On Biased Compression for Distributed Learning

Aleksandr Beznosikov **Samuel Horváth** **Peter Richtárik** **Mher Safaryan**

King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia

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. Our *distributed* SGD method enjoys the ergodic rate $\mathcal{O}\left(\delta L \exp(-K)/\mu + (C+D)/K\mu\right)$, where δ 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 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, see Appendix G. Finally, we propose several new biased compressors with promising theoretical guarantees and practical performance.

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 [1]. The data is typically collected in a distributed manner and stored across a network of edge devices, as is the case in federated learning [2, 3, 4, 5], 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

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

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 $f_i : \mathbb{R}^d \to \mathbb{R}$ often has the form $f_i(x) \coloneqq \mathrm{E}_{\xi\sim\mathcal{P}_i}\left[f_\xi(x)\right]$, with $\mathcal{P}_i$ being the distribution of training data owned by worker i.

1.1 Distributed optimization. A fundamental baseline for solving problem (1) is (distributed) gradient descent (GD), with iterations $x^{k+1} = x^k - \frac{\eta^k}{n}\sum_{i=1}^{n}\nabla f_i(x^k)$, where $\eta^k > 0$ is a stepsize.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), SpicyFL Workshop.

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

Compressor $\mathcal{C}$	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 [17]	✗	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]**	✓					$\frac{1}{4}(b+1/b+2)$
Biased exponential rounding **[NEW]**	✗	$(2/b+1)^2$	$2b/b+1$	$2/b+1$	$(b+1)^2/4b$	
Natural compression [16]	✓					$9/8$
General exponential dithering **[NEW]**	✓					ζ_b
Natural dithering [16]	✓					ζ_2
Top-k + exponential dithering **[NEW]**	✗	k/d	ζ_b	k/d	$d\zeta_b/k$	

Table 2: 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.

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 11	Theorem 12	Theorem 13
Complexity	$\mathcal{O}\left(\frac{\beta^2}{\alpha}\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\beta}{\gamma}\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\delta\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$

Several enhancements to GD have been proposed that can better deal with the communication cost challenges of distributed environments, including acceleration [6, 7, 8], reducing the number of communication rounds, and communication compression [9, 10, 11, 12, 13, 14, 15, 16], reducing the size of communicated messages.

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 [17], or deterministic rounding [18]. 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, and there is no good convergence theory for any gradient-type method with a biased compression in the crucially important $n>1$ 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 they are alternative parameterization of the same collection of operators (the last two more favorable than the first), 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 3. A summary of all compressors considered can be found in Table 1.

(c) In Section 4 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 2.

(d) 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 design a new distributed SGD method equipped with an error-feedback mechanism which can provably handle biased compressors. In our main result (Theorem 14; also see Table 3) 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 compression, mostly focusing on unbi-

Table 3: Ergodic convergence of distributed SGD with biased compression and error-feedback (Algorithm 1) for L-smooth and μ-strongly convex functions with K communications (Theorem 14).

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

ased compressions [10] as these are much easier to analyze. Works concerning biased compressions show strong empirical results with limited or no analysis [19, 20, 21]. There have been several attempts trying to address this issue, e.g., [22] provides analysis for quadratics in distributed setting, [23] gives 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 in [24], but only for one node and under bounded gradient norm assumption, which was later overcome in [25].

1.4 Basic notation and definitions. We use $\langle x, y\rangle \coloneqq \sum_{i=1}^d x_i y_i$ 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 $\|x\|_2 \coloneqq \sqrt{\langle x, x\rangle}$. We denote ℓ_p-norms as $\|x\|_p \coloneqq (\sum_{i=1}^d |x_i|^p)^{1/p}$ for $p \in (1, \infty)$. By $\mathrm{E}\left[\cdot\right]$ we denote mathematical expectation. Function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if it is differentiable and $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2, \forall x, y \in \mathbb{R}^d$. It is μ-strongly convex if it is differentiable and $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2, \forall x, y \in \mathbb{R}^d$.

2 Biased Compressors: Definitions & Theory

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

Definition 1. We say that $\mathcal{C} \in \mathbb{U}(\zeta)$ for $\zeta \geq 1$ if $\mathrm{E}\left[\mathcal{C}(x)\right] = x$, $\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \zeta \|x\|_2^2, \forall x \in \mathbb{R}^d$.

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.

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

$$\alpha \|x\|_2^2 \leq \mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta \langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle, \qquad \text{for all } x \in \mathbb{R}^d. \tag{2}$$

The second ineq. in (2) implies $\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta^2 \|x\|_2^2$.

Definition 3. We say that $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$ for some $\gamma, \beta > 0$ if

$$\max\left\{\gamma \|x\|_2^2, \tfrac{1}{\beta}\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right]\right\} \leq \langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle, \qquad \text{for all } x \in \mathbb{R}^d. \tag{3}$$

Definition 4. We say that $\mathcal{C} \in \mathbb{B}^3(\delta)$ for some $\delta > 0$ if

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \leq \left(1 - \frac{1}{\delta}\right)\|x\|_2^2, \qquad \text{for all } x \in \mathbb{R}^d. \tag{4}$$

This last definition was also considered in [26, 27]. We now establish several basic properties and connections between the classes.

Theorem 1 (Equivalence between biased compressors)**.** *Let $\lambda > 0$ be a free scaling parameter.*
If $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$, then (i) $\beta^2 \geq \alpha$ and $\lambda\mathcal{C} \in \mathbb{B}^1(\lambda^2\alpha, \lambda\beta)$, (ii) $\mathcal{C} \in \mathbb{B}^2(\alpha, \beta^2)$ and $\frac{1}{\beta}\mathcal{C} \in \mathbb{B}^3(\beta^2/\alpha)$.
If $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$, then (i) $\beta \geq \gamma$ and $\lambda\mathcal{C} \in \mathbb{B}^2(\lambda\gamma, \lambda\beta)$, (ii) $\mathcal{C} \in \mathbb{B}^1(\gamma^2, \beta)$ and $\frac{1}{\beta}\mathcal{C} \in \mathbb{B}^3(\beta/\gamma)$.
If $\mathcal{C} \in \mathbb{B}^3(\delta)$, then (i) $\delta \geq 1$ and (ii) $\mathcal{C} \in \mathbb{B}^2\left(\frac{1}{2\delta}, 2\right) \subseteq \mathbb{B}^1\left(\frac{1}{4\delta^2}, 2\right)$.

With a proper scaling any unbiased compressor belongs to all the three classes of biased compressors.

Theorem 2 (From unbiased to biased with scaling)**.** *If $\mathcal{C} \in \mathbb{U}(\zeta)$, then scaled operator $\lambda\mathcal{C}$ belongs to*

$$(i)\ \mathbb{B}^1(\lambda^2, \lambda\zeta) \text{ if } \lambda > 0, \qquad (ii)\ \mathbb{B}^2(\lambda, \lambda\zeta) \text{ if } \lambda > 0, \qquad (iii)\ \mathbb{B}^3\left(\tfrac{1}{\lambda(2-\zeta\lambda)}\right) \text{ if } \zeta\lambda \in (0, 2).$$

3 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 1.

(a) For $k \in [d] := \{1, \dots, d\}$, the **unbiased random (aka Rand-k) sparsification** operator is defined via $\mathcal{C}(x) := \frac{d}{k} \sum_{i \in S} x_i e_i$, 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 3. *The Rand-k sparsification operator belongs to $\mathbb{U}(\frac{d}{k})$.*

(b) Let $S \subseteq [d]$ be a random set, with $p_i := \text{Prob}(i \in S) > 0$, $\forall i \in [d]$ (such a set is called a proper sampling [28]). Define **biased random sparsification** operator via $\mathcal{C}(x) := \sum_{i \in S} x_i e_i$.

Lemma 4. *With $q := \min_i p_i$, the biased random sparsifier belongs to $\mathbb{B}^1(q, 1)$, $\mathbb{B}^2(q, 1)$, $\mathbb{B}^3(1/q)$.*

(c) Adaptive random sparsification is defined via $\mathcal{C}(x) := x_i e_i$ with probability $\frac{|x_i|}{\|x\|_1}$.

Lemma 5. *Adaptive random sparsification operator 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 $\mathcal{C}(x) := \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)}$, where coordinates are ordered by their magnitudes so that $|x_{(1)}| \leq |x_{(2)}| \leq \cdots \leq |x_{(d)}|$.

Lemma 6. *Top-k sparsification operator 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 increasing sequence of positive numbers with $\inf a_k = 0$ and $\sup a_k = \infty$. Then **general unbiased rounding** is defined as follows: if $a_k \leq |x_i| \leq a_{k+1}$ for some $i \in [d]$, then

$$\mathcal{C}(x)_i = \begin{cases} \text{sign}(x_i) a_k & \text{with probability } \frac{a_{k+1} - |x_i|}{a_{k+1} - a_k} \\ \text{sign}(x_i) a_{k+1} & \text{with probability } \frac{|x_i| - a_k}{a_{k+1} - a_k} \end{cases} \tag{5}$$

Lemma 7. *General unbiased rounding operator (5) 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 minimized for exponential roundings $a_k = b^k$ with some basis $b > 1$.

(f) Let $(a_k)_{k \in \mathbb{Z}}$ be defined as in **(e)**. Then **general biased rounding** is defined via

$$\mathcal{C}(x)_i := \text{sign}(x_i) \, \text{argmin}_{t \in (a_k)} |t - |x_i||, \qquad i \in [d]. \tag{6}$$

Lemma 8. *General biased rounding operator (6) belongs to $\mathbb{B}^1(\alpha, \beta)$, $\mathbb{B}^2(\gamma, \beta)$, and $\mathbb{B}^3(\delta)$, where*

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

Remark 1. In the case of exponential rounding $a_k = b^k$, we get $\alpha = \frac{4}{(b+1)^2}$, $\beta = \frac{2b}{b+1}$, $\gamma = \frac{2}{b+1}$, $\delta = \frac{(b+1)^2}{4b}$. Plugging these parameters into the iteration complexities of Table 2, 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 $\mathcal{C}_{nat}$ [16] is the special case of general unbiased rounding operator (5) when $b = 2$. Thus, $\mathcal{C}_{nat} \in \mathbb{U}\left(\frac{9}{8}\right)$.

(h) For base $b > 1$, we define **general exponential dithering** operator with respect to p-norm and with s exponential levels $0 < b^{1-s} < b^{2-s} < \cdots < b^{-1} < 1$ as follows $\mathcal{C}(x) := \|x\|_p \times \text{sign}(x) \times \xi\left(\frac{|x_i|}{\|x\|_p}\right)$, 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, preserving unbiasedness.

Lemma 9. *General exponential dithering operator belongs to $\mathbb{U}(\zeta_b)$ with*

$$\zeta_b = \tfrac{1}{4}(b + \tfrac{1}{b} + 2) + d^{\frac{1}{r}} b^{1-s} \min(1, d^{\frac{1}{r}} b^{1-s}), \quad \text{where } r = \min(p, 2). \tag{7}$$

(i) Natural dithering [16] without norm compression is the spacial case of **(h)** when $b = 2$.

(j) Top-k combined with exponential dithering. Let $\mathcal{C}_{top}$ be the Top-k sparsification operatorand $\mathcal{C}_{dith}$ be general exponential dithering operator with some base $b > 1$ and parameter ζ_b from (7). Define a new compression operator as the composition of these two: $\mathcal{C}(x) := \mathcal{C}_{dith}(\mathcal{C}_{top}(x))$.

Lemma 10. *The composition operator 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* (7)*.*

4 Gradient Descent with Biased Compression (CGD)

We now consider the unconstrained optimization problem $\min_{x\in\mathbb{R}^d} f(x)$, where $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth and μ-strongly convex. We study the method $\boxed{x^{k+1}=x^k-\eta\mathcal{C}^k(\nabla f(x^k))}$, where $\mathcal{C}^k:\mathbb{R}^d\to\mathbb{R}^d$ 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.

4.1 Complexity theory. We now establish three theorems, one for each of the three classes $\mathbb{B}^1$, $\mathbb{B}^2$ and $\mathbb{B}^3$. Let $\mathcal{E}_k \coloneqq \mathrm{E}\left[f(x^k)\right]-f(x^\star)$, with $\mathcal{E}_0=f(x^0)-f(x^\star)$.

Theorem 11. *Let $\mathcal{C}\in\mathbb{B}^1(\alpha,\beta)$. Then as long as $0\le\eta\le\frac{2}{\beta L}$, we have $\mathcal{E}_k\le\left(1-\frac{\alpha}{\beta}\eta\mu(2-\eta\beta L)\right)^k\mathcal{E}_0$. If we choose $\eta=\frac{1}{\beta L}$, then $\mathcal{E}_k\le\left(1-\frac{\alpha}{\beta^2}\frac{\mu}{L}\right)^k\mathcal{E}_0$.*

Theorem 12. *Let $\mathcal{C}\in\mathbb{B}^2(\gamma,\beta)$. Then as long as $0\le\eta\le\frac{2}{\beta L}$, we have $\mathcal{E}_k\le\left(1-\gamma\eta\left(2-\eta\beta\right)L)\right)^k\mathcal{E}_0$. If we choose $\eta=\frac{1}{\beta L}$, then $\mathcal{E}_k\le\left(1-\frac{\gamma}{\beta}\frac{\mu}{L}\right)^k\mathcal{E}_0$.*

Theorem 13. *Let $\mathcal{C}\in\mathbb{B}^3(\delta)$. Then as long as $0\le\eta\le\frac{1}{L}$, we have $\mathcal{E}_k\le\left(1-\frac{1}{\delta}\eta\mu\right)^k\mathcal{E}_0$. If we choose $\eta=\frac{1}{L}$, then $\mathcal{E}_k\le\left(1-\frac{1}{\delta}\frac{\mu}{L}\right)^k\mathcal{E}_0$.*

The iteration complexity for these results can be found in Table 2. 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 1).

4.2 Classes $\mathbb{B}^3$ and $\mathbb{B}^2$ are better than $\mathbb{B}^1$. If $\mathcal{C}\in\mathbb{B}^1(\alpha,\beta)$, then by Theorem 1, $\frac{1}{\beta}\mathcal{C}\in\mathbb{B}^3(\frac{\beta^2}{\alpha})$. Applying Theorem 13, we get the bound $\mathcal{O}\left(\frac{\beta^2}{\alpha}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$. This is the same result as that obtained by Theorem 11. On the other hand, if $\mathcal{C}\in\mathbb{B}^3(\delta)$, then by Theorem 1, $\mathcal{C}\in\mathbb{B}^1(\frac{1}{4\delta^2},2)$. Applying Theorem 11, we get the bound $\mathcal{O}\left(16\delta^2\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$. This is a worse result than what Theorem 13 offers by a factor of 16δ. Hence, while $\mathbb{B}^1$ and $\mathbb{B}^3$ describe the same classes of compressors, for the purposes of CGD it is better to parameterize them as members of $\mathbb{B}^3$.

5 Distributed Setting

We now focus attention on a distributed setup with n machines, each of which owns data defining one loss function f_i. Our goal is to minimize the average loss (1).

5.1 Distributed CGD with unbiased compressors (DCGD). Perhaps the most straightforward extension of CGD to the distributed setting is to consider the method $x^{k+1}=x^k-\eta\frac{1}{n}\sum_{i=1}^n\mathcal{C}_i^k(\nabla f_i(x^k))$. 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 with a suitable stepsize, the method converges to a $\mathcal{O}\left(\eta D(\zeta-1)/\mu n\right)$ neighborhood of the (necessarily unique) solution $x^\star$ with the linear rate $\mathcal{O}\left(\left(L/\mu+L(\zeta-1)/\mu n\right)\log 1/\epsilon\right)$, where $D\coloneqq 1/n\sum_{i=1}^n\|\nabla f_i(x^\star)\|_2^2$ [29].

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

Example 1. Consider $n=d=3$ (see F.1 for an extension to this example) and define

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

where $a=(-3,2,2),b=(2,-3,2),c=(2,2,-3)$. Then, with the initial point $x^0=(t,t,t),\ t>0$

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

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

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

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 $k = 0, 1, 2, \ldots, K$ **do** {In parralel on each machine}
Receive x^k from server and perform $\tilde{g}_i^k = \mathcal{C}_i^k(e_i^k + \eta^k g_i^k)$, $e_i^{k+1} = e_i^k + \eta^k g_i^k - \tilde{g}_i^k$.
Send $\tilde{g}_i^k$ to the server, which aggregates all the updates: $x^{k+1} = x^k - {}^1\!/_n \sum_{i=1}^n \tilde{g}_i^k$
end for
Output: Weighted average of the iterates: $\overline{x}^K$

Repeated application gives $x^k = \left(1 + \frac{11\eta}{6}\right)^k x^0$, which diverges exponentially fast to $+\infty$ as $\eta > 0$.

This example shows that the convergence guarantee cannot be established for problem (1) if Distributed SGD with biased compressor is used as a solver and one needs to devise a different approach.

5.3 Error Feedback. We show that distributed version of Distributed SGD wtih Error-Feedback [24], 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 $g_i^k = \nabla f_i(x^k) + \xi_i^k$, 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 $\overline{x}^K$ is an ergodic average of the form $\overline{x}^K := \frac{1}{W^K}\sum_{k=0}^K w^k x^k$, $W^K := \sum_{k=0}^K w^k$.

5.4 Complexity theory. We assume the stochastic error ξ_i^k satisfies the following condition.

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

$$\mathrm{E}\left[\left\|\xi_i^k\right\|_2^2\right] \leq B\left\|\nabla f_i(x^k)\right\|_2^2 + C, \qquad \text{for all } i \in [n],\ k \geq 0. \tag{8}$$

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.

Theorem 14 (Main). *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 $x^\star$ be the minimizer of f and let $f^\star := f(x^\star)$ and $D := \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|_2^2$. 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) $\mathcal{O}(1/k)$ **stepsizes &** $\mathcal{O}(k)$ **weights.** *Let $\eta^k = {}^4\!/_{\mu(\kappa+k)}$ be the stepsizes and $w^k = \kappa + k$ be the weights for all $k \geq 0$, where $\kappa = 56(2\delta + B)L/\mu$. Then*

$$\boxed{\mathrm{E}\left[f(\overline{x}^K)\right] - f^\star = \mathcal{O}\left({}^{A_1}\!/_{K^2} + {}^{A_2}\!/_K\right)},$$

where $A_1 := {}^{L^2(2\delta+B)^2}\!/_\mu \left\|x^0 - x^\star\right\|_2^2$ and $A_2 := {}^{C\left(1+\frac{1}{n}\right)+D({}^{2B}\!/_n+3\delta)}\!/_\mu$.
(ii) $\mathcal{O}(1)$ **stepsizes &** $\mathcal{O}(e^{-k})$ **weights.** *Let $\eta^k = \eta \leq \frac{1}{14(2\delta+B)L}$ be the stepsizes and $w^k = (1 - \mu\eta/2)^{-(k+1)}$ be the weights for all $k \geq 0$. Then*

$$\boxed{\mathrm{E}\left[f(\overline{x}^K)\right] - f^\star = \tilde{\mathcal{O}}\left(A_3 \exp\left[-{}^K\!/_{A_4}\right] + {}^{A_2}\!/_K\right)},$$

where $A_3 := L(2\delta + B)\left\|x^0 - x^\star\right\|_2^2$ and $A_4 := {}^{28L(2\delta+B)}\!/_\mu$.
(iii) $\mathcal{O}(1)$ **stepsizes & equal weights.** *Let $\eta^k = \eta \leq \frac{1}{14(2\delta+B)L}$ be the stepsizes and $w^k = 1$ be the weights for all $k \geq 0$. Then, letting $A_5 := \sqrt{C\left(1 + {}^1\!/_n\right) + D\left({}^{2B}\!/_n + 3\delta\right)}\left\|x^0 - x^\star\right\|_2$,*

$$\boxed{\mathrm{E}\left[f(\overline{x}^K)\right] - f^\star = \mathcal{O}\left({}^{A_3}\!/_K + {}^{A_5}\!/_{\sqrt{K}}\right)}.$$

Experimental evaluation can be found in Appendix A.

References

[1] 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.

[2] 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.

[3] 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.

[4] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

[5] Peter Kairouz and et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[6] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Hyeontaek Lim, David G Andersen, and Michael Kaminsky. 3lc: Lightweight and effective traffic compression for distributed machine learning. *arXiv preprint arXiv:1802.07389*, 2018.

[13] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5977–5987. Curran Associates, Inc., 2018.

[14] 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.

[15] Mher Safaryan and Peter Richtárik. On stochastic sign descent methods. *arXiv preprint arXiv:1905.12938*, 2019.

[16] Samuel Horváth, Chen-Yu Ho, L'udovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

[17] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. In *32nd Conference on Neural Information Processing Systems*, 2018.

[18] 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.

[19] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. *CoRR*, abs/1905.13727, 2019.

[20] 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, 2017.

[21] 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, Juggapong Natwichai, and Yongxin Tong, editors, *Database Systems for Advanced Applications*, pages 139–155, Cham, 2019. Springer International Publishing.

[22] 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.

[23] 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.

[24] 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*, 2019.

[25] 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.

[26] 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.

[27] Jean-Baptiste Cordonnier. Convex optimization using sparsified stochastic gradient descent with memory. Technical report, 2018.

[28] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

[29] 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*, 2020.

[30] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[31] 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*, 2019.

[32] 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.

[33] Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in order Statistics*. John Wiley and Sons Inc., 1992.

Appendix

Contents

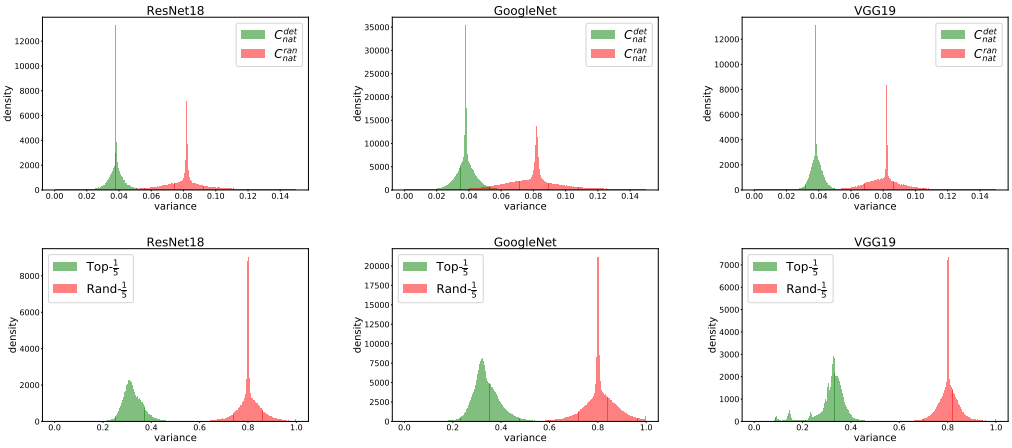


Figure 1: Comparison of empirical variance $\|\mathcal{C}(x) - x\|_2^2 / \|x\|_2^2$ during training procedure of ResNet18, GoogleNet, and VGG19 on CIFAR10 dataset for two pairs of methods–deterministic with classic/unbiased $\mathcal{C}_{\rm nat}$ and Top-k with Rand-k, where $k = 1/5d$.

A Experiments

We conduct several experiments to support our theoretical results. We implement all methods in Python 3.7 using Pytorch [32] and run on a machine with 24 Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz cores, GPU @GeForce GTX 1080 Ti with memory 11264 MB (Cuda 10.1).

As biased compressions were already shown to perform better in distributed settings [14, 12], we rather focus on the reasoning why this is the case. We conduct simulated experiments on one machine which enable us to do rapid direct comparisons against the prior methods. Another issue is that for many methods, there is no public implementation available, which makes it hard to do a fair comparison in distributed settings, thus we focus on simulated experiments.

Motivated by our theoretical results in Section G (Appendix), 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 1, 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 $\mathcal{C}_{\rm nat}$ 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 G.

As the next experiment, we further show that our predicted theoretical behaviour 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 13

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

where

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

Note that this is the direct consequence of our analysis. We apply this property to display the theoretical convergence. Firstly, 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 $Q^\top DQ$, 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

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

For the second experiment, we run standard linear regression on scikit-learn datasets– *Boston* and *Diabetes*. As the preprocessing step, we first do data normalization.

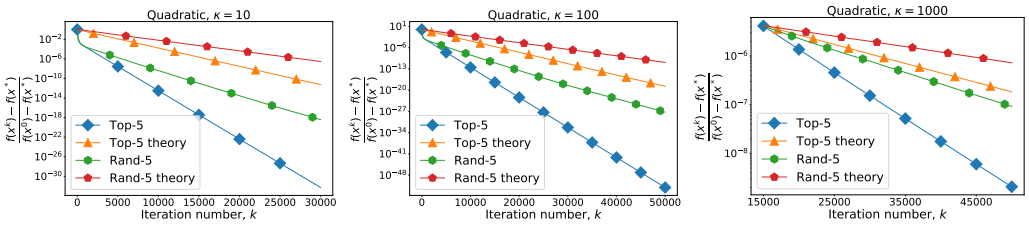


Figure 2: Theoretical vs. Practical Convergence of Compressed Gradient Descent on Quadratics problem with different condition number κ for Top-5 and Rand-5 compression operators.

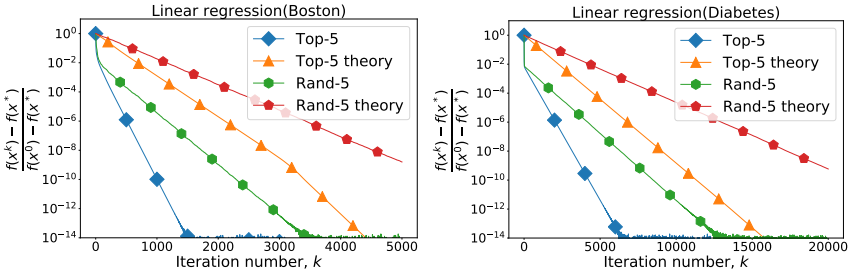


Figure 3: Theoretical vs. Practical Convergence of Compressed Gradient Descent on Linear Regression problem for *Boston* and *Diabetes* datasets with Top-5 and Rand-5 compression operators.

Looking into Figures 2 and 3, 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 as predicted by our theory.

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

In Section 3 we gave a new biased compression operator, where we combined Top-k sparsification operator with the general exponential dithering. 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 6 empirically confirms that it attains the lowest compression parameter $\delta \geq 1$ among all other known compressors (see (4)). 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.

We also conclude an experiment which shows its superiority against current state-of-the-art for low bandwidth approach Top-k for some small k. Figure 5 shows 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 compressors, we train VGG11 on CIFAR10 using plain SGD as an optimizer with default step size 0.01. We can see that adding natural dithering after Top-k has the same effect as the natural dithering comparing

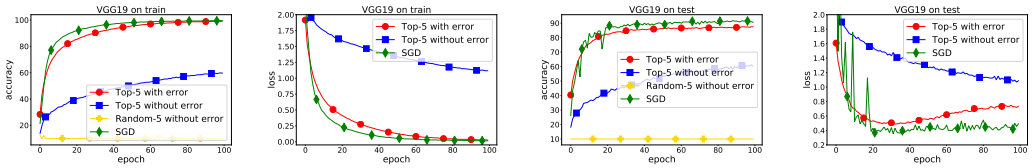


Figure 4: Training/Test loss and accuracy for VGG19 on CIFAR10 distributed among 4 nodes for 4 different compression operators.

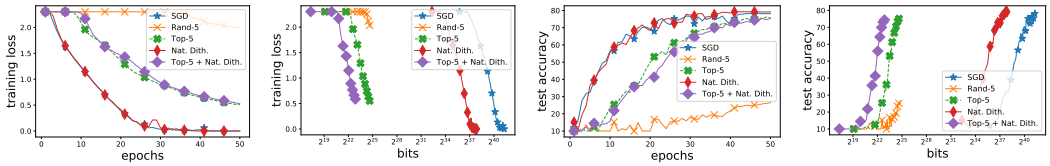


Figure 5: Training loss and test accuracy for VGG11 on CIFAR10 distributed among 4 nodes for 5 different compression operators.

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 6.

Next, we compare two sparsification methods and show the significant empirical advantage of greedy sparsifier against random sparsifier, where we assume that coordinates of to-be-compressed vector are i.i.d. Gaussian random variables. We compare the savings s^k_{top} and s^k_{rnd} of these compressors. For random sparsification, we have

$$\mathrm{E}\left[s^k_{rnd}(x)\right] = k \cdot (\sigma^2 + \mu^2),$$

where μ and σ^2 are the mean and variance of the Gaussian distribution. For computing $\mathrm{E}\left[s^k_{top}(x)\right]$, we use the probability density function of k-th order statistics (see e.g. [33]). Table 4 shows that Top-3 and Top-5 sparsifiers "save" $3\times$–$40\times$ more information in expectation and the factor grows with the dimension. Next we compare normalized variances $\frac{\omega^k_{top}(x)}{\|x\|_2^2}$ and $\frac{\omega^k_{rnd}(x)}{\|x\|_2^2}$ 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. Figure7 reveals the superiority of greedy sparsifier against the random one. In addition to assuming the gradient distribution, we obtained various gradient distributions via logistic regression (*mushrooms* LIBSVM dataset) and least squares. The second moments, i.e. energy "savings", were calculated using formula for density function of k-order statistics, see [33] for reference. We conclude experiments for Top-5 and Rand-5, see Figure 8 for details.

B Basic Facts and Inequalities

B.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:

$$\frac{\mu}{2}\|x-y\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x-y\rangle, \qquad \forall x, y \in \mathbb{R}^d. \tag{9}$$

B.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.

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

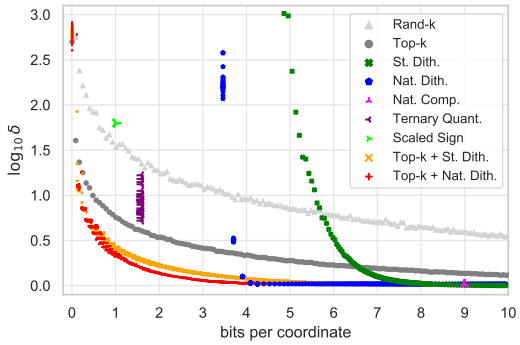


Figure 6: Comparison of various compressors with respect to the parameter $\delta \geq 1$ in $\log_{10}-$scale and the number of encoding bits used for each coordinate on average. Each point/marker represents a single $d = 10^4$ dimensional vector x drawn from Gaussian distribution and then compressed by the specified operator.

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

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

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

$$\frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle, \qquad \forall x, y \in \mathbb{R}^d \tag{10}$$

By plugging $y = x^*$ to (10), we get

$$\left\|\nabla f(x)\right\|_2^2 \leq 2L(f(x) - f(x^*)), \qquad \forall x \in \mathbb{R}^d. \tag{11}$$

B.3 Useful inequalities

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

$$2\langle a, b\rangle \leq \frac{\left\|a\right\|_2^2}{\xi} + \xi\left\|b\right\|_2^2, \tag{12}$$

$$\left\|a + b\right\|_2^2 \leq \left(1 + \frac{1}{\xi}\right)\left\|a\right\|_2^2 + (1 + \xi)\left\|b\right\|_2^2, \tag{13}$$

$$\left\|\sum_{i=1}^n x_i\right\|_2^2 \leq n \cdot \sum_{i=1}^n \left\|x_i\right\|_2^2. \tag{14}$$

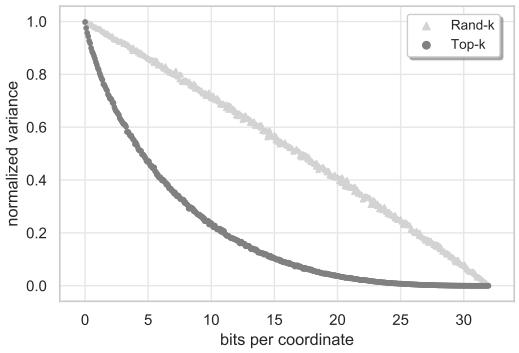
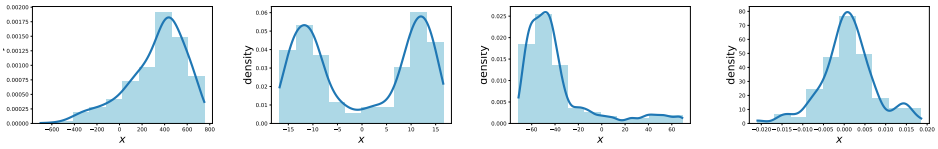


Figure 7: The comparison of Top-k and Rand-k sparsifiers w.r.t. normalized variance and the number of encoding bits used for each coordinate on average. Each point/marker represents a single $d = 10^4$ dimensional vector drawn form Gaussian distribution and then compressed by the specified operator. 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.



(a) $s^5_{rnd} = 9 \cdot 10^5$, $s^5_{top} = 28 \cdot 10^5$

(b) $s^5_{rnd} = 624$, $s^5_{top} = 1357$

(c) $s^5_{rnd} = 11921$, $s^5_{top} = 23488$

(d) $s^5_{rnd} = 2 \cdot 10^{-4}$, $s^5_{top} = 17 \cdot 10^{-4}$

Figure 8: 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.

C Proofs for Section 2

C.1 Lemma

Lemma 15. *For any $x \in \mathbb{R}^d$, if $\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta \langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle$, then*

$$\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta^2 \|x\|_2^2 . \tag{15}$$

Proof. Fix any $x \in \mathbb{R}^d$. Applying Jensen's inequality, the second inequality in (2) and Cauchy-Schwarz, we get

$$\|\mathrm{E}\left[\mathcal{C}(x)\right]\|_2^2 \leq \mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \overset{(2)}{\leq} \beta \langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle \leq \beta \|\mathrm{E}\left[\mathcal{C}(x)\right]\|_2 \|x\|_2 . \tag{16}$$

If $\mathrm{E}\left[\mathcal{C}(x)\right] \neq 0$, this implies $\|\mathrm{E}\left[\mathcal{C}(x)\right]\|_2 \leq \beta \|x\|_2$. Plugging this back into (16), we get (15). If $\mathrm{E}\left[\mathcal{C}(x)\right] = 0$, then from (2) we see that $\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] = 0$, and (15) holds trivially. □

C.2 Proof of Theorem 1

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

(i) Let us choose any $x \neq 0$ and observe that (2) implies that $\mathrm{E}\left[\mathcal{C}(x)\right] \neq 0$. Further, from (2) we get the bounds

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

Finally,

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

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 (2).

(ii) 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

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

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

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

(i) Using (3) we get

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

where the first and third inequalities follow from (3) 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 (3).

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

$$\gamma^2\|x\|_2^4 \overset{(3)}{\leq} \langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle^2 \leq \|\mathrm{E}\left[\mathcal{C}(x)\right]\|_2^2\|x\|_2^2 \leq \mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right]\|x\|_2^2,$$

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

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

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

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

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

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

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

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

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

which implies that

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

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

□

C.3 Proof of Theorem 2

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

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

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

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

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

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

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

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

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

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

□

D Proofs for Section 3

D.1 Proof of Lemma 3: Unbiased Random Sparsification

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

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

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

D.2 Proof of Lemma 4: Biased Random Sparsification

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

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

Letting $q \coloneqq \min_i p_i$, we get

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

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

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

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

D.3 Proof of Lemma 5: Adaptive Random Sparsification

From the definition of the compression operator, we have

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

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

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

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

D.4 Proof of Lemma 6: Top-k sparsification

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

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

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

D.5 Proof of Lemma 7: General Unbiased Rounding

The unbiasedness follows immediately from the definition (5)

$$\mathrm{E}\left[\mathcal{C}(x)\right] = \sum_{i=1}^{d} \mathrm{E}\left[\mathcal{C}(x)_i\right] e_i = \sum_{i=1}^{d} \mathrm{sign}(x_i)\left(a_k \frac{a_{k+1} - |x_i|}{a_{k+1} - a_k} + a_{k+1}\frac{|x_i| - a_k}{a_{k+1} - a_k}\right) e_i = \sum_{i=1}^{d} x_i e_i = x. \tag{17}$$

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 $\mathrm{E}\left[\mathcal{C}(t)^2\right] \leq \zeta \cdot t^2$. From the definition we compute the second moment as follows

$$\mathrm{E}\left[\mathcal{C}(t)^2\right] = a_k^2 \frac{a_{k+1} - t}{a_{k+1} - a_k} + a_{k+1}^2 \frac{t - a_k}{a_{k+1} - a_k} = (a_k + a_{k+1})t - a_k a_{k+1} = t^2 + (t - a_k)(a_{k+1} - t), \tag{18}$$

from which

$$\frac{\mathrm{E}\left[\mathcal{C}(t)^2\right]}{t^2} = 1 + \left(1 - \frac{a_k}{t}\right)\left(\frac{a_{k+1}}{t} - 1\right), \quad a_k \le t \le a_{k+1}. \tag{19}$$

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

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

Thus, the parameter ζ for general unbiased rounding would be

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

D.6 Proof of Lemma 8: General Biased Rounding

From the definition (6) of compression operator $\mathcal{C}$ we derive the following inequalities

$$\inf_{k\in\mathbb{Z}}\left(\frac{2a_k}{a_k + a_{k+1}}\right)^2 \|x\|_2^2 \le \|\mathcal{C}(x)\|_2^2,$$

$$\|\mathcal{C}(x)\|_2^2 \le \sup_{k\in\mathbb{Z}} \frac{2a_{k+1}}{a_k + a_{k+1}} \langle \mathcal{C}(x), x\rangle,$$

$$\inf_{k\in\mathbb{Z}} \frac{2a_k}{a_k + a_{k+1}} \|x\|_2^2 \le \langle \mathcal{C}(x), x\rangle,$$

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

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

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 (6), we get

$$\frac{(\mathcal{C}(t) - t)^2}{t^2} = \min\left[\left(1 - \frac{a_k}{t}\right)^2, \left(1 - \frac{a_{k+1}}{t}\right)^2\right], \qquad a_k \le t \le a_{k+1}. \tag{20}$$

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 (20), 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

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

which gives

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

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

D.7 Proof of Lemma 9: General Exponential Dithering

The proof goes with the same steps as in Theorem 4 of [16]. 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 (17) was done. Then we note that

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

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 $\mathcal{C}$:

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

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 $p \geq 2$.

D.8 Proof of Lemma 10: Top-k Combined with Exponential Dithering

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

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

from which we conclude $\langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle = \langle \mathcal{C}_{top}(x), x\rangle = \|\mathcal{C}_{top}(x)\|_2^2$. Next, using Lemma 9 on exponential dithering we get

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

which implies $\beta = \zeta_b$. Using Lemma 6 we show $\gamma = \frac{k}{d}$ as $\langle \mathrm{E}\left[\mathcal{C}(x)\right], x\rangle = \|\mathcal{C}_{top}(x)\|_2^2 \geq \frac{k}{d}\|x\|_2^2$. Utilizing the derivations (18) and (19) it can be shown that $\mathrm{E}\left[\|\mathcal{C}_{dith}(x)\|_2^2\right] \geq \|x\|_2^2$ and therefore

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

Hence, $\alpha = \frac{k}{d}$. To compute the parameter δ we use Theorem 1, which yields $\delta = \frac{\beta}{\gamma} = \frac{d}{k}\zeta_b$.

E Proofs for Section 4

We now perform analysis of CGD for compression operators in $\mathbb{B}^1$, $\mathbb{B}^2$ and $\mathbb{B}^3$, establishing Theorems 11, 12 and 13, respectively.

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

Lemma 16. *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*

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

Proof. Letting $g = \nabla f(x)$, we have[1]

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

□

Proof of Theorem 11

Proof. Since f is μ-strongly convex, $\left\|\nabla f(x^k)\right\|_2^2 \geq 2\mu(f(x^k) - f(x^\star))$. Combining this with Lemma 16 applied to $x = x^k$ and $g = \nabla f(x^k)$, we get

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

□

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

Lemma 17. *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*

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

[1] Alternatively, we can write

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

Bot approaches lead to the same bound.

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

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

□

Proof of Theorem 12

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

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

□

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

Lemma 18. *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*

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

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

$$\begin{aligned}
\mathrm{E}\left[f\left(x - \eta\mathcal{C}(g)\right)\right] &\leq \mathrm{E}\left[f(x) + \langle g, -\eta\mathcal{C}(g)\rangle + \frac{L}{2}\left\|-\eta\mathcal{C}(g)\right\|_2^2\right] \\
&= f(x) - \eta\langle \mathrm{E}\left[\mathcal{C}(g)\right], g\rangle + \frac{\eta^2 L}{2}\mathrm{E}\left[\left\|\mathcal{C}(g)\right\|_2^2\right]. \qquad (21)
\end{aligned}$$

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

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

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

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

Assuming that $\eta L \leq 1$, we can combine this with (21) and the lemma is proved.

□

Proof of Theorem 13

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

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

□

F Proofs for Section 5

F.1 Failure of DCGD with biased compressors: an extension to Example 1

Here we extend the example given in Section 5 showing a potential divergence of DCGD with biased compression. 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

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

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

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

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

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

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 $+\infty$.

F.2 Proof of Theorem 14 (Main)

In this section, we include our analysis for the Distributed SGD with biased compression. Our analysis is closely related to the analysis of [25].

We start with the definition of some auxiliary objects:

Definition 5. The sequence $\{a^k\}_{k \geq 0}$ of positive values is τ-slow decreasing for parameter τ:

$$a^{k+1} \leq a^k, \quad a^{k+1}\left(1 + \frac{1}{2\tau}\right) \geq a^k, \quad \forall k \geq 0 \tag{22}$$

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

$$a^{k+1} \geq a^k, \quad a^{k+1} \leq a^k\left(1 + \frac{1}{2\tau}\right), \quad \forall k \geq 0 \tag{23}$$

And let:

$$\tilde{x}^k = x^k - \frac{1}{n} \sum_{i=1}^n e_i^k, \quad \forall k \geq 0 \tag{24}$$

$$g^k = \frac{1}{n}\sum_{i=1}^n g_i^k \tag{25}$$

It is easy to see:

$$\begin{aligned}
\tilde{x}^{k+1} &= x^{k+1} - \frac{1}{n}\sum_{i=1}^n e_i^{k+1} \\
&\overset{Alg.\ 1}{=} \left(x^k - \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^k\right) - \left(\frac{1}{n}\sum_{i=1}^n [e_i^k + \eta^k g_i^k - \tilde{g}_i^k]\right) \\
&= \tilde{x}^k - \frac{\eta^k}{n}\sum_{i=1}^n g_i^k
\end{aligned} \tag{26}$$

Lemma 19. *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* (24)*,*

$$\begin{aligned}
\mathrm{E}\left[\left\|\tilde{x}^{k+1} - x^*\right\|_2^2\right] &\leq \left(1 - \frac{\mu\eta^k}{2}\right)\mathrm{E}\left[\left\|\tilde{x}^k - x^*\right\|_2^2\right] - \frac{\eta^k}{2}\mathrm{E}\left[f(x^k) - f^*\right] \\
&\quad + 3L\eta^k\mathrm{E}\left[\left\|x^k - \tilde{x}^k\right\|_2^2\right] + (\eta^k)^2\frac{C + 2BD}{n}
\end{aligned} \tag{27}$$

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

$$\begin{aligned}
\left\|\tilde{x}^{k+1} - x^*\right\|_2^2 &\overset{(25),(26)}{=} \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k\langle g^k, \tilde{x}^k - x^*\rangle + (\eta^k)^2\left\|g^k\right\|_2^2 \\
&= \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k\langle g^k, x^k - x^*\rangle + (\eta^k)^2\left\|g^k\right\|_2^2 + 2\eta^k\langle g^k, x^k - \tilde{x}^k\rangle.
\end{aligned}$$

Taking the conditional expectation conditioned on previous iterates, we get

$$\begin{aligned}
\mathrm{E}\left[\left\|\tilde{x}^{k+1} - x^*\right\|_2^2\right] &= \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - x^*\rangle + (\eta^k)^2\cdot\mathrm{E}\left[\left\|g^k\right\|_2^2\right] + 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - \tilde{x}^k\rangle \\
&\overset{(25)}{=} \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - x^*\rangle \\
&\quad + (\eta^k)^2\cdot\mathrm{E}\left[\left\|\nabla f(x^k) + \frac{1}{n}\sum_{i=1}^n\xi_i^k\right\|_2^2\right] + 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - \tilde{x}^k\rangle \\
&= \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - x^*\rangle \\
&\quad + (\eta^k)^2\cdot\mathrm{E}\left[\left\|\nabla f(x^k)\right\|_2^2 + 2\langle\nabla f(x^k), \frac{1}{n}\sum_{i=1}^n\xi_i^k\rangle + \left\|\frac{1}{n}\sum_{i=1}^n\xi_i^k\right\|_2^2\right] + 2\eta^k\langle\mathrm{E}\left[g^k\right], x^k - \tilde{x}^k\rangle.
\end{aligned}$$

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

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

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

$$
\begin{aligned}
&\overset{(14)}{\leq} \quad \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
&\quad +(\eta^k)^2 \cdot \left\|\nabla f(x^k)\right\|_2^2 + (\eta^k)^2 \cdot \frac{1}{n^2} \sum_{i=1}^n \mathrm{E}\left[\left\|\xi_i^k\right\|_2^2\right] + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \\
&\overset{(8)}{\leq} \quad \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
&\quad +(\eta^k)^2 \cdot \left\|\nabla f(x^k)\right\|_2^2 + \frac{(\eta^k)^2}{n^2} \sum_{i=1}^n \left[B \left\|\nabla f_i(x^k)\right\|_2^2\right] + \frac{(\eta^k)^2}{n} C \\
&\quad +2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \\
&\overset{(11)}{\leq} \quad \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle \\
&\quad +(\eta^k)^2 \cdot 2L(f(x^k) - f(x^*)) + \frac{(\eta^k)^2}{n^2} \sum_{i=1}^n \left[B \left\|\nabla f_i(x^k)\right\|_2^2\right] + \frac{(\eta^k)^2}{n} C \\
&\quad +2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle. \qquad (28)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|_2^2 &= \frac{1}{n} \sum_{i=1}^n \left\|\nabla f_i(x^k) - \nabla f_i(x_*) + \nabla f_i(x^*)\right\|_2^2 \\
&\overset{(14)}{\leq} \frac{2}{n} \sum_{i=1}^n \left(\left\|\nabla f_i(x^k) - \nabla f_i(x^*)\right\|_2^2 + \left\|\nabla f_i(x^*)\right\|_2^2\right) \\
&\overset{(10)}{\leq} \frac{2}{n} \sum_{i=1}^n \left[2L\left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^* \rangle\right) + \left\|\nabla f_i(x^*)\right\|_2^2\right].
\end{aligned}
$$

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

$$
\frac{1}{n} \sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|_2^2 \leq 4L\left(f(x^k) - \nabla f(x^*)\right) + 2D \qquad (29)
$$

Substituting (29) to (28):

$$
\begin{aligned}
\mathrm{E}\left[\left\|\tilde{x}^{k+1} - x^*\right\|_2^2\right] &= \left\|\tilde{x}^k - x^*\right\|_2^2 - 2\eta^k \langle \nabla f(x^k), x^k - x^* \rangle + (\eta^k)^2 \cdot 2L\left(1 + \frac{2B}{n}\right)(f(x^k) - f(x^*)) \\
&\quad +(\eta^k)^2 \frac{C + 2BD}{n} + 2\eta^k \langle \nabla f(x^k), x^k - \tilde{x}^k \rangle \qquad (30)
\end{aligned}
$$

By (9) we have for f:

$$
-2\langle \nabla f(x^k), x^k - x^* \rangle \leq -\mu \left\|x^k - x^*\right\|_2^2 - 2(f(x^k) - f^*). \qquad (31)
$$

Using (12) with $\xi = 1/2L$ and L-smothness of f (11):

$$
2\langle \nabla f(x^k), \tilde{x}^k - x^k \rangle \leq \frac{1}{2L} \left\|\nabla f(x^k)\right\|_2^2 + 2L \left\|x^k - \tilde{x}^k\right\|_2^2 \leq f(x^k) - f^* + 2L \left\|x^k - \tilde{x}^k\right\|_2^2. \qquad (32)
$$

By (14) for $\left\|\tilde{x}^k - x^*\right\|_2^2$, we get:

$$
-\left\|x^k - x^*\right\|_2^2 \leq -\frac{1}{2} \left\|\tilde{x}^k - x^*\right\|_2^2 + \left\|x^k - \tilde{x}^k\right\|_2^2. \qquad (33)
$$

Plugging (31), (32), (33) into (30):

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

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

Lemma 20. *$\eta^k \leq \frac{1}{14(2\delta+B)L}$, $\forall k \geq 0$ and $\{(\eta^k)^2\}_{k\geq 0}$ – 2δ-slow decreasing. Then*

$$\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n e_i^{k+1}\right\|_2^2\right] \leq \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{j=0}^k\left[\left(1-\frac{1}{4\delta}\right)^{k-j}(f(x^j)-f(x^*))\right]+\eta^k\frac{2(\delta-1)}{7L}\left(2D+\frac{C}{2\delta+B}\right). \quad (34)$$

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

$$3L\cdot\sum_{k=0}^K w^k\cdot\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n e_i^k\right\|_2^2\right] \leq \frac{1}{4}\sum_{k=0}^K w^k(\mathrm{E}\left[f(x^k)\right]-f(x_*))+\left(3\delta D+\frac{3C}{4}\right)\sum_{k=0}^K w^k\eta^k. \quad (35)$$

Proof. We prove the first part of the statement:

$$\begin{aligned}\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n e_i^{k+1}\right\|_2^2\right] &\overset{(14)}{\leq} \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^n\left\|e_i^{k+1}\right\|_2^2\right]\\ &\overset{Alg.\ 1}{=} \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^n\left\|e_i^k+\eta^k g_i^k-\tilde{g}_i^k\right\|_2^2\right]\\ &\overset{Alg.\ 1}{=} \frac{1}{n}\sum_{i=1}^n\mathrm{E}\left[\left\|e_i^k+\eta^k g_i^k-\mathcal{C}(e_i^k+\eta^k g_i^k)\right\|_2^2\right]\\ &\overset{(4)}{\leq} \frac{1-1/\delta}{n}\sum_{i=1}^n\mathrm{E}_\nabla\left[\left\|e_i^k+\eta^k g_i^k\right\|_2^2\right]\\ &= \frac{1-1/\delta}{n}\sum_{i=1}^n\mathrm{E}_\nabla\left[\left\|e_i^k+\eta^k\nabla f_i(x^k)+\eta^k\xi_i^k\right\|_2^2\right]\end{aligned}$$

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. $\mathrm{E}\left[\cdot\right]=\mathrm{E}_{\mathrm{C}}\left[\mathrm{E}_\nabla\left[\cdot\right]\right]$. Given the unbiased stochastic gradient ($\mathrm{E}\left[\xi_i^k\right]=0$):

$$\begin{aligned}\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n e_i^{k+1}\right\|_2^2\right] &\leq \frac{1-1/\delta}{n}\sum_{i=1}^n\left[\left\|e_i^k+\eta^k\nabla f_i(x^k)\right\|_2^2+\mathrm{E}_\nabla\left[\left\|\eta^k\xi_i^k\right\|_2^2\right]\right]\\ &\overset{(8)}{\leq} \frac{1-1/\delta}{n}\sum_{i=1}^n\left[\left\|e_i^k+\eta^k\nabla f_i(x^k)\right\|_2^2+(\eta^k)^2\left(B\left\|\nabla f_i(x^k)\right\|_2^2+C\right)\right]\end{aligned}$$

Using (13) with some ξ:

$$\begin{aligned}
\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^{k+1}\right\|_2^2\right] &\leq \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left\|e_i^{k+1}\right\|_2^2\right] \\
&\leq \frac{1-\frac{1}{\delta}}{n}\sum_{i=1}^{n}\left[(1+\xi)\left\|e_i^k\right\|_2^2 + (\eta^k)^2\left(1+\frac{1}{\xi}\right)\left\|\nabla f_i(x^k)\right\|_2^2 + (\eta^k)^2 B\left\|\nabla f_i(x^k)\right\|_2^2 + (\eta^k)^2 C\right] \\
&= \left(1-\frac{1}{\delta}\right)\left[(1+\xi)\left(\frac{1}{n}\sum_{i=1}^{n}\left\|e_i^k\right\|_2^2\right) + (\eta^k)^2\left(1+\frac{1}{\xi}+B\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^k)\right\|_2^2\right) + (\eta^k)^2 C\right] \\
&\overset{(29)}{\leq} \left(1-\frac{1}{\delta}\right)\left[(1+\xi)\left(\frac{1}{n}\sum_{i=1}^{n}\left\|e_i^k\right\|_2^2\right)\right] \\
&\quad + \left(1-\frac{1}{\delta}\right)\left[(\eta^k)^2\left(1+\frac{1}{\xi}+B\right)\left(4L(f(x^k)-f(x^*))+2D\right) + (\eta^k)^2 C\right]
\end{aligned}$$

Using the recurrence for $\frac{1}{n}\sum\limits_{i=1}^{n}\left\|e_i^k\right\|_2^2$, 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

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

For 2δ-slow decreasing $\{(\eta^k)^2\}_{k\geq 0}$ by definition (22) we get that $(\eta^j)^2 \leq (\eta^k)^2\left(1+\frac{1}{4\delta}\right)^{k-j}$. Due to the fact that $(1-1/2\delta)(1+1/4\delta) \leq (1-1/4\delta)$, we have:

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

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 $\eta^k \leq \frac{1}{14(2\delta+B)L}$:

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

which concludes the proof of (34). For the second part, we use the previous results. Summing over all k:

$$\begin{aligned}
\sum_{k=0}^{K} w^k\cdot\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^k\right\|_2^2\right] &\overset{(34)}{\leq} \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{k=0}^{K} w^k\sum_{j=0}^{k-1}\left(1-\frac{1}{4\delta}\right)^{k-j-1}\left(\mathrm{E}\left[f(x^j)\right]-f(x^*)\right) \\
&\quad+\frac{2(\delta-1)}{7L}\left(2D+\frac{C}{2\delta+B}\right)\sum_{k=0}^{K} w^k\eta^{k-1}
\end{aligned}$$

For 2δ-slow decreasing $\{(\eta^k)^2\}_{k\geq 0}$, it holds $(\eta^{k-1})^2 \leq (\eta^k)^2\left(1+\frac{1}{4\delta}\right)$ which follows from (22) and $\eta^{k-1}\leq \eta^k\left(1+\frac{1}{4\delta}\right)$ and for 4δ-slow increasing $\{w^k\}_{k\geq 0}$ by (23) we have $w^k \leq w^{k-j}\left(1+\frac{1}{8\delta}\right)^j$. Then

$$\begin{aligned}
\sum_{k=0}^{K} w^k\cdot\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^k\right\|_2^2\right] &\overset{(34)}{\leq} \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{k=0}^{K} w^k\sum_{j=0}^{k-1}\left(1-\frac{1}{4\delta}\right)^{k-j-1}\left(\mathrm{E}\left[f(x^j)\right]-f(x^*)\right) \\
&\quad+\frac{2(\delta-1)}{7L}\left(2D+\frac{C}{2\delta+B}\right)\left(1+\frac{1}{4\delta}\right)\sum_{k=0}^{K} w^k\eta^k \\
&\leq \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{k=0}^{K}\sum_{j=0}^{k-1} w^j\left(1+\frac{1}{8\delta}\right)^{k-j}\left(1-\frac{1}{4\delta}\right)^{k-j}\left(\mathrm{E}\left[f(x^j)\right]-f(x^*)\right) \\
&\quad+\frac{\delta-1}{2L}\left(2D+\frac{C}{2\delta+B}\right)\sum_{k=0}^{K} w^k\eta^k \\
&\leq \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{k=0}^{K}\sum_{j=0}^{k-1} w_j\left(1-\frac{1}{8\delta}\right)^{k-j}\left(\mathrm{E}\left[f(x^j)\right]-f(x^*)\right) \\
&\quad+\frac{\delta-1}{2L}\left(2D+\frac{C}{2\delta+B}\right)\sum_{k=0}^{K} w^k\eta^k \\
&\leq \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{k=0}^{K} w^k\left(\mathrm{E}\left[f(x^k)\right]-f(x^*)\right)\sum_{j=0}^{\infty}\left(1-\frac{1}{8\delta}\right)^j \\
&\quad+\frac{\delta-1}{2L}\left(2D+\frac{C}{2\delta+B}\right)\sum_{k=0}^{K} w^k\eta^k.
\end{aligned}$$

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. □

Lemma 21 (Lemma 11, [25]). *For decreasing stepsizes $\left\{\eta^k \coloneqq \frac{2}{a(\kappa+k)}\right\}_{k\geq 0}$, and weights $\{w_k \coloneqq (\kappa+k)\}_{k\geq 0}$ 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*

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

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

Proof. We start by observing that

$$\frac{w^k}{\eta^k}\left(1-a\eta^k\right)r^k = \frac{a}{2}(\kappa+k)(\kappa+k-2)r^k = \frac{a}{2}\left((\kappa+k-1)^2-1\right) \leq \frac{a}{2}(\kappa+k-1)^2\,. \quad (36)$$

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

$$\Psi^K \overset{(36)}{\leq} \frac{1}{W^K}\sum_{k=0}^{K}\left(\frac{a}{2}(\kappa+k-1)^2r^k - \frac{a}{2}(\kappa+k)^2r^{k+1}\right) + \sum_{k=0}^{K}\frac{2c}{aW^K} \leq \frac{a(\kappa-1)^2r^0}{2W^K} + \frac{2c(K+1)}{aW^K}.$$

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

Lemma 22 (Lemma 12, [25])**.** *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 $w^k \coloneqq (1-a\eta)^{-(k+1)}$ it holds*

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

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

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

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

Lemma 23 (Lemma 13, [25])**.** *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:*

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

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. □

The proof of the main theorem follows

Proof of the Theorem 14. It is easy to see that $1/14(2\delta+B)L \leq 1/4L(1+2B/n)$. This means that the Lemma 19 is satisfied. With the notation $r^k := \mathrm{E}\left[\left\|\tilde{x}^{k+1}-x^\star\right\|_2^2\right]$ and $s^k := \mathrm{E}\left[f(x^k)\right]-f^\star$ we have for any $w^k>0$:

$$\frac{w^k}{2}s^k \overset{(27)}{\leq} \frac{w^k}{\eta^k}\left(1-\frac{\mu\eta^k}{2}\right)r^k - \frac{w^k}{\eta^k}r^{k+1} + \eta^k w^k \frac{C+2BD}{n} + 3w^k L\cdot \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n e_i^k\right\|_2^2\right].$$

Substituting (35) 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

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

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 21 are satisfied, and we obtain the desired statement. For the second case, the conditions of Lemma 22 are easy to check (see the previous paragraph). The claim follows by this lemma. Finally, for the third claim, we invoke Lemma 23. □

G Superiority of Biased Compressors Under Statistical Assumptions

(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 several compressors under various synthetic and empirical statistical assumptions on the distribution of the entries of gradient vectors which need to be compressed. We quantify the gains of the Top-k sparsifier when compared against the unbiased Rand-k sparsifier, for example (see Section G).

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.

G.1 Top-k vs Rand-k

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

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

and

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

and the energy "saving" via

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

and

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

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. $x_1 = \cdots = x_d$. 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 show that in the case of uniform and exponentially distributed entries, the difference is significant.

Lemma 24. *Assume the coordinates of $x \in \mathbb{R}^d$ are i.i.d.*
(a) If they follow uniform distribution over $[0,1]$, then

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

(b) If they follow standard exponential distribution, then

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

Proof. **(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 $\mathrm{E}\left[\omega_{rnd}^k\right]$ for Rand-k is easy to compute as all coordinates are independent and uniformly distributed on $[0,1]$:

$$\mathrm{E}\left[x_i^2\right] \equiv \int_{[0,1]^d} x_i^2\, dx = \int_0^1 x_i^2\, dx_i = \frac{1}{3}, \tag{37}$$

which implies

$$\mathrm{E}\left[\omega_{rnd}^k(x)\right] = \left(1 - \frac{k}{d}\right)\sum_{i=1}^{d}\mathrm{E}\left[x_i^2\right] = \left(1 - \frac{k}{d}\right)\frac{d}{3} = \frac{d-k}{3}. \tag{38}$$

In order to compute the expected variance $\mathrm{E}\left[\omega_{top}^k\right]$ for Top-k, we use the following formula from order statistics[2] (see e.g. [33])

$$\mathrm{E}\left[x_{(i)}^2\right] \equiv \int_{[0,1]^d} x_{(i)}^2\, dx = \frac{\Gamma(i+2)\Gamma(d+1)}{\Gamma(i)\Gamma(d+3)} = \frac{i(i+1)}{(d+1)(d+2)}, \tag{39}$$

[2]see `https://en.wikipedia.org/wiki/Order_statistic`, `https://www.sciencedirect.com/science/article/pii/S0167715212001940`

from which we derive

$$\begin{aligned}\mathrm{E}\left[\omega_{top}^{k}\right] &= \sum_{i=1}^{d-k}\mathrm{E}\left[x_{(i)}^{2}\right] = \frac{1}{(d+1)(d+2)}\sum_{i=1}^{d-k} i(i+1) \\ &= \frac{1}{(d+1)(d+2)}\cdot\frac{(d-k)(d-k+1)(d-k+2)}{3} \\ &= \frac{d-k}{3}\left(1-\frac{k}{d+1}\right)\left(1-\frac{k}{d+2}\right).\end{aligned} \tag{40}$$

Combining (38) and (40) 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 (37) and (39) for $i=d$ and get

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

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 $\mathrm{E}\left[s_{rnd}^{1}\right]$ can be computed directly:

$$\mathrm{E}\left[s_{rnd}^{1}(x)\right] = \mathrm{E}\left[x_{d}^{2}\right] = \mathrm{Var}\left[x_{d}\right] + \mathrm{E}\left[x_{d}\right]^{2} = 2.$$

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

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

$$y_i \coloneqq (d-i+1)(x_{(i)}-x_{(i-1)}), \quad 1\le i\le d,$$

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

Proof. The joint density function of $x_{(1)},\ldots,x_{(d)}$ is given by (see [33])

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

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

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

with the transformation matrix

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

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

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

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

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

which means that variables $y_1, \ldots y_d$ 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

$$\mathrm{E}\left[x_{(d)}\right] = \sum_{i=1}^{d} \mathrm{E}\left[\frac{y_i}{d-i+1}\right] = \sum_{i=1}^{d} \frac{\mathrm{E}\left[y_i\right]}{d-i+1} = \sum_{i=1}^{d} \frac{1}{i},$$
$$\mathrm{Var}\left[x_{(d)}\right] = \sum_{i=1}^{d} \mathrm{Var}\left[\frac{y_i}{d-i+1}\right] = \sum_{i=1}^{d} \frac{\mathrm{Var}\left[y_i\right]}{(d-i+1)^2} = \sum_{i=1}^{d} \frac{1}{i^2},$$

from which we conclude the lemma as

$$\mathrm{E}\left[s^1_{top}(x)\right] = \mathrm{E}\left[x^2_{(d)}\right] = \mathrm{Var}\left[x_{(d)}\right] + \mathrm{E}\left[x_{(d)}\right]^2 = \sum_{i=1}^{d} \frac{1}{i^2} + \left(\sum_{i=1}^{d} \frac{1}{i}\right)^2 \approx \mathcal{O}(\log^2 d).$$

□