{1}{2L} \|\nabla f(x^k)\|_2^2 + 2L \|x^k - \tilde{x}^k\|_2^2 \leq f(x^k) - f^* + 2L \|x^k - \tilde{x}^k\|_2^2. \quad (52)$$

By (30) for $\|\tilde{x}^k - x^*\|_2^2$, we get:

$$-\|x^k - x^*\|_2^2 \leq -\frac{1}{2} \|\tilde{x}^k - x^*\|_2^2 + \|x^k - \tilde{x}^k\|_2^2. \quad (53)$$

Plugging (51), (52), (53) into (50):

$$\begin{aligned} \|\tilde{x}^{k+1} - x^*\|_2^2 &\leq \left(1 - \frac{\mu\eta^k}{2} \right) \|\tilde{x}^k - x^*\|_2^2 - \eta^k \left[1 - \eta^k \cdot 2L \left(1 + \frac{2B}{n} \right) \right] (f(x^k) - f^*) \\ &\quad + \eta^k (2L + \mu) \|x^k - \tilde{x}^k\|_2^2 + (\eta^k)^2 \frac{C + 2BD}{n} \end{aligned}$$

The lemma follows by the choice $\eta^k \leq \frac{1}{4L(1+2B/n)}$ and $L \geq \mu$. ■

Lemma 28 $\eta^k \leq \frac{1}{14(2\delta+B)L}$, $\forall k \geq 0$ and $\{(\eta^k)^2\}_{k \geq 0}$ -2δ -slow decreasing. Then

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] &\leq \frac{(1-1/\delta)}{49L(2\delta+B)} \sum_{j=0}^k \left[\left(1 - \frac{1}{4\delta}\right)^{k-j} (f(x^j) - f(x^*)) \right] \\ &\quad + \eta^k \frac{2(\delta-1)}{7L} \left(2D + \frac{C}{2\delta+B} \right). \end{aligned} \quad (54)$$

Furthermore, for any 4δ -slow increasing non-negative sequence $\{w^k\}_{k \geq 0}$ it holds:

$$3L \cdot \sum_{k=0}^K w^k \cdot \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^k \right\|_2^2 \right] \leq \frac{1}{4} \sum_{k=0}^K w^k (\mathbb{E} [f(x^k)] - f(x_*)) + \left(3\delta D + \frac{3C}{4} \right) \sum_{k=0}^K w^k \eta^k. \quad (55)$$

Proof

We prove the first part of the statement:

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] &\stackrel{(30)}{\leq} \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \left\| e_i^{k+1} \right\|_2^2 \right] \\ &\stackrel{(22)}{=} \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \left\| e_i^k + \eta^k g_i^k - \tilde{g}_i^k \right\|_2^2 \right] \\ &\stackrel{(21)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| e_i^k + \eta^k g_i^k - \mathcal{C}(e_i^k + \eta^k g_i^k) \right\|_2^2 \right] \\ &\stackrel{(7)}{\leq} \frac{1-1/\delta}{n} \sum_{i=1}^n \mathbb{E}_\nabla \left[\left\| e_i^k + \eta^k g_i^k \right\|_2^2 \right] \\ &\stackrel{(19)}{=} \frac{1-1/\delta}{n} \sum_{i=1}^n \mathbb{E}_\nabla \left[\left\| e_i^k + \eta^k \nabla f_i(x^k) + \eta^k \xi_i^k \right\|_2^2 \right] \end{aligned}$$

Here we have taken into account that the operator of full expectation is a combination of operators of expectation by the randomness of the operator and the randomness of the stochastic gradient, i.e. $\mathbb{E}[\cdot] = \mathbb{E}_C[\mathbb{E}_\nabla[\cdot]]$. Given the unbiased stochastic gradient ($\mathbb{E}[\xi_i^k] = 0$):

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] &\leq \frac{1-1/\delta}{n} \sum_{i=1}^n \left[\left\| e_i^k + \eta^k \nabla f_i(x^k) \right\|_2^2 + \mathbb{E}_\nabla \left[\left\| \eta^k \xi_i^k \right\|_2^2 \right] \right] \\ &\stackrel{(24)}{\leq} \frac{1-1/\delta}{n} \sum_{i=1}^n \left[\left\| e_i^k + \eta^k \nabla f_i(x^k) \right\|_2^2 + (\eta^k)^2 \left(B \left\| \nabla f_i(x^k) \right\|_2^2 + C \right) \right] \end{aligned}$$

Using (29) with some ξ :

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] \\
 & \leq \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \|e_i^{k+1}\|_2^2 \right] \\
 & \leq \frac{1 - \frac{1}{\delta}}{n} \sum_{i=1}^n \left[(1 + \xi) \|e_i^k\|_2^2 + (\eta^k)^2 \left(1 + \frac{1}{\xi}\right) \|\nabla f_i(x^k)\|_2^2 + (\eta^k)^2 B \|\nabla f_i(x^k)\|_2^2 + (\eta^k)^2 C \right] \\
 & = \left(1 - \frac{1}{\delta}\right) \left[(1 + \xi) \left(\frac{1}{n} \sum_{i=1}^n \|e_i^k\|_2^2\right) + (\eta^k)^2 \left(1 + \frac{1}{\xi} + B\right) \left(\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k)\|_2^2\right) + (\eta^k)^2 C \right] \\
 & \stackrel{(49)}{\leq} \left(1 - \frac{1}{\delta}\right) \left[(1 + \xi) \left(\frac{1}{n} \sum_{i=1}^n \|e_i^k\|_2^2\right) \right] \\
 & \quad + \left(1 - \frac{1}{\delta}\right) \left[(\eta^k)^2 \left(1 + \frac{1}{\xi} + B\right) (4L(f(x^k) - f(x^*)) + 2D) + (\eta^k)^2 C \right]
 \end{aligned}$$

Using the recurrence for $\frac{1}{n} \sum_{i=1}^n \|e_i^k\|_2^2$, and let $\xi = \frac{1}{2(\delta-1)}$, then $(1 + 1/\xi) \leq 2\delta$, and $(1 - 1/\delta)(1 + \xi) = (1 - 1/2\delta)$ we have

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] \\
 & \leq \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \|e_i^{k+1}\|_2^2 \right] \\
 & \leq \left(1 - \frac{1}{\delta}\right) \sum_{j=0}^k (\eta^j)^2 \left[\left(1 - \frac{1}{\delta}\right) (1 + \xi) \right]^{k-j} \left(1 + \frac{1}{\xi} + B\right) (4L(\mathbb{E}[f(x^j)] - f(x^*)) + 2D) \\
 & \quad + \left(1 - \frac{1}{\delta}\right) \sum_{j=0}^k (\eta^j)^2 \left[\left(1 - \frac{1}{\delta}\right) (1 + \xi) \right]^{k-j} C \\
 & \leq \left(1 - \frac{1}{\delta}\right) \sum_{j=0}^k (\eta^j)^2 \left(1 - \frac{1}{2\delta}\right)^{k-j} ((2\delta + B) (4L(\mathbb{E}[f(x^j)] - f(x^*)) + 2D) + C) .
 \end{aligned}$$

For 2δ -slow decreasing $\{(\eta^k)^2\}_{k \geq 0}$ by definition (42) we get that $(\eta^j)^2 \leq (\eta^k)^2 (1 + \frac{1}{4\delta})^{k-j}$. Due to the fact that $(1 - 1/2\delta)(1 + 1/4\delta) \leq (1 - 1/4\delta)$, we have:

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] \\
 & \leq \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \left\| e_i^{k+1} \right\|_2^2 \right] \\
 & \leq \left(1 - \frac{1}{\delta}\right) \sum_{j=0}^k (\eta^j)^2 \left(1 + \frac{1}{4\delta}\right)^{k-j} \left(1 - \frac{1}{2\delta}\right)^{k-j} (2\delta + B) (4L(\mathbb{E}[f(x^j)] - f(x^*)) + 2D) \\
 & \quad + \left(1 - \frac{1}{\delta}\right) \sum_{j=0}^k (\eta^j)^2 \left(1 + \frac{1}{4\delta}\right)^{k-j} \left(1 - \frac{1}{2\delta}\right)^{k-j} C \\
 & \leq (\eta^k)^2 \left(1 - \frac{1}{\delta}\right) (2\delta + B) \sum_{j=0}^k \left[\left(1 - \frac{1}{4\delta}\right)^{k-j} 4L(\mathbb{E}[f(x^j)] - f(x^*)) \right] \\
 & + (\eta^k)^2 \left(1 - \frac{1}{\delta}\right) 4\delta[C + 2D(2\delta + B)].
 \end{aligned}$$

As the last step, we use formula for geometric progression in the following way:

$$\sum_{j=0}^k \left(1 - \frac{1}{4\delta}\right)^{k-j} = \sum_{j=0}^k \left(1 - \frac{1}{4\delta}\right)^j \leq \sum_{j=0}^{\infty} \left(1 - \frac{1}{4\delta}\right)^j = 4\delta$$

By observing that the choice of the stepsize $\eta^k \leq \frac{1}{14(2\delta+B)L}$:

$$\begin{aligned}
 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^{k+1} \right\|_2^2 \right] & \leq \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \left\| e_i^{k+1} \right\|_2^2 \right] \\
 & \leq \frac{(1 - 1/\delta)}{49L(2\delta + B)} \sum_{j=0}^k \left[\left(1 - \frac{1}{4\delta}\right)^{k-j} (\mathbb{E}[f(x^j)] - f(x^*)) \right] \\
 & \quad + \eta^k \frac{2(\delta - 1)}{7L} \left(2D + \frac{C}{2\delta + B} \right),
 \end{aligned}$$

which concludes the proof of (54). For the second part, we use the previous results. Summing over all k :

$$\begin{aligned}
 \sum_{k=0}^K w^k \cdot \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^k \right\|_2^2 \right] & \stackrel{(54)}{\leq} \frac{(1 - 1/\delta)}{49L(2\delta + B)} \sum_{k=0}^K w^k \sum_{j=0}^{k-1} \left(1 - \frac{1}{4\delta}\right)^{k-j-1} (\mathbb{E}[f(x^j)] - f(x^*)) \\
 & \quad + \frac{2(\delta - 1)}{7L} \left(2D + \frac{C}{2\delta + B} \right) \sum_{k=0}^K w^k \eta^{k-1}
 \end{aligned}$$

For 2δ -slow decreasing $\{(\eta^k)^2\}_{k \geq 0}$, it holds $(\eta^{k-1})^2 \leq (\eta^k)^2(1 + \frac{1}{4\delta})$ which follows from (42) and $\eta^{k-1} \leq \eta^k(1 + \frac{1}{4\delta})$ and for 4δ -slow increasing $\{w^k\}_{k \geq 0}$ by (43) we have $w^k \leq w^{k-j}(1 + \frac{1}{8\delta})^j$. Then

$$\begin{aligned}
 \sum_{k=0}^K w^k \cdot \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^k \right\|_2^2 \right] &\stackrel{(54)}{\leq} \frac{(1-1/\delta)}{49L(2\delta+B)} \sum_{k=0}^K w^k \sum_{j=0}^{k-1} \left(1 - \frac{1}{4\delta}\right)^{k-j-1} (\mathbb{E}[f(x^j)] - f(x^*)) \\
 &\quad + \frac{2(\delta-1)}{7L} \left(2D + \frac{C}{2\delta+B}\right) \left(1 + \frac{1}{4\delta}\right) \sum_{k=0}^K w^k \eta^k \\
 &\leq \frac{(1-1/\delta)}{49L(2\delta+B)} \sum_{k=0}^K \sum_{j=0}^{k-1} w^j \left(1 + \frac{1}{8\delta}\right)^{k-j} \left(1 - \frac{1}{4\delta}\right)^{k-j} (\mathbb{E}[f(x^j)] - f(x^*)) \\
 &\quad + \frac{\delta-1}{2L} \left(2D + \frac{C}{2\delta+B}\right) \sum_{k=0}^K w^k \eta^k \\
 &\leq \frac{(1-1/\delta)}{49L(2\delta+B)} \sum_{k=0}^K \sum_{j=0}^{k-1} w_j \left(1 - \frac{1}{8\delta}\right)^{k-j} (\mathbb{E}[f(x^j)] - f(x^*)) \\
 &\quad + \frac{\delta-1}{2L} \left(2D + \frac{C}{2\delta+B}\right) \sum_{k=0}^K w^k \eta^k \\
 &\leq \frac{(1-1/\delta)}{49L(2\delta+B)} \sum_{k=0}^K w^k (\mathbb{E}[f(x^k)] - f(x^*)) \sum_{j=0}^{\infty} \left(1 - \frac{1}{8\delta}\right)^j \\
 &\quad + \frac{\delta-1}{2L} \left(2D + \frac{C}{2\delta+B}\right) \sum_{k=0}^K w^k \eta^k.
 \end{aligned}$$

Observing $\sum_{j=0}^{\infty} (1 - 1/8\delta)^j \leq 8\delta$ and using $\delta^{-1}/2\delta+B \leq 1/2$ concludes the proof. \blacksquare

Lemma 29 (Lemma 11, Stich and Karimireddy (2019)) *For decreasing stepsizes $\{\eta^k := \frac{2}{a(\kappa+k)}\}_{k \geq 0}$, and weights $\{w_k := (\kappa+k)\}_{k \geq 0}$ for parameters $\kappa \geq 1$, it holds for every non-negative sequence $\{r^k\}_{k \geq 0}$ and any $a > 0$, $c \geq 0$ that*

$$\Psi^K := \frac{1}{WK} \sum_{k=0}^K \left(\frac{w^k}{\eta^k} (1 - a\eta^k) r^k - \frac{w^k}{\eta^k} r^{k+1} + c\eta^k w^k \right) \leq \frac{a\kappa^2 r^0}{K^2} + \frac{4c}{aK},$$

where $W^K := \sum_{k=0}^K w^k$.

Proof We start by observing that

$$\frac{w^k}{\eta^k} (1 - a\eta^k) r^k = \frac{a}{2} (\kappa+k)(\kappa+k-2) r^k = \frac{a}{2} ((\kappa+k-1)^2 - 1) \leq \frac{a}{2} (\kappa+k-1)^2. \quad (56)$$

By plugging in the definitions of η^k and w^k in Ψ^K , we end up with the following telescoping sum:

$$\Psi^K \stackrel{(56)}{\leq} \frac{1}{WK} \sum_{k=0}^K \left(\frac{a}{2} (\kappa + k - 1)^2 r^k - \frac{a}{2} (\kappa + k)^2 r^{k+1} \right) + \sum_{k=0}^K \frac{2c}{aWK} \leq \frac{a(\kappa - 1)^2 r^0}{2WK} + \frac{2c(K+1)}{aWK}.$$

The lemma now follows from $(\kappa - 1)^2 \leq \kappa^2$ and $WK = \sum_{k=0}^K (\kappa + k) = \frac{(2\kappa + K)(K+1)}{2} \geq \frac{K(K+1)}{2} \geq \frac{K^2}{2}$. \blacksquare

Lemma 30 (Lemma 12, Stich and Karimireddy (2019)) *For every non-negative sequence $\{r^k\}_{k \geq 0}$ and any parameters $d \geq a > 0$, $c \geq 0$, $K \geq 0$, there exists a constant $\eta \leq \frac{1}{d}$, such that for constant stepsizes $\{\eta^k = \eta\}_{k \geq 0}$ and weights $w^k := (1 - a\eta)^{-(k+1)}$ it holds*

$$\Psi^K := \frac{1}{WK} \sum_{k=0}^K \left(\frac{w^k}{\eta^k} \left(1 - a\eta^k \right) r^k - \frac{w^k}{\eta^k} r^{k+1} + c\eta^k w^k \right) = \tilde{\mathcal{O}} \left(dr_0 \exp \left[-\frac{aK}{d} \right] + \frac{c}{aK} \right).$$

Proof By plugging in the values for η^k and w^k , we observe that we again end up with a telescoping sum and estimate

$$\Psi^K = \frac{1}{\eta WK} \sum_{k=0}^K \left(w^{k-1} r^k - w^k r^{k+1} \right) + \frac{c\eta}{WK} \sum_{k=0}^K w^k \leq \frac{r^0}{\eta WK} + c\eta \leq \frac{r^0}{\eta} \exp[-a\eta K] + c\eta,$$

where we used the estimate $WK \geq w^K \geq (1 - a\eta)^{-K} \geq \exp[a\eta K]$ for the last inequality. The lemma now follows by carefully tuning η . \blacksquare

Lemma 31 (Lemma 13, Stich and Karimireddy (2019)) *For every non-negative sequence $\{r^k\}_{k \geq 0}$ and any parameters $d \geq 0$, $c \geq 0$, $K \geq 0$, there exists a constant $\eta \leq \frac{1}{d}$, such that for constant stepsizes $\{\eta^k = \eta\}_{k \geq 0}$ it holds:*

$$\Psi^K := \frac{1}{K+1} \sum_{k=0}^K \left(\frac{(1 - a\eta^k) r^k}{\eta^k} - \frac{r^{k+1}}{\eta^k} + c\eta^k \right) \leq \frac{(d - a)r^0}{K+1} + \frac{2\sqrt{cr^0}}{\sqrt{K+1}}$$

Proof For constant stepsizes $\eta^t = \eta$ we can derive the estimate

$$\Psi^K = \frac{1}{\eta(K+1)} \sum_{k=0}^K \left((1 - a\eta) r^k - r^{k+1} \right) + c\eta \leq \frac{(1 - a\eta)r^0}{\eta(K+1)} + c\eta.$$

We distinguish two cases: if $\frac{r^0}{c(K+1)} \leq \frac{1}{d^2}$, then we chose the stepsize $\eta = \sqrt{\frac{r^0}{c(K+1)}}$ and get

$$\Psi^K \leq \frac{\sqrt{r^0}}{(K+1)} (2 \cdot \sqrt{c(K+1)} - a\sqrt{r^0}),$$

on the other hand, if $\frac{r^0}{c(K+1)} > \frac{1}{d^2}$, then we choose $\eta = \frac{1}{d}$ and get

$$\Psi^K \leq \frac{r^0(d-a)}{K+1} + \frac{c}{d} \leq \frac{r^0(d-a)}{K+1} + \frac{\sqrt{cr^0}}{\sqrt{K+1}},$$

which concludes the proof. \blacksquare

The proof of the main theorem follows

Proof [Proof of the Theorem 21] It is easy to see that $1/14(2\delta+B)L \leq 1/4L(1+2B/n)$. This means that the Lemma 27 is satisfied. With the notation $r^k := \mathbb{E} \left[\|\tilde{x}^{k+1} - x^*\|_2^2 \right]$ and $s^k := \mathbb{E} [f(x^k)] - f^*$ we have for any $w^k > 0$:

$$\frac{w^k}{2} s^k \stackrel{(47)}{\leq} \frac{w^k}{\eta^k} \left(1 - \frac{\mu\eta^k}{2} \right) r^k - \frac{w^k}{\eta^k} r^{k+1} + \eta^k w^k \frac{C + 2BD}{n} + 3w^k L \cdot \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n e_i^k \right\|_2^2 \right].$$

Substituting (55) and summing over k we have:

$$\frac{1}{2} \sum_{k=0}^K w^k s^k \leq \sum_{k=0}^K \left(\frac{w^k}{\eta^k} \left(1 - \frac{\mu\eta^k}{2} \right) r^k - \frac{w^k}{\eta^k} r^{k+1} + \eta^k w^k \tilde{C} \right) + \frac{1}{4} \sum_{k=0}^K w^k s^k.$$

where $\tilde{C} = C \left(1 + \frac{1}{n} \right) + D \left(\frac{2B}{n} + 3\delta \right)$.

This can be rewritten as

$$\frac{1}{WK} \sum_{k=0}^K w^k s^k \leq \frac{4}{WK} \sum_{k=0}^K \left(\frac{w^k}{\eta^k} \left(1 - \frac{\mu\eta^k}{2} \right) r^k - \frac{w^k}{\eta^k} r^{k+1} + \eta^k w^k \tilde{C} \right).$$

First, when the stepsizes $\eta^k = \frac{4}{\mu(\kappa+k)}$, it is easy to see that $\eta^k \leq \frac{1}{14(2\delta+B)L}$:

$$\eta^k \leq \eta^0 = \frac{4}{\mu\kappa} \leq \frac{4}{\mu} \cdot \frac{\mu}{56(2\delta+B)L} = \frac{1}{14(2\delta+B)L}$$

Not difficult to check that $\{(\eta^k)^2\}_{k \geq 0}$ is 2δ slow decreasing:

$$\frac{(\eta^{k+1})^2}{(\eta^k)^2} = \left(\frac{\kappa+k+1}{\kappa+k} \right)^2 \leq \left(1 + \frac{1}{\kappa+k} \right)^2 \leq \left(1 + \frac{1}{\kappa} \right)^2 = \left(1 + \frac{\mu}{56(2\delta+B)L} \right)^2 \leq 1 + \frac{1}{4\delta}$$

Furthermore, the weights $\{w^k = \kappa+k\}_{k \geq 0}$ are 4δ -slow increasing:

$$\frac{w^{k+1}}{w^k} = \frac{\kappa+k+1}{\kappa+k} = 1 + \frac{1}{\kappa+k} \leq 1 + \frac{1}{\kappa} = 1 + \frac{\mu}{56(2\delta+B)L} \leq 1 + \frac{1}{8\delta}.$$

The conditions for Lemma 29 are satisfied, and we obtain the desired statement. For the second case, the conditions of Lemma 30 are easy to check (see the previous paragraph). The claim follows by this lemma. Finally, for the third claim, we invoke Lemma 31. \blacksquare

References

- Saurabh Agarwal, Hongyi Wang, Kangwook Lee, Shivaram Venkataraman, and Dimitris Papailiopoulos. Accordion: Adaptive gradient communication via critical learning regime identification. *arXiv preprint arXiv:2010.16248*, 2020.
- Ahmad Ajalloeian and Sebastian U. Stich. On the Convergence of SGD with Biased Gradients. *arXiv preprint arXiv:2008.00051*, 2021.
- Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.
- Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5977–5987, 2018.
- Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.
- Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in order Statistics*. John Wiley and Sons Inc., 1992.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- Tom B. Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Jean-Baptiste Cordonnier. Convex optimization using sparsified stochastic gradient descent with memory. Technical report, École Polytechnique Fédérale de Lausanne, 2018.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback. *arXiv preprint arXiv:2110.03294*, 2021.
- Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020a.
- Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. In *34th Conference on Neural Information Processing Systems*, 2020b.

- Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=vYVI1CHPaQg>.
- Samuel Horváth, Chen-Yu Ho, Ľudovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019a.
- Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019b.
- Peter Kairouz and et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.
- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *ArXiv*, abs/1910.06378, 2019a.
- Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019b.
- Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020a.
- Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *arXiv preprint arXiv:2006.11573*, 2020b.
- Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.
- Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.
- Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.
- Hyeontaek Lim, David G Andersen, and Michael Kaminsky. 3lc: Lightweight and effective traffic compression for distributed machine learning. *arXiv preprint arXiv:1802.07389*, 2018.

- Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *CoRR*, abs/1712.01887, 2017a. URL <http://arxiv.org/abs/1712.01887>.
- Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017b.
- Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *ICLR 2018 - International Conference on Learning Representations*, 2018.
- H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.
- Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- 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. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.
- Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In *35nd Conference on Neural Information Processing Systems*, 2021.
- Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
- Amedeo Sapio, Marco Canini, Chen-Yu Ho, Jacob Nelson, Panos Kalnis, Changhoon Kim, Arvind Krishnamurthy, Masoud Moshref, Dan R. K. Ports, and Peter Richtárik. Scaling distributed machine learning with in-network aggregation. *arXiv preprint arXiv:1903.06701*, 2019.
- Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and application to data-parallel distributed training of speech dnns. In *Interspeech 2014*, September 2014.

- Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.
- 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 31*, pages 4447–4458. Curran Associates, Inc., 2018. URL <http://papers.nips.cc/paper/7697-sparsified-sgd-with-memory.pdf>.
- Haobo Sun, Yingxia Shao, Jiawei Jiang, Bin Cui, Kai Lei, Yu Xu, and Jiang Wang. Sparse gradient compression for distributed SGD. In Guoliang Li, Jun Yang, Joao Gama, Jugapong Natwichai, and Yongxin Tong, editors, *Database Systems for Advanced Applications*, pages 139–155, Cham, 2019. Springer International Publishing. ISBN 978-3-030-18579-4.
- Sharan Vaswani, Francis Bach, and Mark Schmidt. 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*, pages 1195–1204, 2019.
- Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- 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.
- Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5325–5333, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4035–4043, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- Shen-Yi Zhao, Yinpeng Xie, Hao Gao, and Wu-Jun Li. Global momentum compression for sparse communication in distributed SGD. *arXiv preprint arXiv:1905.12948*, 2019.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations

by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.