
Adaptive Compression for Communication-Efficient Distributed Training

Maksim Makarenko
KAUST

Elnur Gasanov
KAUST

Rustem Islamov
Institut Polytechnique de Paris

Abdurakhmon Sadiev
MIPT

Peter Richtárik
KAUST

Abstract

We propose Adaptive Compressed Gradient Descent (**AdaCGD**) – a novel optimization algorithm for communication-efficient training of supervised machine learning models with adaptive compression level. Our approach is inspired by the recently proposed three point compressor (**3PC**) framework of Richtárik et al. (2022), which includes error feedback (**EF21**), lazily aggregated gradient (**LAG**), and their combination as special cases, and offers the current state-of-the-art rates for these methods under weak assumptions. While the above mechanisms offer a fixed compression level, or adapt between two extremes only, our proposal is to perform a much finer adaptation. In particular, we allow the user to choose any number of arbitrarily chosen contractive compression mechanisms, such as Top- K sparsification with a user-defined selection of sparsification levels K , or quantization with a user-defined selection of quantization levels, or their combination. **AdaCGD** chooses the appropriate compressor and compression level adaptively during the optimization process. Besides i) proposing a theoretically-grounded multi-adaptive communication compression mechanism, we further ii) extend the 3PC framework to bidirectional compression, i.e., we allow the server to compress as well, and iii) provide sharp convergence bounds in the strongly convex, convex and nonconvex settings. The convex regime results are new even for several key special cases of our general mechanism, including **3PC** and **EF21**. In all regimes, our rates are superior compared to all existing adaptive compression methods.

1 Introduction

Training machine learning models is computationally expensive and time-consuming. In the recent years, researchers have tended to use increasing datasets, often distributed over several devices, throughout the training process in order to improve the effective generalization ability of contemporary machine learning frameworks (Vaswani et al., 2019). By word “device” or “node” we refer to any gadget that can store data, compute loss values and gradients (or stochastic gradients), and communicate with other different devices. For example, this distributed setting appears in *federated learning* (FL) (Konečný et al., 2016; McMahan et al., 2017; Lin et al., 2018). FL describes machine learning in a setting where a substantial number of strongly heterogeneous clients attempt to cooperatively train a model using the varied data stored on these devices without violating clients’ information privacy (Richtárik et al., 2022). In this setting, distributed methods can be very efficient (Goyal et al., 2017; You et al., 2019) and therefore federated frameworks have attracted tremendous attention in recent years.

Dealing with the distributed environment, we consider optimization problem of the form

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

Algorithm 1 DCGD method with master compression

```
1: Input: starting point  $x^0 \in \mathbb{R}^d$ ;  $g^0, \tilde{g}_i^0 \in \mathbb{R}^d$  for  $i = 1, \dots, n$  (known by nodes),  $\tilde{g}^0 = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^0$ 
   (known by master); learning rate  $\gamma > 0$ , worker compressor  $\mathcal{M}^W$ , master compressor  $\mathcal{M}^M$ .
2: for  $t = 0, 1, 2, \dots, T - 1$  do
3:   Server broadcasts  $g^t$  to all workers
4:   for all devices  $i = 1, \dots, n$  in parallel do
5:      $x^{t+1} = x^t - \gamma g^t$ 
6:      $\tilde{g}_i^{t+1} = \mathcal{M}^W(\nabla f_i(x^{t+1}))$ 
7:     Communicate  $\tilde{g}_i^{t+1}$  to the server
8:   end for
9:   Server aggregates received gradient estimators  $\tilde{g}^{t+1} = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^{t+1}$ 
10:   $g^{t+1} = \mathcal{M}^M(\tilde{g}^{t+1})$ 
11: end for
```

where $x \in \mathbb{R}^d$ is the parameter vector of training model, d is the dimensionality of the problem (number of trainable features), n is the number of workers/devices/nodes, and $f_i(x)$ is the loss incurred by model x on data stored on worker i . The loss function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ often has the form of expectation of some random function $f_i(x) := \mathbb{E}_{\xi \sim \mathcal{D}_i} [f_\xi(x)]$ with \mathcal{D}_i being the distribution of training data owned by worker i . In federated learning, these distributions are allowed to be different (so-called *heterogeneous* case). This finite sum function form allows us to capture the distributed nature of the problem in a very efficient way.

1.1 Communication-efficient distributed learning via gradient compression

The most effective models are frequently over-parameterized, which means that they contain more parameters than there are training data samples (Arora et al., 2018).

In this case, distributed methods may experience *communication bottleneck*: the situation when the communication cost for the workers to transfer sensitive information in joint optimization can exceed by multiple orders of magnitude the cost of local computation (Dutta et al., 2020). One of the practical methods to transfer information more efficiently is to apply a local compression operator (Seide et al., 2014; Suresh et al., 2017; Konečný & Richtárik, 2018) to the model’s parameters (gradients or tensors) needs to be communicated across different clients. The compression operator could be formalized as (possibly randomized) mapping $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, where d is the size of the tensor that has to be transmitted, with the feature that transmission of compressed tensor $\mathcal{C}(x)$ requires much less communication cost than the transfer of initial tensor x . While compression decreases the number of bits that are transferred during each communication cycle, it also brings in errors. As a result, the number of rounds necessary to obtain a solution with the appropriate accuracy typically increases. However, as the trade-off frequently appears to favor compression over no compression, compression has been proven to be effective in practice.

Distributed Compressed Gradient Descent (DCGD) (Khairat et al., 2018) provides the simplest and universal mechanism for distributed communication-efficient training with compression. With the given learning rate γ , DCGD implements the following update rule

$$x^{t+1} = x^t - \gamma \frac{1}{n} \sum_{i=1}^n g_i^t, \quad g_i^t = \mathcal{M}_i^t(\nabla f_i(x^t)). \quad (2)$$

Here, g_i^t represents an estimated gradient, result of mapping of original dense and high-dimensional gradient $\nabla f_i(x^t) \in \mathbb{R}^d$ into a vector of same size that can be transferred efficiently with far fewer bits via \mathcal{M}_i^t compression mechanism.

1.2 DCGD with bidirectional compression

In some cases (Tang et al., 2020; Philippenko & Dieuleveut, 2020; Fatkhullin et al., 2021) it is desirable to add compression on the server side to have efficient communication between server and clients in both directions. One could easily extend the general framework of DCGD to the case of bidirectional compression. If we define the general master compression mechanism as $\mathcal{M}^{M,t}$ and

worker compression mechanism as $\mathcal{M}_i^{W,t}$ we could formally write the general bidirectional DCGD algorithm as Algorithm 1.

2 Motivation and Background

The main motivation of this work is to generalize the ideas presented in (Richtárik et al., 2022) to allow compression level evolve during the optimization process based on some local information about client’s individual cost function.

2.1 Constant contractive compressors

The majority of methods employing gradient compression mechanisms use static compressors with constant compression level. In this approach (Richtárik et al., 2022), one sets

$$\mathcal{M}_i^t(x) \equiv \mathcal{C}(x), \quad (3)$$

where $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a (possibly randomized) operator. There are two large classes of operators (or compressors) that have been analyzed in the literature: i) *unbiased* compression operators and ii) *biased* or *contractive* compression operators. In this work we deal with *contractive* compressors only. Here we explicitly give the definition of this class.

Definition 1 (Biased or contractive compression operator). A mapping $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called *biased* or *contractive* compression operator there exists $0 < \alpha \leq 1$ such that

$$\mathbb{E} [\|\mathcal{C}(x) - x\|^2] \leq (1 - \alpha) \|x\|^2, \quad \forall x \in \mathbb{R}^d. \quad (4)$$

Rank- K (Khirirat et al., 2018) and Top- K (Alistarh et al., 2018) sparsification compressors are typical examples of contractive compressors. Due to the biased nature of these compressors, until recently, there was a gap between experimental and theoretical development of gradient descent methods based on contractive compressors. Thus, during the last years, algorithmic approaches have provided several methods of high practical importance, most notable of which is the so-called error feedback mechanism (Seide et al., 2014), fixing a divergence issue that appeared in practice. In contrast, in the theoretical development, until very recently, analytical studies offered weak sublinear (Stich et al., 2018; Karimireddy et al., 2019; Horváth & Richtárik, 2021) convergence rates under, in some cases, strong unrealistic assumptions (Richtárik et al., 2022). Recently, Richtárik et al. (2021) fixed this by providing a novel algorithmic and theoretical development that recovers GD $\mathcal{O}(1/T)$ rates, with the analysis using standard assumptions only. Fatkhullin et al. (2021) subsequently extended the EF21 framework by including several algorithmic and theoretical extensions, such as bidirectional compression and client stochasticity, which makes this method of high practical interest. Despite these advances, there are still many challenges in the theoretical understanding of these classes of methods. One of such challenges is a lack of precise theoretical study with the strong rates for error feedback methods in a convex setting.

2.2 Existing adaptive compressors

Using a static compression level of the compressor for all clients could limitate the optimization framework’s capability. Indeed, compression in the FL scenario can depend on the client it is applied on. For example, in heterogeneous hardware cases, *i.e.* when machines participating in collaborative training have very different computational capabilities, adjusting the compression level of a compressor for every client could help to reduce overall training time (Horváth et al., 2021; Abdelmoniem & Canini, 2021). Ideally, the optimizer should be able to define the particular compression level for each client separately based on the local information from the client.

Despite the significant practical interest in the development of such methods, there is currently very limited research and understanding of adaptive mechanisms of this type. Only a few papers (Qu et al., 2021; Höning et al., 2021; Mishchenko et al., 2022) provide any convergence guarantees with explicit rates, and most of them are designed for the specific type of compressors only, mostly quantizers. So, in (Jhunjunwala et al., 2021), the authors design a mechanism for adaptive change of quantization level $s^k \sim \sqrt{\frac{f(x^0)}{f(x^k)}}$ of a random dithering operator (Alistarh et al., 2017). DAdaQuant (Höning et al., 2021) and FedDQ (Qu et al., 2021) suggest ascending and descending quantizations throughout

Table 1 Summary of adaptive compressed methods. n is a number of workers, L and μ are smoothness and strong convexity constants respectively of $f_i \forall i \in \{1, \dots, n\}$, $\kappa = \frac{L}{\mu}$ is a condition number, C_i s are constants, $\Delta_x = \|x^0 - x^*\|^2$, $\Delta_f = f(x_0) - f^*$, $M_1 = \max\{L_- + L_+ \sqrt{\frac{2B_{\max}}{A_{\min}}}, \frac{1}{A_{\min}}\}$, $M_2 = L_- + L_+ \sqrt{\frac{B_{\max}}{A_{\min}}}$ (see Lemma 2). str cvx = strongly convex, cvx = convex, noncvx = nonconvex.

Paper	Any \mathcal{C} ?	Theory?	Str cvx / PL noncvx rate	Cvx rate	General noncvx rate
Jhunjunwala et al. (2021)	✗	✗	✗	✗	✗
Abdelmoniem & Canini (2021)	✗	✗	✗	✗	✗
Hönig et al. (2021)	✗	✓ ⁽¹⁾	$\frac{\max\{\kappa, \frac{\kappa^2}{n}, \frac{n}{\mu^2}\}}{T^2}$	✗	$\mathcal{O}(\frac{L\Delta_f}{\sqrt{T}} + \frac{C_1}{T})$
Qu et al. (2021)	✗	✓	✗	✗	$\mathcal{O}(\frac{L\Delta_f}{\sqrt{T}})$ ⁽²⁾
Zhao et al. (2022)	✗	✓	linear ⁽³⁾	✗	✗
Mao et al. (2021)	✗	✓	linear ⁽³⁾	✗	✗
Khairat et al. (2021)	✓	✗	✗	✗	✗
Mishchenko et al. (2022)	✗	✓ ⁽⁴⁾	✗	$\mathcal{O}(\frac{L\Delta_x}{T} + \frac{\sigma_*^2 + \varepsilon}{Ln})$	$\mathcal{O}(\frac{L\Delta_f}{T} + \frac{\varepsilon}{Ln})$
THIS WORK	✓	✓	$(1 - \min\{\frac{\mu}{M_2}, A_{\min}\})^T$	$\mathcal{O}(\frac{M_1}{T})$	$\mathcal{O}(\frac{2\Delta_f M_2 + C_3}{T A_{\min}})$

⁽¹⁾ The rates, as stated in the paper, are derived from Reisizadeh et al. (2020). We consider non-local full gradient setup, i.e. $\sigma^2 = 0$ and $\tau = 1$.

⁽²⁾ We show the rate for non-local full gradient setup, i.e. $\sigma^2 = 0$ and $\tau = 1$.

⁽³⁾ Their work does not present any *explicit* rate.

⁽⁴⁾ $\varepsilon > 0$ is a parameter of IntSGD algorithm.

the training. AQUILA (Zhao et al., 2022) and AGQ (Mao et al., 2021) build an adaptive quantization on top of the Lazily Aggregated Quantized (LAQ) gradient algorithm (Sun et al., 2019). IntSGD (Mishchenko et al., 2022) adaptively sets the scaling parameter α^k of a vector plugged in a randomized integer rounding operator. CAT S+Q (Khairat et al., 2021) proposes an adaptive way to choose k : the top- k elements of the gradient at iteration i , only *signs* of which clients send to the server along with the gradient norm. Table 1 provides a detailed comparison of these works.

2.3 Adaptive compression via selective (*lazy*) aggregation

The LAG mechanism proposed by Chen et al. (2018) is an alternative way to embed adaptivity into the framework by introducing communication "skipping". According to the lazy aggregation communication mechanism, each worker i only shares its local gradient if it is significantly different from the last gradient shared previously. Otherwise, the worker decides to "skip" the communication round. In some sense, it is an adaptive mechanism that chooses between two extremes for each client: sending a full gradient or skipping the communication round.

Richtárik et al. (2022) recently generalized this idea by introducing CLAG, which connects particular contractive \mathcal{C} compressor with a pre-defined compression level with LAG mechanism. In Richtárik et al.'s CLAG method all n workers send the compressed gradient $g_i^0 = \mathcal{C}(\nabla f_i(x^0))$ for all $i \in [n]$, at the beginning of the training. The workers $i \in [n]$ define g_i^{t+1} , which can be viewed as a shifted and compressed version of the client's gradient $\nabla f_i(x^{t+1})$ using the *lazy aggregation rule* combined with EF21 shift

$$g_i^{t+1} = \begin{cases} g_i^t + \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t), & \text{if } \|\nabla f_i(x^{t+1}) - g_i^t\|^2 > \zeta D_i^t, \\ g_i^t, & \text{otherwise} \end{cases} \quad (5)$$

where $D_i^t := \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2$ and $\zeta > 0$ is the *trigger*. Trigger parameter ζ controls how frequently trigger condition will be satisfied and how often clients skip communication rounds. CLAG method includes LAG as a special case with \mathcal{C} compressor being identity operator (*no compression*). This approach also could be seen as adaptive which interpolates between two extremes: compressed gradient with pre-defined compression level or entirely skipping communication.

Although both LAG and CLAG perform well in practice, their fixed and limited compression levels could restrict their performance and make these methods sub-optimal. It is of particular practical interest to create a more general method with evolving fine-tuned compression level individual for every client. From the perspective of the convergence theory, one of the issues *lazy* methods have is

Table 2 Comparison of available convergence guarantee results of methods employing lazy aggregation.

Method	Adaptive compression?	Bidirectional case	Str cvx case	Cvx case	PL noncvx case	General noncvx case
LAQ (Sun et al., 2019)	✗	✗	✓	✗	✗	✗
LENA (Ghadikolaei et al., 2021)	✗	✗	✓	✓	✓	✓
LAG (Richtárik et al., 2022)	✗	✗	✓	✗	✓	✓
CLAG (Richtárik et al., 2022)	✗	✗	✓	✗	✓	✓
AdaCGD (NEW, 2022)	✓	✓	✓	✓	✓	✓

the difficulty of determining how often communication skips occur because the trigger is conditional. Thus, there are no theoretical studies examining the frequency of communication skips.

3 Summary of Contributions

We highlight our main contributions as follows:

- **New class of adaptive compressors.** In Richtárik et al. (2022), the authors propose the different classes of compressors unified through a single theory. Despite the large variability of the compression mechanisms, including the algorithms with *lazy* aggregation rule, the compression level in all of the considered algorithms is pre-defined before and kept constant during the training. In this work, we take a step further and formulate an extended class of an adaptive 3PC compressors (*ada3PC*) with tunable compression levels defined by some general trigger rules. As an original 3PC compressors, this class of compressors are very general and includes a number of specific compressors such as *AdaCGD* which includes *EF21* and *CLAG* as special cases. This method is applicable for a large class of biased compressors, such as Top- K and Rank- K sparsification.

- **Convergence guarantees with strong rates unified by a single 3PC theory.** We provide a strong convergence bound for strongly convex, convex, and non-convex settings. Comparing with the adaptive methods outside the 3PC context, we provide a more elaborate theory with better convergence rates. For *AdaCGD* we recover $\mathcal{O}(1/T)$ rate of *GD* up to a certain constant in general non-convex case. It is superior in comparison with $\mathcal{O}(1/\sqrt{T})$ rate (Hönig et al., 2021; Qu et al., 2021) for SOTA in adaptive compression outside 3PC context. The convergence theory in a convex set is of particular interest since due to its novelty even in the case of 3PC for some key cases of *AdaCGD*, such as *EF21* and *CLAG*. In other words, it is a new SOTA result for the error-feedback method in the convex setting.

- **Extension of 3PC theory with bidirectional compression.** We extend 3PC methods with bidirectional compression i.e., we allow the server to compress as well.

Table 2 compares *AdaCGD* with other described in the literature lazy algorithms.

4 Ada3PC: A Compression-Adaptive 3PC Method

4.1 3PC compressor

Richtárik et al. (2022) introduces the general concept of three point compressors. Here we provide its formal definition for consistency:

Definition 2. We say that a (possibly randomized) map $\mathcal{C}_{h,y}(x) : \underbrace{\mathbb{R}^d}_{h \in} \times \underbrace{\mathbb{R}^d}_{y \in} \times \underbrace{\mathbb{R}^d}_{x \in} \rightarrow \mathbb{R}^d$ is a three point compressor (3PC) if there exist constants $0 < A \leq 1$ and $B \geq 0$ such that the following relation holds for all $x, y, h \in \mathbb{R}^d$

$$\mathbb{E} [\|\mathcal{C}_{h,y}(x) - x\|^2] \leq (1 - A)\|h - y\|^2 + B\|x - y\|^2. \quad (6)$$

Authors show that *EF21* compression mechanism satisfies Definition 2. Let $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a contractive compressor with contraction parameter α , and define

$$\mathcal{C}_{h,y}^{\text{EF}}(x) := h + \mathcal{C}(x - h). \quad (7)$$

If we use this mapping to define a compression mechanism \mathcal{M}_i^t via (2) within **DCGD**, we obtain the **EF21** method.

Another variant of **3PC** compressors introduced in (Richtárik et al., 2022) is **CLAG** compressor. Let $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a contractive compressor with contraction parameter α . Choosing a trigger $\zeta > 0$, authors define the **CLAG** rule

$$\mathcal{C}_{h,y}^{\text{CL}}(x) := \begin{cases} h + \mathcal{C}(x - h), & \text{if } \|x - h\|^2 > \zeta \|x - y\|^2, \\ h, & \text{otherwise,} \end{cases} \quad (8)$$

If we employ this mapping into **DCGD** method (2) as communication mechanism \mathcal{M}_i^t , we obtain **CLAG**. It includes **LAG** compressor \mathcal{C}^L as a special case when compressor \mathcal{C} is identity.

4.2 Adaptive 3PC compressor

We are now ready to introduce the Adaptive Three Point (**Ada3PC**) Compressor.

Definition 3 (**Ada3PC** compressor). Let $\mathcal{C}^1, \mathcal{C}^2, \dots, \mathcal{C}^m$ be 3PC compressors: $\mathcal{C}^i : \mathbb{R}^{3d} \rightarrow \mathbb{R}^d$ for all i . Let P_1, P_2, \dots, P_{m-1} be conditions depending on h, y, x , i.e. $P_j : \underbrace{\mathbb{R}^d}_{h \in} \times \underbrace{\mathbb{R}^d}_{y \in} \times \underbrace{\mathbb{R}^d}_{x \in} \rightarrow \{0, 1\}$

for all j . Then, the adaptive 3PC compressor, associated with above 3PC compressors and conditions, is defined as follows:

$$\mathcal{C}_{h,y}^{\text{ad}}(x) = \begin{cases} \mathcal{C}_{h,y}^1(x) & \text{if } P_1(h, y, x), \\ \mathcal{C}_{h,y}^2(x) & \text{else if } P_2(h, y, x), \\ \dots, \\ \mathcal{C}_{h,y}^{m-1}(x) & \text{else if } P_{m-1}(h, y, x), \\ \mathcal{C}_{h,y}^m(x) & \text{otherwise.} \end{cases} \quad (9)$$

Ada3PC compressor first finds the smallest index j for which $P_j(h, y, x) = 1$ (if such index does not exist, we set $j = m$). Then, **Ada3PC** applies \mathcal{C}^j compressor on vector x .

4.3 Adaptive Compressed Gradient Descent

There are many ways how to define meaningful and practical compressors in the context of the adaptive **3PC** framework. Here we provide one particular, perhaps the simplest scheme, which we define as **AdaCGD**. In this scheme we have a set of m **EF21** compressors $\{\mathcal{C}_{h,y}^{\text{EF},j}(x)\}_{j \in 1 \dots m}$ sorted in order from the highest compression level to the lowest, i.e. $\alpha_1 \leq \alpha_2 \dots \leq \alpha_m$, where α_j is a corresponding contractive parameter of the j -th compressor. For example, if we use Top- K in $\mathcal{C}_{h,y}^{\text{EF}}$ compressors, first and last indices correspond to the compressors with the smallest and the largest K , respectively. We choose the first compressor, i.e. with the strongest compression, which satisfies a trigger rule. We design the j -th trigger rule following an intuition of *lazy aggregation* rule:

$$P_j := \|x - \mathcal{C}_{h,y}^{\text{EF},j}(x)\|^2 \leq \zeta \|x - y\|^2. \quad (10)$$

As in the original **LAG** rule, the left side of (10) presents the difference between the true gradient and its estimate, while the right side compares gradient differences on neighboring iterations. The key difference of (10) trigger from **LAG** and **CLAG** rule (5) is that the left side of this trigger condition depends explicitly from the level of compression. This feature is essential as it enables the desired rule-based compressor selection. Altogether, we can define **AdaCGD** compressor formally.

Definition 4 (**AdaCGD** compressor). Given the set of m **EF21** compressors $\{\mathcal{C}_{h,y}^{\text{EF},j}(x)\}_{j \in 1 \dots m}$, sorted in descending order of compression level and $\zeta \geq 0$, adaptive **AdaCGD** compressor is defined with a switch condition as follows:

$$\mathcal{C}_{h,y}^{\text{AC}}(x) = \begin{cases} h & \text{if } \|x - h\|^2 \leq \zeta \|x - y\|^2, \\ \mathcal{C}_{h,y}^{\text{EF},1}(x) & \text{else if } \|x - \mathcal{C}_{h,y}^{\text{EF},1}(x)\|^2 \leq \zeta \|x - y\|^2, \\ \dots, \\ \mathcal{C}_{h,y}^{\text{EF},m-1}(x) & \text{else if } \|x - \mathcal{C}_{h,y}^{\text{EF},m-1}(x)\|^2 \leq \zeta \|x - y\|^2, \\ \mathcal{C}_{h,y}^{\text{EF},m}(x) & \text{otherwise.} \end{cases} \quad (11)$$

If $C_{h,y}^{\text{EF},m}$ uses Top- d compression, i.e., identity operator, **AdaCGD** is an adaptive compressor composed of the whole spectrum of compressors from full compression, i.e., communication "skip", to zero compression, i.e., sending full gradient. Since standalone "skip" connection is clearly not **3PC** operator, it may not be obvious that **AdaCGD** is a special case of **Ada3PC**. For this reason, here we provide the following lemma:

Lemma 1. *AdaCGD is a special case of Ada3PC compressor.*

It is easy to see that if $\zeta = 0$ **AdaCGD** reduces to **EF21**. Similarly, **CLAG** is a special case of **AdaCGD** when $m = 1$.

5 Theory

In this section, we present theoretical convergence guarantees for Algorithm 1 with **Ada3PC** compressors in two new cases presented in Table 2. The results for general and PL nonconvex cases can be found in the appendix.

5.1 Assumptions

To get convergence rates of Algorithm 1, we rely on standard assumptions.

Assumption 1. *The functions $f_1, \dots, f_n : \mathbb{R}^d \rightarrow \mathbb{R}$ are convex, i.e.*

$$f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^d, \forall i. \quad (12)$$

Assumption 2. *The function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L_- -smooth, i.e.*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_- \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \quad (13)$$

Assumption 3. *There exists $L_+ > 0$ such that $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_+^2 \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$. Let L_+ be the smallest such number.*

We borrow $\{L_-, L_+\}$ notation from (Szlendak et al., 2022). Assumption 3 avoids a stronger assumption on Lipschitz smoothness of individual functions f_i . Moreover, one can easily prove that $L_- \leq L_+$.

The next assumption is standard for analysis of practical methods (Ahn et al., 2020), Rajput et al. (2020). However, compared to previous works, we require a more general version.

Assumption 4. *We assume that there exists a constant $\Omega > 0$ such that $\mathbb{E}[\|x^t - x^*\|^2] \leq \Omega^2$, where x^t is any iterate generated by Algorithm 1.*

Assumption 5. *The functions f_1, \dots, f_n are differentiable. Moreover, f is bounded from below by an infimum $f^{\text{inf}} \in \mathbb{R}$, i.e. $f(x) \geq f^{\text{inf}}$ for all $x \in \mathbb{R}^d$.*

5.2 Adaptive 3PC is a 3PC compressor

While this may not be obvious at first glance, Adaptive 3PC compressor itself belongs to the class of 3PC compressors. We formalize this statement in the following lemma.

Lemma 2. *Let C^{ad} be an adaptive 3PC compressor. Let $\{C^i\}_{i=1}^m$ be 3PC compressors associated with C^{ad} , i.e. for all i there exists constants $0 < A_i \leq 1$ and $B_i \geq 0$, such that for all $h, y, x \in \mathbb{R}^d$*

$$\mathbb{E}\|C_{h,y}^i(x) - x\|^2 \leq (1 - A_i)\|h - y\|^2 + B_i\|x - y\|^2.$$

Then, C^{ad} is a 3PC compressor with constants $A_{\min} := \min\{A_1, \dots, A_m\}$ and $B_{\max} := \max\{B_1, \dots, B_m\}$.

Proof. Let us fix $h, y, x \in \mathbb{R}^d$ and let j be the index, such that $P_i(h, y, x) = 0$ for all $i < j$ and, if $j < m$, $P_j(h, y, x) = 1$. Then,

$$\begin{aligned} \mathbb{E}\|C_{h,y}^{\text{ad}}(x) - x\|^2 &= \mathbb{E}\|C_{h,y}^j(x) - x\|^2 \stackrel{(6)}{\leq} (1 - A_j)\|h - y\|^2 + B_j\|x - y\|^2 \\ &\leq (1 - A_{\min})\|h - y\|^2 + B_{\max}\|x - y\|^2. \end{aligned}$$

□

In the definition of **Ada3PC** compressor, we never specify what conditions P_i s are. In fact, they are completely arbitrary! This enables us to build infinitely many new compressors out of few notable examples, presented in (Richtárik et al., 2022).

5.3 Convergence

In this subsection, we show how assumptions we make about minimized functions and compressors affect the convergence rate of Algorithm 1.

Convergence for convex functions. The first result assumes that \mathcal{M}^W in Algorithm 1 is a **3PC** compressor, \mathcal{M}^M is an identity compressor: $\mathcal{M}^M(x) = x \forall x \in \mathbb{R}^d$.

Theorem 5. *Let Assumptions 1, 2, 3 and 4 hold. In Algorithm 1, assume \mathcal{M}^W is a **3PC** compressor, \mathcal{M}^M is an identity compressor, and the stepsize γ satisfies $0 < \gamma \leq 1/M$, where $M = L_- + L_+ \sqrt{\frac{2B}{A}}$. Then, for any $T \geq 1$ we have*

$$\mathbb{E} [f(x^T)] - f(x^*) \leq \max \left\{ \frac{1}{\gamma}, \frac{1}{A} \right\} \frac{2(\Omega^2 + \Phi^0)}{T},$$

where $\Phi^t := f(x^t) - f(x^*) + \frac{\gamma}{A} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^t) - g_i^t\|^2$ for any $t \geq 0$.

The theorem combined with Lemma 2 implies the following fact.

Corollary 1. *Let the assumptions of Theorem 5 hold, assume \mathcal{M}^W is an **Ada3PC** compressor, \mathcal{M}^M is an identity compressor, and choose the stepsize $\gamma = \frac{1}{L_- + L_+ \sqrt{\frac{2B_{\max}}{A_{\min}}}}$. Then, for any $T \geq 1$ we have*

$$\mathbb{E} [f(x^T)] - f(x^*) \leq \max \left\{ L_- + L_+ \sqrt{\frac{2B_{\max}}{A_{\min}}}, \frac{1}{A_{\min}} \right\} \frac{2(\Omega^2 + \Phi^0)}{T}.$$

Thus, to achieve $\mathbb{E} [f(x^T)] - f(x^*) \leq \varepsilon$ for some $\varepsilon > 0$, the **Ada3PC** method requires

$$T = \mathcal{O} \left(\max \left\{ L_- + L_+ \sqrt{\frac{2B_{\max}}{A_{\min}}}, \frac{1}{A_{\min}} \right\} \frac{2(\Omega^2 + \Phi_0^2)}{\varepsilon} \right)$$

iterations.

Convergence for bidirectional method. Here, we analyze the case when meaningful compressors applied on both communication directions, i.e., both \mathcal{M}^M and \mathcal{M}^W are **3PC** compressors.

Theorem 6. *Let Assumptions 3 and 5 hold. Let \mathcal{M}^M and \mathcal{M}^W be **3PC** compressors and the stepsize in Algorithm 1 be set as*

$$0 < \gamma \leq \frac{1}{L_- + L_+ \sqrt{\frac{6B^M(B^W+1)}{A^M} + \frac{2B^W}{A^W} \left(1 + \frac{3B^M(2-A^W)}{A^M}\right)}}. \quad (14)$$

Fix T and let \hat{x}^T be chosen uniformly from $\{x^0, x^1, \dots, x^{T-1}\}$ uniformly at random. Then

$$\mathbb{E} \left[\|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Psi^0}{\gamma T}. \quad (15)$$

where $\Psi^t = f(x^t) - f^{\inf} + \frac{\gamma}{A^M} \|g^t - \tilde{g}^t\|^2 + \frac{\gamma}{A^W} \left(1 + \frac{3B^M(2-A^W)}{A^M}\right) \frac{1}{n} \sum_{i=1}^n \|\tilde{g}_i^t - \nabla f_i(x^t)\|^2$ for any $t \geq 0$.

In the theorem, superscripts ‘‘M’’ and ‘‘W’’ denote master and worker compressor parameters, respectively. Theorem 6 implies the following corollary.

Corollary 2. *Let the assumptions of Theorem 6 hold, assume \mathcal{M}^M and \mathcal{M}^W are **Ada3PC** compressors and the stepsize*

$$\gamma = \frac{1}{L_- + L_+ \sqrt{\frac{6B_{\max}^M(B_{\max}^W+1)}{A_{\min}^M} + \frac{2B_{\max}^W}{A_{\min}^W} \left(1 + \frac{3B_{\max}^M(2-A_{\min}^W)}{A_{\min}^M}\right)}}.$$

Fix T and let \hat{x}^T be chosen uniformly from $\{x^0, x^1, \dots, x^{T-1}\}$ uniformly at random. Then, we have

$$\mathbb{E} \left[\|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Psi^0 \left(L_- + L_+ \sqrt{\frac{6B_{\max}^M(B_{\max}^W+1)}{A_{\min}^M} + \frac{2B_{\max}^W}{A_{\min}^W} \left(1 + \frac{3B_{\max}^M(2-A_{\min}^W)}{A_{\min}^M}\right)} \right)}{T}.$$

Thus, to achieve $\mathbb{E} [\|\nabla f(\hat{x}^T)\|^2] \leq \varepsilon^2$ for some $\varepsilon > 0$, Algorithm 1 requires

$$T = \mathcal{O} \left(\frac{2\Psi^0 \left(L_- + L_+ \sqrt{\frac{6B_{\max}^M (B_{\max}^W + 1)}{A_{\min}^M} + \frac{2B_{\max}^W}{A_{\min}^M} \left(1 + \frac{3B_{\max}^M (2 - A_{\min}^W)}{A_{\min}^M} \right)} \right)}{T} \right)$$

iterations.

6 Experiments

In this work we use the similar setup described in (Richtárik et al., 2022), namely we aim to solve logistic regression problem with non-convex regularizer:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{N} \sum_{i=1}^N \log(1 + e^{-y_i a_i^\top x}) + \lambda \sum_{j=1}^d \frac{x_j^2}{1+x_j^2} \right],$$

where $a_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ are the training data and labels, and $\lambda > 0$ is a regularization parameter, which is fixed to $\lambda = 0.1$. In training we use LIBSVM Chang & Lin (2011) datasets *phishing*, *ala*, *a9a*. Each dataset has been split into $n = 20$ equal parts, each representing a different client.

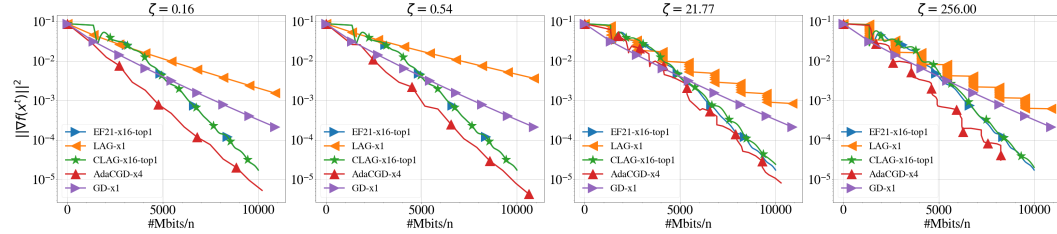


Figure 1: Comparison of LAG, CLAG, EF21 and GD with AdaCGD on phishing dataset. $1 \times, 2 \times, 4 \times$ (and so on) indicates the multiplication factor we use for the optimal stepsizes predicted by theory.

Figure 1 compares AdaCGD with other 3PC methods. We fine-tune the stepsize of each considered algorithm with $(2^0, 2^1, \dots, 2^8)$ multiples of corresponding theoretical stepsize. As contractive compressor we use Top- k operator. For EF21 and CLAG we use top-1 compressor, which usually the best in practice for these methods. In the experiments, AdaCGD is shown to be comparable and in some cases superior to CLAG and always superior to LAG. In other words, AdaCGD efficiently complements CLAG and other 3PC methods. Additional experiments and details are available in the appendix.

7 Discussion and Limitations

The main limitations of the work are assumptions we make upon functions f_i of the problem 1. But, on the other hand, these assumptions govern the convergence rates we report: for example, we cannot show linear rate for convex functions due to the fundamental lower bound (Nesterov et al., 2018).

Another limitation comes from the analysis of Bidirectional 3PC algorithm (Theorem 6). We show the analysis only for general nonconvex functions.

References

- Abdelmoniem, A. M. and Canini, M. Towards Mitigating Device Heterogeneity in Federated Learning via Adaptive Model Quantization. In *Proceedings of EuroMLSys'21*, Apr 2021. (Cited on 3, 4)
- Ahn, K., Yun, C., and Sra, S. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33:17526–17535, 2020. (Cited on 7)

- Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1709–1720, 2017. (Cited on 3)
- Alistarh, D., Hoefler, T., Johansson, M., Khirirat, S., Konstantinov, N., and Renggli, C. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (Cited on 3)
- Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. (Cited on 2)
- Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011. (Cited on 9, 22)
- Chen, T., Giannakis, G., Sun, T., and Yin, W. LAG: Lazily aggregated gradient for communication-efficient distributed learning. *Advances in Neural Information Processing Systems*, 2018. (Cited on 4)
- Dutta, A., Bergou, E. H., Abdelmoniem, A. M., Ho, C.-Y., Sahu, A. N., Canini, M., and Kalnis, P. On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3817–3824, 2020. (Cited on 2)
- Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. Ef21 with bells & whistles: practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021. (Cited on 2, 3)
- Ghadikolaie, H. S., Stich, S., and Jaggi, M. LENA: Communication-efficient distributed learning with self-triggered gradient uploads. In *International Conference on Artificial Intelligence and Statistics*, pp. 3943–3951. PMLR, 2021. (Cited on 5)
- Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. (Cited on 1)
- Hönig, R., Zhao, Y., and Mullins, R. D. Dadaquant: Doubly-adaptive quantization for communication-efficient federated learning. *CoRR*, abs/2111.00465, 2021. URL <https://arxiv.org/abs/2111.00465>. (Cited on 3, 4, 5)
- Horváth, S. and Richtárik, P. A better alternative to error feedback for communication-efficient distributed learning. In *9th International Conference on Learning Representations (ICLR)*, 2021. (Cited on 3)
- Horváth, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., and Lane, N. D. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4fLr7H5D_eT. (Cited on 3)
- Jhunjhunwala, D., Gadhikar, A., Joshi, G., and Eldar, Y. C. Adaptive quantization of model updates for communication-efficient federated learning. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3110–3114, 2021. doi: 10.1109/ICASSP39728.2021.9413697. (Cited on 3, 4)
- Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes SignSGD and other gradient compression schemes. In *36th International Conference on Machine Learning (ICML)*, 2019. (Cited on 3)
- Khairat, S., Feyzmahdavian, H. R., and Johansson, M. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018. (Cited on 2, 3)

- Khairat, S., Magnússon, S., Aytekin, A., and Johansson, M. A flexible framework for communication-efficient machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 8101–8109, May 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16987>. (Cited on 4)
- Konečný, J. and Richtárik, P. Randomized distributed mean estimation: Accuracy vs. communication. *Frontiers in Applied Mathematics and Statistics*, 4:62, 2018. (Cited on 2)
- Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. (Cited on 1)
- Li, Z., Bao, H., Zhang, X., and Richtárik, P. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2008.10898. (Cited on 15)
- Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, B. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018. (Cited on 1)
- Mao, Y., Zhao, Z., Yan, G., Liu, Y., Lan, T., Song, L., and Ding, W. Communication efficient federated learning with adaptive quantization. *CoRR*, abs/2104.06023, 2021. URL <https://arxiv.org/abs/2104.06023>. (Cited on 4)
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. (Cited on 1)
- Mishchenko, K., Wang, B., Kovalev, D., and Richtárik, P. IntSGD: Adaptive floatless compression of stochastic gradients. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=pFyXqxChZc>. (Cited on 3, 4)
- Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on 9, 23)
- Philippenko, C. and Dieuleveut, A. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2020. (Cited on 2)
- Pishro-Nik, H. *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014. ISBN 9780990637202. URL https://books.google.com.sa/books?id=3yq_0QEACAAJ. (Cited on 14)
- Qu, L., Song, S., and Tsui, C. Feddq: Communication-efficient federated learning with descending quantization. *CoRR*, abs/2110.02291, 2021. URL <https://arxiv.org/abs/2110.02291>. (Cited on 3, 4, 5)
- Rajput, S., Gupta, A., and Papailiopoulos, D. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, 2020. (Cited on 7)
- Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2021–2031. PMLR, 26–28 Aug 2020. URL <https://proceedings.mlr.press/v108/reisizadeh20a.html>. (Cited on 4)
- Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *arXiv preprint arXiv:2106.05203*, 2021. (Cited on 3, 14)
- Richtárik, P., Sokolov, I., Fatkhullin, I., Gasanov, E., Li, Z., and Gorbunov, E. A. 3pc: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. *CoRR*, abs/2202.00998, 2022. URL <https://arxiv.org/abs/2202.00998>. (Cited on 1, 3, 4, 5, 6, 8, 9, 15, 21, 22)

- Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. (Cited on 2, 3)
- Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (Cited on 3)
- Sun, J., Chen, T., Giannakis, G., and Yang, Z. Communication-efficient distributed learning via lazily aggregated quantized gradients. *Advances in Neural Information Processing Systems*, 32: 3370–3380, 2019. (Cited on 4, 5)
- Suresh, A. T., Felix, X. Y., Kumar, S., and McMahan, H. B. Distributed mean estimation with limited communication. In *International Conference on Machine Learning*, pp. 3329–3337. PMLR, 2017. (Cited on 2)
- Szlendak, R., Tyurin, A., and Richtárik, P. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=GugZ5DzzAu>. (Cited on 7)
- Tang, H., Lian, X., Yu, C., Zhang, T., and Liu, J. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2020. (Cited on 2)
- Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR, 2019. (Cited on 1)
- You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. (Cited on 1)
- Zhao, Z., Mao, Y., Zeeshan, M., Liu, Y., Lan, T., and Ding, W. AQUILA: Communication efficient federated learning with adaptive quantization of lazily-aggregated gradients, 2022. URL <https://openreview.net/forum?id=cdZLe5S0ur>. (Cited on 4)

APPENDIX

In Appendix A, we state the basic facts needed for detailed proofs of the propositions. In Appendix B, we provide the proofs missing in the main part of the paper. Appendix C contains experimental details and extra experiments. We briefly discuss the main limitations of the paper in Appendix D.

Contents

1	Introduction	1
1.1	Communication-efficient distributed learning via gradient compression	2
1.2	DCGD with bidirectional compression	2
2	Motivation and Background	3
2.1	Constant contractive compressors	3
2.2	Existing adaptive compressors	3
2.3	Adaptive compression via selective (<i>lazy</i>) aggregation	4
3	Summary of Contributions	5
4	Ada3PC: A Compression-Adaptive 3PC Method	5
4.1	3PC compressor	5
4.2	Adaptive 3PC compressor	6
4.3	Adaptive Compressed Gradient Descent	6
5	Theory	7
5.1	Assumptions	7
5.2	Adaptive 3PC is a 3PC compressor	7
5.3	Convergence	8
6	Experiments	9
7	Discussion and Limitations	9
A	Basic facts	14
B	Proofs for Sections 4 and 5	15
B.1	Lemma 1	15
B.2	Theorem 5	15
B.3	Theorem 6	17
B.4	Convergence for general nonconvex functions	21
B.5	Convergence for μ -nonconvex functions	21
C	Experimental details and extra experiments	22
D	Limitations	23

A Basic facts

We start the appendix with common math facts. Lemmas 3 and 4 present classic Cauchy-Schwartz inequality for vectors in metric space and random variables in probabilistic space, respectively. Lemma 5 shows a classic upper bound on quadratics. Lemma 6 provides a sufficient condition that ensures a quadratic inequality appearing in our convergence proofs holds.

Lemma 3 (Cauchy-Schwartz inequality for arbitrary vectors). *Let $x, y \in \mathbb{R}^d$ be arbitrary vectors. Then, the following inequality holds*

$$|\langle x, y \rangle| \leq \|x\| \|y\|, \quad (16)$$

where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and the induced norm, respectively.

Lemma 4 (Cauchy-Schwartz inequality for random variables; section 6.2.4 of (Pishro-Nik, 2014)). *For any two random variables X and Y , we have*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}, \quad (17)$$

where equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in \mathbb{R}$.

Lemma 5. *Let $a, b, c, d \in \mathbb{R}^d$ be arbitrary vectors. Then, the following inequalities hold*

$$\|a - b\|^2 \leq 2(\|a - c\|^2 + \|c - b\|^2), \quad (18)$$

$$\|a - b\|^2 \leq 3(\|a - c\|^2 + \|c - d\|^2 + \|d - b\|^2). \quad (19)$$

Lemma 6 (Lemma 5 of (Richtárik et al., 2021)). *If $0 < \gamma \leq \frac{1}{\sqrt{a+b}}$, then $a\gamma^2 + b\gamma \leq 1$. Moreover, the bound is tight up to the factor of 2 since $\frac{1}{\sqrt{a+b}} \leq \min\{\frac{1}{\sqrt{a}}, \frac{1}{b}\} \leq \frac{2}{\sqrt{a+b}}$.*

B Proofs for Sections 4 and 5

B.1 Lemma 1

At first glance, **AdaCGD** does not seem to be an **Ada3PC** compressor. However, we can construct an **Ada3PC** compressor, which is equivalent to **AdaCGD**.

Lemma 1. *AdaCGD is a special case of Ada3PC compressor.*

Proof. Let us consider the following **Ada3PC** compressor constructed from one **LAG** compressor and m **EF21** compressors.

$$\mathcal{C}_{h,y}(x) = \begin{cases} \mathcal{C}_{h,y}^{\text{LAG}} = \begin{cases} h & \text{if } \|x - h\|^2 \leq \zeta \|x - y\|^2, \\ x & \text{otherwise.} \end{cases} & \text{if } \|x - h\|^2 \leq \zeta \|x - y\|^2, \\ \mathcal{C}_{h,y}^{\text{EF},1}(x) & \text{else if } \|x - \mathcal{C}_{h,y}^{\text{EF},1}(x)\|^2 \leq \zeta \|x - y\|^2, \\ \dots, \\ \mathcal{C}_{h,y}^{\text{EF},m-1}(x) & \text{else if } \|x - \mathcal{C}_{h,y}^{\text{EF},m-1}(x)\|^2 \leq \zeta \|x - y\|^2, \\ \mathcal{C}_{h,y}^{\text{EF},m}(x) & \text{otherwise.} \end{cases}$$

If $\|x - h\|^2 \leq \zeta \|x - y\|^2$, then $\mathcal{C}_{h,y}$ applies the **LAG** compressor to x . This **LAG** compressor in turn outputs h , as it does $\mathcal{C}_{h,y}^{\text{AC}}$ for the same condition. If the opposite is true, *i.e.*, $\|x - h\|^2 > \zeta \|x - y\|^2$, $\mathcal{C}_{h,y}$ checks the same conditions and chooses the same compressor as $\mathcal{C}_{h,y}^{\text{AC}}$. Thus, $\mathcal{C}_{h,y}^{\text{AC}}$ is equivalent to **Ada3PC** compressor $\mathcal{C}_{h,y}$. \square

B.2 Theorem 5

The proof of Theorem 5 requires several auxiliary results. Lemma 7 states the descent lemma typical for the analysis of biased compressors. Lemma 8 shows how individual **3PC** compressors, applied at clients, affect the aggregated divergence of gradient estimates from gradients. Lemma 9 presents a technical upper bound on Lyapunov function Ψ^t .

Lemma 7 (Lemma 2 of (Li et al., 2021)). *Suppose the function f is L_- -smooth and $x^{t+1} = x^t - \gamma g^t$, where $g^t \in \mathbb{R}^d$ is any vector, and $\gamma > 0$ is any scalar. Then we have*

$$f(x^{t+1}) - f(x^t) \leq -\frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2. \quad (20)$$

Lemma 8 (Lemma B.3 of (Richtárik et al., 2022)). *Let Assumption 3 hold. Consider Algorithm 1 with **3PC** compressor \mathcal{M}^W and identity compressor \mathcal{M}^M . Then for all $t \geq 0$ the sequence*

$$G^t = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^t) - g_i^t\|^2 \quad (21)$$

satisfies

$$\mathbb{E}[G^{t+1}] \leq (1 - A)\mathbb{E}[G^t] + BL_+^2 \mathbb{E}[\|x^{t+1} - x^t\|^2], \quad (22)$$

where A and B are parameters of \mathcal{M}^W .

Lemma 9. *Let Assumption 1 hold. Let Lyapunov function $\Psi^t := f(x^t) - f^* + \frac{\gamma}{A} G^t$. Then, for any $t \geq 0$, the following inequality holds*

$$\mathbb{E}\Psi^t \leq \sqrt{\left(\mathbb{E}[\|\nabla f(x^t)\|^2] + \frac{\gamma}{A} \mathbb{E}G^t \right) \left(\mathbb{E}[\|x^t - x^*\|^2] + \frac{\gamma}{A} \mathbb{E}[G^t] \right)}, \quad (23)$$

where x^* is any point belonging to $\text{Argmin } f(x)$.

Proof. By definition of convexity we get

$$\begin{aligned} \mathbb{E}\Psi^t &= \mathbb{E}f(x^t) - f^* + \frac{\gamma}{A} \mathbb{E}G^t \\ &\stackrel{(12)}{\leq} \mathbb{E}\langle \nabla f(x^t), x^t - x^* \rangle + \frac{\gamma}{A} \mathbb{E}G^t \\ &= \mathbb{E} \left\langle \left[\nabla f(x^t), \sqrt{\frac{\gamma}{A} \mathbb{E}G^t} \right], \left[x^t - x^*, \sqrt{\frac{\gamma}{A} \mathbb{E}G^t} \right] \right\rangle \end{aligned}$$

By applying Cauchy-Schwartz inequality on vectors and random variables we finish the proof

$$\begin{aligned}
& \mathbb{E} \left\langle \left[\nabla f(x^t), \sqrt{\frac{\gamma}{A} \mathbb{E} G^t} \right], \left[x^t - x^*, \sqrt{\frac{\gamma}{A} \mathbb{E} G^t} \right] \right\rangle \\
& \stackrel{(16)}{\leq} \mathbb{E} \left[\sqrt{\|\nabla f(x^t)\|^2 + \frac{\gamma}{A} \mathbb{E} G^t} \sqrt{\|x^t - x^*\|^2 + \frac{\gamma}{A} \mathbb{E} G^t} \right] \\
& \stackrel{(17)}{\leq} \sqrt{\left(\mathbb{E} [\|\nabla f(x^t)\|^2] + \frac{\gamma}{A} \mathbb{E} G^t \right) \left(\mathbb{E} [\|x^t - x^*\|^2] + \frac{\gamma}{A} \mathbb{E} [G^t] \right)}.
\end{aligned}$$

□

Now we are ready to prove the main theorem.

Theorem 5. *Let Assumptions 1, 2, 3 and 4 hold. Assume the stepsize γ of algorithm satisfies $0 < \gamma \leq 1/M$, where $M = L_- + L_+ \sqrt{\frac{2B}{A}}$. Then, for any $T \geq 0$ we have*

$$\mathbb{E} [f(x^T)] - f(x^*) \leq \max \left\{ \frac{1}{\gamma}, \frac{1}{A} \right\} \frac{2(\Omega^2 + \Psi^0)}{T}.$$

Proof. Combining Lemma 7, Jensen's inequality, we get

$$\begin{aligned}
f(x^{t+1}) - f(x^t) & \leq -\frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \left\| \frac{1}{n} \sum_{i=1}^n g_i^t - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^t) \right\|^2 \\
& \leq -\frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \frac{1}{n} \sum_{i=1}^n \|g_i^t - \nabla f_i(x^t)\|^2 \\
& = -\frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} G^t.
\end{aligned} \tag{24}$$

Now applying Equation (24) and Lemma 8 on the difference of Lyapunov functions, we obtain

$$\begin{aligned}
\mathbb{E} [\Psi^{t+1}] - \mathbb{E} [\Psi^t] & = \mathbb{E} [f(x^{t+1}) - f(x^t)] + \frac{\gamma}{A} \mathbb{E} [G^{t+1}] - \frac{\gamma}{A} \mathbb{E} [G^t] \\
& \stackrel{(24)}{\leq} -\frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] + \frac{\gamma}{2} \mathbb{E} [G^t] \\
& \quad + \frac{\gamma}{A} \mathbb{E} [G^{t+1}] - \frac{\gamma}{A} \mathbb{E} [G^t] \\
& \stackrel{(22)}{\leq} -\frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] - \left(\frac{1}{2\gamma} - \frac{L_-}{2} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& \quad + \frac{\gamma}{A} [(1-A)\mathbb{E}[G^t] + BL_+^2 \mathbb{E} \|x^{t+1} - x^t\|^2 - \mathbb{E}[G^t]].
\end{aligned}$$

Rearranging the term, we get

$$\mathbb{E} [\Psi^{t+1}] - \mathbb{E} [\Psi^t] \leq -\frac{\gamma}{2} [\|\nabla f(x^t)\|^2] - \left(\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{\gamma BL_+^2}{A} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] - \frac{A}{2} \frac{\gamma}{A} \mathbb{E} [G^t].$$

We further note that

$$\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{\gamma BL_+^2}{A} \geq 0 \Leftrightarrow L_+^2 \frac{2B}{A} \gamma^2 + L_- \gamma \leq 1 \stackrel{\text{Lemma 6}}{\Leftrightarrow} \gamma \leq \frac{1}{L_- + L_+ \sqrt{\frac{2B}{A}}}.$$

Appropriately chosen stepsize gives

$$\mathbb{E} [\Psi^{t+1}] - \mathbb{E} [\Psi^t] \leq -\min \left\{ \frac{\gamma}{2}, \frac{A}{2} \right\} \left(\mathbb{E} [\|\nabla f(x^t)\|^2] + \frac{\gamma}{A} \mathbb{E} [G^t] \right).$$

Rearranging the terms, we have

$$\mathbb{E} [\|\nabla f(x^t)\|^2] + \frac{\gamma}{A} \mathbb{E} [G^t] \leq \frac{\mathbb{E} [\Psi^t] - \mathbb{E} [\Psi^{t+1}]}{\min \left\{ \frac{\gamma}{2}, \frac{A}{2} \right\}}. \quad (25)$$

from what we deduce that $\mathbb{E} [\Psi^{t+1}] \leq \mathbb{E} [\Psi^t]$.

Using Lemma 9 and (25), we have

$$\begin{aligned} \mathbb{E} \Psi^{t+1} \mathbb{E} \Psi^t &\leq (\mathbb{E} \Psi^t)^2 \leq \left(\mathbb{E} [\|\nabla f(x^t)\|^2] + \frac{\gamma}{A} \mathbb{E} G^t \right) \left(\mathbb{E} [\|x^t - x^*\|^2] + \frac{\gamma}{A} \mathbb{E} [G^t] \right) \\ &\leq \frac{\mathbb{E} [\|x^t - x^*\|^2] + \frac{\gamma}{A} \mathbb{E} [G^t]}{\min \left\{ \frac{\gamma}{2}, \frac{A}{2} \right\}} (\mathbb{E} [\Psi^t] - \mathbb{E} [\Psi^{t+1}]). \end{aligned}$$

Using that $\frac{\gamma}{A} \mathbb{E} [G^t] \leq \mathbb{E} \Psi^t \leq \Psi_0$ and $\mathbb{E} [\|x^t - x^*\|^2] \leq \Omega^2$, we obtain

$$\mathbb{E} \Psi^{t+1} \mathbb{E} \Psi^t \leq \frac{\Omega^2 + \Psi^0}{\min \left\{ \frac{\gamma}{2}, \frac{A}{2} \right\}} (\mathbb{E} [\Psi^t] - \mathbb{E} [\Psi^{t+1}]).$$

Rearranging again, we get

$$\frac{\min \left\{ \frac{\gamma}{2}, \frac{A}{2} \right\}}{\Omega^2 + \Psi^0} \leq \left(\frac{1}{\mathbb{E} [\Psi^{t+1}]} - \frac{1}{\mathbb{E} [\Psi^t]} \right).$$

Summing up from $t = 0$ to $t = T - 1$, we finish the proof

$$\mathbb{E} [f(x^T)] - f(x^*) \leq \mathbb{E} [\Psi^T] \leq \max \left\{ \frac{2}{\gamma}, \frac{2}{A} \right\} \frac{\Omega^2 + \Psi^0}{T}. \quad (26)$$

□

B.3 Theorem 6

Algorithm 2 3PC-BD (Bidirectional 3PC algorithm)

- 1: **Input:** starting point $x^0 \in \mathbb{R}^d$; $g^0, \tilde{g}_i^0 \in \mathbb{R}^d$ for $i = 1, \dots, n$ (known by nodes), $\tilde{g}^0 = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^0$ (known by master); learning rate $\gamma > 0$.
 - 2: **for** $t = 0, 1, 2, \dots, T - 1$ **do**
 - 3: Broadcast g^t to all workers
 - 4: **for for all devices** $i = 1, \dots, n$ **in parallel do**
 - 5: $x^{t+1} = x^t - \gamma g^t$
 - 6: $\tilde{g}_i^{t+1} = \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1}))$
 - 7: Communicate \tilde{g}_i^{t+1} to the server
 - 8: **end for**
 - 9: $\tilde{g}^{t+1} = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^{t+1}$
 - 10: $g^{t+1} = \mathcal{C}_{g^t, \tilde{g}^t}^M(\tilde{g}^{t+1})$
 - 11: **end for**
-

For Theorem 6, we assume that both compressors \mathcal{M}^W and \mathcal{M}^M in Algorithm 1 are 3PC compressors. The main steps of the algorithm are:

$$\begin{aligned} x^{t+1} &= x^t - \gamma g^t \\ \tilde{g}_i^{t+1} &= \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) \\ \tilde{g}^{t+1} &= \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^{t+1} \\ g^{t+1} &= \mathcal{C}_{g^t, \tilde{g}^t}^M(\tilde{g}^{t+1}) \end{aligned}$$

Unlike in the previous subsection, we use additional notations: $P_i^t := \|\tilde{g}_i^t - \nabla f_i(x^t)\|^2$, $P^t := \frac{1}{n} \sum_{i=1}^n P_i^t$ and $R^t := \|x^{t+1} - x^t\|^2$.

Lemma 10 is an analogue of Lemma 8 (in bidirectional case we need slightly different arguments at some steps). Lemma 11 is another technical lemma that upper bounds $\mathbb{E}[\|g^t - \tilde{g}^t\|^2]$.

Lemma 10. *Let Assumption 3 hold, C^w be a 3PC compressor, and \tilde{g}_i^{t+1} be an 3PC-BD estimator of $\nabla f_i(x^{t+1})$, i.e.*

$$\tilde{g}_i^{t+1} = C_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) \quad (27)$$

for arbitrary \tilde{g}_i^0 for all $i \in [n]$, $t \geq 0$. Then

$$\mathbb{E}[P^{t+1}] \leq (1 - A^W)\mathbb{E}[P^t] + B^W L_+^2 \mathbb{E}[R^t]. \quad (28)$$

Proof. Define $W^t := \{\tilde{g}_1^t, \dots, \tilde{g}_n^t, x^t, x^{t+1}\}$, then

$$\begin{aligned} \mathbb{E}[P_i^{t+1}] &= \mathbb{E}[\mathbb{E}[P_i^{t+1} | W^t]] \\ &= \mathbb{E}\left[\mathbb{E}\left[\|\tilde{g}_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \mid W^t\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\|C_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) - \nabla f_i(x^{t+1})\right\|^2 \mid W^t\right]\right] \\ &\stackrel{(6)}{\leq} (1 - A^W)\mathbb{E}\left[\|\tilde{g}_i^t - \nabla f_i(x^t)\|^2\right] + B^W \mathbb{E}\left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2\right] \end{aligned} \quad (29)$$

Averaging the above inequalities over $i \in [n]$, we obtain (28). Indeed,

$$\begin{aligned} \mathbb{E}[P^{t+1}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n P_i^{t+1}\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[P_i^{t+1}] \\ &\stackrel{(29)}{\leq} \frac{1}{n} \sum_{i=1}^n (1 - A^W)\mathbb{E}\left[\|\tilde{g}_i^t - \nabla f_i(x^t)\|^2\right] + \frac{1}{n} \sum_{i=1}^n B^W \mathbb{E}\left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2\right] \\ &= (1 - A^W)\mathbb{E}[P^t] + B^W \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2\right] \\ &\stackrel{\text{Assumption 3}}{\leq} (1 - A^W)\mathbb{E}[P^t] + B^W L_+^2 \mathbb{E}\|x^{t+1} - x^t\|^2 \\ &= (1 - A^W)\mathbb{E}[P^t] + B^W L_+^2 \mathbb{E}[R^t]. \end{aligned}$$

□

Lemma 11. *Let Assumptions 3 and 5 hold, C^M, C^w be 3PC compressors. Let \tilde{g}_i^{t+1} be an 3PC-BD estimator of $\nabla f_i(x^{t+1})$, i.e.*

$$\tilde{g}_i^{t+1} = C_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) \quad (30)$$

and let g^{t+1} be an 3PC-BD estimator of $\tilde{g}^{t+1} = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^{t+1}$, i.e.

$$g_i^{t+1} = C_{g_i^t, \tilde{g}_i^t}^M(\tilde{g}_i^{t+1}) \quad (31)$$

for arbitrary g^0, \tilde{g}_i^0 for all $i \in [n]$, $t \geq 0$. Then

$$\mathbb{E}\left[\|g^{t+1} - \tilde{g}^{t+1}\|^2\right] \leq (1 - A^M)\mathbb{E}\left[\|g^t - \tilde{g}^t\|^2\right] + 3B^M(2 - A^W)\mathbb{E}[P^t] + 3B^M(B^W + 1)L_+^2 \mathbb{E}[R^t], \quad (32)$$

where $g^t = \frac{1}{n} \sum_{i=1}^n g_i^t$, $\tilde{g}^t = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^t$.

Proof. Similarly to the proof of Lemma 10, we define $W^t := \{g_1^t, \dots, g_n^t, x^t, x^{t+1}\}$ and bound $\mathbb{E} \left[\|g^{t+1} - \tilde{g}^{t+1}\|^2 \right]$:

$$\begin{aligned} \mathbb{E} \left[\|g^{t+1} - \tilde{g}^{t+1}\|^2 \right] &= \mathbb{E} \left[\mathbb{E} \left[\|g^{t+1} - \tilde{g}^{t+1}\|^2 \mid W^t \right] \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\|\mathcal{C}_{g^t, \tilde{g}^t}^M(\tilde{g}^{t+1}) - \tilde{g}^{t+1}\|^2 \mid W^t \right] \right] \\ &\stackrel{(6)}{\leq} (1 - A^M) \mathbb{E} \left[\|g^t - \tilde{g}^t\|^2 \right] + B^M \mathbb{E} \left[\|\tilde{g}^{t+1} - \tilde{g}^t\|^2 \right], \end{aligned} \quad (33)$$

Further, we bound the last term in (33). Recall that

$$\tilde{g}^{t+1} = \frac{1}{n} \sum_{i=1}^n \tilde{g}_i^{t+1} = \frac{1}{n} \sum_{i=1}^n \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})). \quad (34)$$

Then,

$$\begin{aligned} \mathbb{E} \left[\|\tilde{g}^{t+1} - \tilde{g}^t\|^2 \right] &= \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) - \tilde{g}_i^t \right\|^2 \right] \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) - \tilde{g}_i^t \right\|^2 \right] \\ &\stackrel{(19)}{\leq} \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \mathcal{C}_{\tilde{g}_i^t, \nabla f_i(x^t)}^w(\nabla f_i(x^{t+1})) - \nabla f_i(x^{t+1}) \right\|^2 \right] \\ &\quad + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \right] + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^t) - \tilde{g}_i^t\|^2 \right] \\ &\stackrel{(6)}{\leq} 3(1 - A^W) \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^t) - \tilde{g}_i^t\|^2 \right] + 3B^W \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \right] \\ &\quad + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \right] + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x^t) - \tilde{g}_i^t\|^2 \right] \\ &\stackrel{\text{Assumption 3}}{\leq} 3(2 - A^W) \mathbb{E} [P^t] + 3(B^W + 1)L_+^2 \mathbb{E} \left[\|x^{t+1} - x^t\|^2 \right] \\ &= 3(2 - A^W) \mathbb{E} [P^t] + 3(B^W + 1)L_+^2 \mathbb{E} [R^t], \end{aligned} \quad (35)$$

where the first inequality follows from Young's inequality. Plugging (35) into (33) we finish the proof:

$$\begin{aligned} \mathbb{E} \left[\|g^{t+1} - \tilde{g}^{t+1}\|^2 \right] &\leq (1 - A^M) \mathbb{E} \left[\|g^t - \tilde{g}^t\|^2 \right] + 3B^M(2 - A^W) \mathbb{E} [P^t] \\ &\quad + 3B^M(B^W + 1)L_+^2 \mathbb{E} [R^t]. \end{aligned}$$

□

Having proved the previous lemmas, we can now show the convergence of bidirectional 3PC algorithm.

Theorem 6. *Let Assumptions 3 and 5 hold, and let the stepsize in Algorithm 2 be set as*

$$0 \leq \gamma < \left(L_- + L_+ \sqrt{\frac{6B^M(B^W + 1)}{A^M} + \frac{2B^W}{A^M} \left(1 + \frac{3B^M(2 - A^W)}{A^M} \right)} \right)^{-1}. \quad (36)$$

Fix T and let \hat{x}^T be chosen uniformly from $\{x^0, x^1, \dots, x^{T-1}\}$ uniformly at random. Then

$$\mathbb{E} \left[\|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Psi^0}{\gamma T}. \quad (37)$$

where $\Psi^T = f(x^t) - f^{\text{inf}} + \frac{\gamma}{A^M} \|g^t - \tilde{g}^t\|^2 + \frac{\gamma}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M} \right) \frac{1}{n} \sum_{i=1}^n \|\tilde{g}_i^t - \nabla f_i(x^t)\|^2$.

Proof. We apply Lemma 7 and split the error $\|g^t - \nabla f(x^t)\|^2$ into two parts

$$\begin{aligned}
f(x^{t+1}) &\stackrel{(20)}{\leq} f(x^t) - \frac{\gamma}{2} \|\nabla f(x)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) R^t + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2 \\
&\stackrel{(18)}{\leq} f(x^t) - \frac{\gamma}{2} \|\nabla f(x)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) R^t + \gamma \|\tilde{g}^t - \nabla f(x^t)\|^2 + \gamma \|\tilde{g}^t - g^t\|^2 \\
&\leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) R^t + \frac{\gamma}{n} \sum_{i=1}^n \|\tilde{g}_i^t - \nabla f_i(x^t)\|^2 + \gamma \|\tilde{g}^t - g^t\|^2 \\
&= f(x^t) - \frac{\gamma}{2} \|\nabla f(x)\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) R^t + \gamma P^t + \gamma \|\tilde{g}^t - g^t\|^2, \tag{38}
\end{aligned}$$

where in the last inequality we applied Young's inequality. Subtracting f^{inf} from both sides of the above inequality, taking expectation and using the notation $\delta^t = f(x^t) - f^{\text{inf}}$, we get

$$\mathbb{E}[\delta^{t+1}] \leq \mathbb{E}[\delta^t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) \mathbb{E}[R^t] + \gamma \mathbb{E}[P^t] + \gamma \mathbb{E}[\|\tilde{g}^t - g^t\|^2]. \tag{39}$$

Further, Lemmas 10 and 11 provide the recursive bounds for the last two terms of (39)

$$\mathbb{E}[P^{t+1}] \leq (1 - A^W) \mathbb{E}[P^t] + B^W L_+^2 \mathbb{E}[R^t], \tag{40}$$

$$\begin{aligned}
\mathbb{E}[\|g^{t+1} - \tilde{g}^{t+1}\|^2] &\leq (1 - A^M) \mathbb{E}[\|g^t - \tilde{g}^t\|^2] + 3B^M(2 - A^W) \mathbb{E}[P^t] \\
&\quad + 3B^M(B^W + 1)L_+^2 \mathbb{E}[R^t]. \tag{41}
\end{aligned}$$

Summing up (39) with a $\frac{\gamma}{A^M}$ multiple of (41) we obtain

$$\begin{aligned}
\mathbb{E}[\delta^{t+1}] + \frac{\gamma}{A^M} \mathbb{E}[\|g^t - \tilde{g}^t\|^2] &\leq \mathbb{E}[\delta^t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right) \mathbb{E}[R^t] \\
&\quad + \gamma \mathbb{E}[P^t] + \gamma \mathbb{E}[\|\tilde{g}^t - g^t\|^2] \\
&\quad + \frac{\gamma}{A^M} \left((1 - A^M) \mathbb{E}[\|g^t - \tilde{g}^t\|^2] \right) \\
&\quad + \frac{\gamma}{A^M} \left(3B^M(2 - A^W) \mathbb{E}[P^t] + 3B^M(B^W + 1)L_+^2 \mathbb{E}[R^t] \right) \\
&\leq \mathbb{E}[\delta^t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] + \frac{\gamma}{A^M} \mathbb{E}[\|g^t - \tilde{g}^t\|^2] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{3\gamma B^M(B^W + 1)L_+^2}{A^M}\right) \mathbb{E}[R^t] \\
&\quad + \gamma \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right) \mathbb{E}[P^t].
\end{aligned}$$

Then adding the above inequality with a $\frac{\gamma}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right)$ multiple of (40), we get

$$\begin{aligned}
\mathbb{E}[\Psi^{t+1}] &= \mathbb{E}[\delta^{t+1}] + \frac{\gamma}{A^M} \mathbb{E}[\|g^t - \tilde{g}^t\|^2] + \frac{\gamma}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right) \mathbb{E}[P^{t+1}] \\
&\leq \mathbb{E}[\delta^t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] + \frac{\gamma}{A^M} \mathbb{E}[\|g^t - \tilde{g}^t\|^2] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{3\gamma B^M(B^W + 1)L_+^2}{A^M}\right) \mathbb{E}[R^t] + \gamma \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right) \mathbb{E}[P^t] \\
&\quad + \frac{\gamma}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right) \left((1 - A^W) \mathbb{E}[P^t] + B^W L_+^2 \mathbb{E}[R^t] \right) \\
&\leq \mathbb{E}[\delta^t] + \frac{\gamma}{A^M} \mathbb{E}[\|g^t - \tilde{g}^t\|^2] + \frac{\gamma}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right) \mathbb{E}[P^t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{3\gamma B^M(B^W + 1)L_+^2}{A^M} - \frac{\gamma B^W L_+^2}{A^W} \left(1 + \frac{3B^M(2 - A^W)}{A^M}\right)\right) \mathbb{E}[R^t]. \tag{42}
\end{aligned}$$

Thus by Lemma 6 and the choice of the stepsize

$$0 \leq \gamma < \left(L + L_+ \sqrt{\frac{6B^M(B^W + 1)}{A^M} + \frac{2B^W}{A^M} \left(1 + \frac{3B^M(2 - A^W)}{A^M} \right)} \right)^{-1}, \quad (43)$$

the last term in (42) is not positive. By summing up inequalities for $t = 0, 1, \dots, T - 1$, we get

$$0 \leq \mathbb{E} [\Psi^T] \leq \Psi^0 - \frac{\gamma}{2} \sum_{i=1}^{T-1} \mathbb{E} [\|\nabla f(x^t)\|^2].$$

Multiplying both sides by $\frac{2}{\gamma T}$ and rearranging we get

$$\frac{1}{T} \sum_{i=1}^{T-1} \mathbb{E} [\|\nabla f(x^t)\|^2] \leq \frac{2\Psi^0}{\gamma T}.$$

□

B.4 Convergence for general nonconvex functions

The results in two subsequent subsections set \mathcal{M}^W as a 3PC compressor and \mathcal{M}^M as an identity one. According to Lemma 2, Adaptive 3PC is a 3PC compressor. Thus, convergence results from (Richtárik et al., 2022) are valid for Adaptive 3PC compressor. It leads us to the following corollary.

Corollary 3 (Corollary 5.6 of (Richtárik et al., 2022)). *Let Assumptions 2, 3 and 5 hold. Let \mathcal{M}^W and \mathcal{M}^M in Algorithm 1 be Ada3PC and identity compressors, respectively, and choose the stepsize $\gamma = \frac{1}{L_- + L_+ \sqrt{\frac{B_{\max}}{A_{\min}}}}$. Then, for any $T \geq 1$ we have*

$$\mathbb{E} [\|\nabla f(\hat{x}^T)\|^2] \leq \frac{2(f(x^0) - f(x^{\text{inf}})) \left(L_- + L_+ \sqrt{\frac{B_{\max}}{A_{\min}}} \right)}{T} + \frac{\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \|g_i^0 - \nabla f_i(x^0)\|^2 \right]}{A_{\min} T}.$$

That is, to achieve $\mathbb{E} [\|\nabla f(\hat{x}^T)\|^2] \leq \varepsilon^2$ for some $\varepsilon > 0$, Algorithm 1 requires

$$T = \mathcal{O} \left(\frac{2(f(x^0) - f(x^{\text{inf}})) \left(L_- + L_+ \sqrt{\frac{B_{\max}}{A_{\min}}} \right)}{\varepsilon^2} + \frac{\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \|g_i^0 - \nabla f_i(x^0)\|^2 \right]}{A_{\min} \varepsilon^2} \right)$$

iterations.

B.5 Convergence for PŁ nonconvex functions

The setup here is the same as in the previous subsection, except we add the following assumption.

Assumption 6 (PŁ condition). *Function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the Polyak-Łojasiewicz (PŁ) condition with parameter $\mu > 0$, i.e.,*

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^d,$$

where $x^* := \arg \min_{x \in \mathbb{R}^d} f(x)$ and $f^* := f(x^*)$.

Corollary 4 (Corollary 5.9 of (Richtárik et al., 2022)). *Let Assumptions 2, 3, 5 and 6 hold. Let \mathcal{M}^W and \mathcal{M}^M in Algorithm 1 be Ada3PC and identity compressors, respectively, and choose the stepsize*

$$\gamma = \min \left\{ \frac{1}{L_- + L_+ \sqrt{\frac{2B_{\max}}{A_{\min}}}}, \frac{A_{\min}}{2\mu} \right\}.$$

Then, to achieve $\mathbb{E} [f(x^T)] - f^* \leq \varepsilon$ for some $\varepsilon > 0$ the method requires

$$\mathcal{O} \left(\max \left\{ \frac{L_- + L_+ \sqrt{\frac{B_{\max}}{A_{\min}}}}{\mu}, A_{\min} \right\} \log \frac{f(x^0) - f(x^{\text{inf}}) + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \|g_i^0 - \nabla f_i(x^0)\|^2 \gamma / A_{\min} \right]}{\varepsilon} \right)$$

iterations.

C Experimental details and extra experiments

All simulations are implemented in Python 3.8 and run on Intel(R) Xeon(R) Gold 6230R CPU cluster with 48 nodes. We fine-tune the stepsize of each considered algorithm with $(2^0, 2^1, \dots, 2^8)$ multiples of the corresponding theoretical stepsize. As contractive compressor we use Top- k operator. For **EF21** and **CLAG** we use top-1 compressor, which usually the best in practice for these methods. For **AdaCGD** we choose compressors varying from full compression (skip communication) to compression of 50% of features. In order to provide fair comparisons, we choose master compressor \mathcal{M}^M as identity operator in these experiments. For the stopping criterion we choose communication cost of the algorithm.

We use the setup described in Richtárik et al. (2022), namely logistic regression with non-convex regularizer:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{N} \sum_{i=1}^N \log(1 + e^{-y_i a_i^\top x}) + \lambda \sum_{j=1}^d \frac{x_j^2}{1+x_j^2} \right],$$

where $a_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ are the training data and labels, and $\lambda > 0$ is a regularization parameter, which is fixed to $\lambda = 0.1$. We solve this problem using LIBSVM Chang & Lin (2011) datasets *phishing*, *ala*, *a9a*. Each dataset has been evenly split into $n = 20$ equal parts where each part represents a separate client. Figures 2-4 compare **AdaCGD** with **LAG**, **EF21** and their generalization **CLAG**. In the experiments, **AdaCGD** is shown to be comparable and in some cases superior to **CLAG** and always superior to **LAG**. In other words, **AdaCGD** efficiently complements **CLAG** and other 3PC methods.

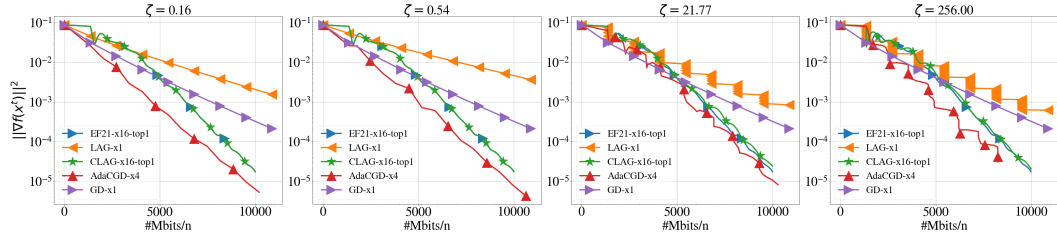


Figure 2: Comparison of **LAG**, **CLAG**, **EF21** and **GD** with **AdaCGD** on *phishing* dataset.

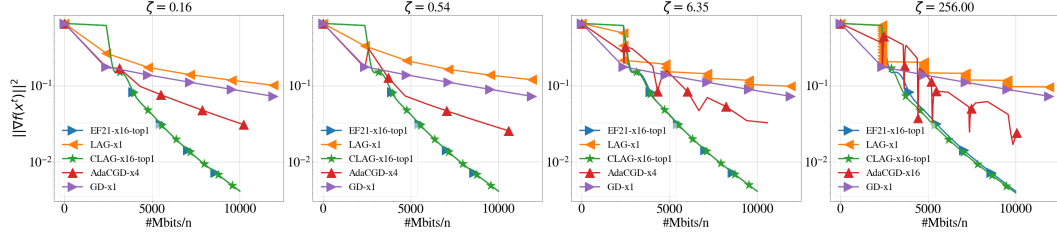


Figure 3: Comparison of **LAG**, **CLAG**, **EF21** and **GD** with **AdaCGD** on *ala* dataset.

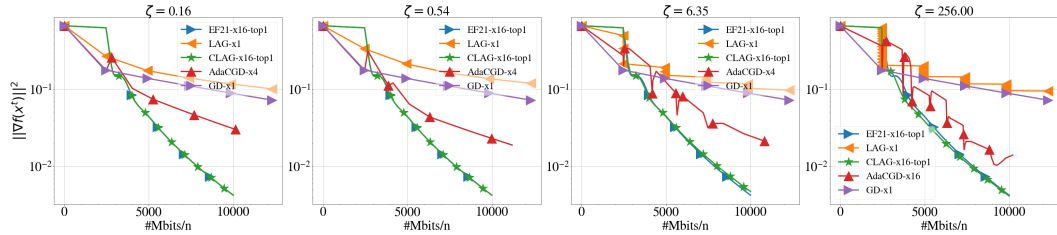


Figure 4: Comparison of **LAG**, **CLAG**, **EF21** and **GD** with **AdaCGD** on *a9a* dataset.

D Limitations

The main limitations of the work are assumptions we make upon functions f_i of the problem 1. But, on the other hand, these assumptions govern the convergence rates we report: for example, we cannot show linear rate for convex functions due to the fundamental lower bound (Nesterov et al., 2018).

Another limitation comes from the analysis of Bidirectional 3PC algorithm (Theorem 6). We show the analysis only for general nonconvex functions.