Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization

Zhize Li[1], Dmitry Kovalev[1], Xun Qian[1], and Peter Richtárik[1]

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

February 26, 2020

Abstract

Due to the high communication cost in distributed and federated learning problems, methods relying on compression of communicated messages are becoming increasingly popular. While in other contexts the best performing gradient-type methods invariably rely on some form of acceleration/momentum to reduce the number of iterations, there are no methods which combine the benefits of both gradient compression and acceleration. In this paper, we remedy this situation and propose the first *accelerated compressed gradient descent (ACGD)* methods. In the single machine regime, we prove that ACGD enjoys the rate $O((1+\omega)\sqrt{L/\mu}\log 1/\epsilon)$ for μ-strongly convex problems and $O((1+\omega)\sqrt{L/\epsilon})$ for convex problems, respectively, where L is the smoothness constant and ω is the compression parameter. Our results improve upon the existing non-accelerated rates $O\left((1+\omega)L/\mu \log 1/\epsilon\right)$ and $O\left((1+\omega)L/\epsilon\right)$, respectively, and recover the optimal rates of accelerated gradient descent as a special case when no compression ($\omega = 0$) is applied. We further propose a distributed variant of ACGD (called ADIANA) and prove the convergence rate $\widetilde{O}\left(\omega + \sqrt{L/\mu} + \sqrt{(\omega/n + \sqrt{\omega/n})\omega L/\mu}\right)$, where n is the number of devices/workers and $\widetilde{O}$ hides the logarithmic factor $\log 1/\epsilon$. This improves upon the previous best result $\widetilde{O}\left(\omega + L/\mu + \omega L/n\mu\right)$ achieved by the DIANA method of Mishchenko et al. (2019b). Finally, we conduct several experiments on real-world datasets which corroborate our theoretical results and confirm the practical superiority of our methods.

Contents

1 Introduction

With the proliferation of edge devices such as mobile phones, wearables and smart home devices comes an increase in the amount of data rich in potential information which can be mined for the benefit of the users. One of the approaches of turning the raw data into information is via federated learning (Konečný et al., 2016; McMahan et al., 2017), where typically a single global supervised model is trained in a massively distributed manner over a network of heterogeneous devices.

Training supervised federated learning models is typically performed by solving an optimization problem of the form

$$\min_{x\in\mathbb{R}^d} \Big\{ P(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x) \Big\}, \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is smooth loss associated with data stored on device i and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a relatively simple but possibly nonsmooth regularizer.

In distributed learning in general, and federated learning in particular, communication of messages across a network forms the bottleneck of the training system. It is thus very important to devise novel strategies for reducing the number of communication rounds. Two of the most common strategies are i) local computations (Ma et al., 2017; Stich, 2019; Khaled et al., 2020) and ii) communication compression (Seide et al., 2014; Alistarh et al., 2017; Wangni et al., 2018; Horváth et al., 2019a). The former is used to perform more local computations on each device before communication and subsequent model averaging, hoping that this will reduce the total number of communications. The latter is used to reduce the size of communicated messages, saving precious time spent in each communication round, and hoping that this will not increase the total number of communications.

1.1 Theoretical inefficiency of local methods

Despite their practical success, local methods are poorly understood and there is much to be discovered. For instance, there exist no theoretical results which would suggest that any local method (e.g., local gradient descent (GD) or local SGD) can achieve better communication complexity than its standard non-local variant (e.g., GD, SGD). In fact, until recently, no complexity results existed for local SGD in environments with heterogeneous data (Khaled et al., 2019, 2020), a key regime in federated learning settings (Li et al., 2019). In the important regime when all participating devices compute full gradients based on their local data, the recently proposed stochastic controlled averaging (SCAFFOLD) method (Karimireddy et al., 2019) offers no improvement on the number of communication as the number of local steps grows despite the fact that this is a rather elaborate method combining local stochastic gradient descent with control variates for reducing the model drift among clients.

1.2 Methods with compressed communication

However, the situation is much brighter with methods employing communication compression. Indeed, several recent theoretical results suggest that by combining an appropriate (typically randomized) compression operator with a suitably designed gradient-type method, one can obtain improvement in the the total communication complexity over comparable baselines not performing any compression. For instance, this is the case for distributed compressed gradient descent (CGD) (Alistarh et al., 2017; Khirirat et al., 2018; Horváth et al., 2019a) and distributed CGD methods which employ variance reduction to tame the variance introduced by compression (Hanzely et al., 2018; Mishchenko et al., 2019a; Horváth et al., 2019b; Hanzely & Richtárik, 2019b).

While in the case of CGD compression leads to a decrease in the size of communicated messages per communication round, it leads to an increase in the number of communications. Yet, certain compression operators, such as natural dithering (Horváth et al., 2019a), were shown to be better than no compression in terms of the overall communication complexity.

The variance-reduced CGD method DIANA (Mishchenko et al., 2019a; Horváth et al., 2019b) enjoys even better behavior: the number of communication rounds for this method is unaffected up to a certain level of compression when the variance induced by compression reaches a certain threshold. This threshold can be very large in practice, which means that massive reduction is often possible in the number of communicated bits without this having any adverse effect on the number of communication rounds.

Recall that $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth (for $L>0$) if

$$\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|, \tag{2}$$

and μ-strongly convex (for $\mu\geq 0$) if

$$f(x)-f(y)-\langle\nabla f(y),x-y\rangle\geq\frac{\mu}{2}\|x-y\|^2 \tag{3}$$

for all $x,y\in\mathbb{R}^d$. The $\mu=0$ case corresponds to the standard convexity.

In particular, for L-smooth and μ-strongly convex f with n machines, DIANA enjoys the iteration bound $O\left(\left(\omega+\frac{L}{\mu}+\frac{L}{\mu}\frac{\omega}{n}\right)\log\frac{1}{\epsilon}\right)$, where $\frac{L}{\mu}$ is the condition number and ω is the variance parameter associated with the compressor. If $\omega=0$, which corresponds to no compression, one recovers the $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$ rate of gradient descent. On the other hand, as long as $\omega=O\left(\min\left\{\frac{L}{\mu},n\right\}\right)$, the rate is still $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$, which shows that DIANA is able to retain the same number of communication rounds as gradient descent and yet save on bit transmission in each round. The higher ω is allowed to be, the more compression can be applied.

2 Contributions

Discouraged by the lack of theoretical results suggesting that local methods indeed help to reduce the number of communications, and encouraged by the theoretical success of CGD methods, in this paper we seek to enhance CGD methods with a mechanism which, unlike local updating, can provably lead to a decrease of the number of communication rounds.

What mechanism could achieve further improvements?

In the world of deterministic gradient methods, one technique for such a reduction is well known: *Nesterov acceleration / momentum* (Nesterov, 1983, 2004). In case of SGD for finite sum problems, the momentum method Katyusha (Allen-Zhu, 2017) achieves the optimal rate; for enhancements, see also (Kovalev et al., 2020; Qian et al., 2019; Lan et al., 2019). Essentially all state-of-the-art methods for training deep learning models, including Adam (Kingma & Ba, 2014), rely on the use of momentum in one form or another, albeit lacking in theoretical support.

However, the successful combination of gradient compression and acceleration/momentum has so far remained elusive, and to the best of our knowledge, no algorithms nor theoretical results exist in this space. Given the omnipresence of momentum in modern machine learning, this is surprising.

We now summarize our key contributions:

Table 1: Convergence results for the special case with $n = 1$ device (i.e., problem (4))

Algorithm	μ-strongly convex f	convex f
Compressed Gradient Descent (CGD (Khirirat et al., 2018))	$O\left((1+\omega)\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left((1+\omega)\frac{L}{\epsilon}\right)$
ACGD (this paper)	$O\left((1+\omega)\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$O\left((1+\omega)\sqrt{\frac{L}{\epsilon}}\right)$

Table 2: Convergence results for the general case with n devices (i.e., problem (1)). Our results are always better than previous results.

Algorithm	$n \leq \omega$ (few devices or high compression)	$n > \omega$ (lots of devices or low compression)
Distributed CGD (DIANA (Mishchenko et al., 2019b))	$O\left(\omega\left(1+\frac{L}{n\mu}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\left(\omega+\frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$
ADIANA (this paper)	$O\left(\omega\left(1+\sqrt{\frac{L}{n\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\left(\omega+\sqrt{\frac{L}{\mu}}+\sqrt{\sqrt{\frac{\omega}{n}}\frac{\omega L}{\mu}}\right)\log\frac{1}{\epsilon}\right)$

2.1 First combination of gradient compression and acceleration

We develop the first gradient-type optimization methods combining the benefits of gradient compression and acceleration: i) ACGD (Algorithm 1) in the single device case, and ii) ADIANA (Algorithm 2) in the distributed case.

2.2 Single device setting

We first study the single-device setting, and design an accelerated CGD method (ACGD - Algorithm 1) for solving the unconstrained smooth minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \tag{4}$$

in the regimes when f is L-smooth and i) μ-strongly convex, and ii) convex. Our theoretical results are summarized in Table 1. In the strongly convex case, we improve the complexity of CGD (Khirirat et al., 2018) from $O\left((1+\omega)\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$ to $O\left((1+\omega)\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$. In the convex case, the improvement is from $O\left((1+\omega)\frac{L}{\epsilon}\right)$ to $O\left((1+\omega)\sqrt{\frac{L}{\epsilon}}\right)$, where $\omega \geq 0$ denotes the compression parameter (see Def. 1).

2.3 Distributed setting

We further study the distributed setting with n devices/workers and focus on problem (1) in its full generality. The presence of multiple nodes ($n > 1$) and of the regularizer ψ poses additional challenges. In order to address them, we need to not only combine acceleration and compression, but also introduce a DIANA-like variance reduction mechanism to remove the variance introduced by the compression operators.

In particular, we have developed an accelerated variant of the DIANA method for solving problem (1), which we call ADIANA (Algorithm 2). Our complexity results covering the strongly convex case for ADIANA are summarized in Table 2. Note that our results always improve upon the non-accelerated DIANA method. Indeed, in the regime when the compression parameter ω is larger than the number of nodes n, we improve the DIANA rate $O\left(\omega\left(1+\frac{L}{n\mu}\right)\log\frac{1}{\epsilon}\right)$ to $O\left(\omega\left(1+\sqrt{\frac{L}{n\mu}}\right)\log\frac{1}{\epsilon}\right)$.

On the other hand, in the regime when the compression parameter ω is smaller than the number of nodes n, we improve the DIANA rate $O\left(\left(\omega+\frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ to $O\left(\left(\omega+\sqrt{\frac{L}{\mu}}+\sqrt{\sqrt{\frac{\omega}{n}}\frac{\omega L}{\mu}}\right)\log\frac{1}{\epsilon}\right)$. Our rate is better because of the inequalities $\omega+\frac{L}{\mu} \geq 2\sqrt{\omega\frac{L}{\mu}}$ and $\sqrt{\frac{\omega}{n}} < 1$, which follows from $n > \omega$.

Note that if $\omega \leq n^{1/3}$, which is more often true in federated learning than in classical distributed learning as the number of devices in federated learning is typically very large, our result reduces to $O\left(\left(\omega+\sqrt{\frac{L}{\mu}}\right)\log\frac{1}{\epsilon}\right)$. In

particular, if $\omega = O\left(\min\left\{n^{1/3}, \sqrt{\frac{L}{\mu}}\right\}\right)$, then the communication complexity of our ADIANA is $O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$; the same as that of accelerated gradient descent (AGD) (Nesterov, 2004). It means that ADIANA benefits from cheaper communication due to compression *for free* without hurting the convergence rate, and is therefore better suited for distributed optimization.

3 Randomized Compression Operators

3.1 Unbiased compression operators with bounded variance

We now introduce the notion of a randomized compression operator which is used to compress the gradients.

Definition 1 (Compression operator) *A randomized map $\mathcal{C}:\mathbb{R}^d \mapsto \mathbb{R}^d$ is an ω-compression operator if*

$$\mathbb{E}[\mathcal{C}(x)] = x, \qquad \mathbb{E}[\|\mathcal{C}(x)-x\|^2] \leq \omega\|x\|^2, \qquad \forall x \in \mathbb{R}^d. \tag{5}$$

In particular, no compression ($\mathcal{C}(x)\equiv x$) implies $\omega = 0$.

Note that the conditions (5) require the compression operator to be unbiased and its variance uniformly bounded by a multiple of the squared norm of the vector which we are compressing.

3.2 Examples

We now give a few examples of compression operators without attempting to be exhaustive.
Example 1 (Random sparsification): Given $x \in \mathbb{R}^d$, the random-k sparsification operator is defined by

$$\mathcal{C}(x) := \frac{d}{k}(\xi_k \odot x),$$

where $\odot$ denotes the Hadamard (element-wise) product and $\xi_k \in \{0,1\}^d$ is a uniformly random binary vector with k nonzero entries ($\|\xi_k\|_0 = k$). This random-k sparsification operator $\mathcal{C}$ satisfies (5) with $\omega = \frac{d}{k} - 1$.
Example 2 (Quantization): Given $x \in \mathbb{R}^d$, the (p,s)-quantization operator is defined by

$$\mathcal{C}(x) := \operatorname{sign}(x)\cdot\|x\|_p \cdot \frac{1}{s}\cdot\xi_s,$$

where $\xi_s \in \mathbb{R}^d$ is a random vector with i-th element

$$\xi_s(i) := \begin{cases} l+1, & \text{with probability } \frac{|x_i|}{\|x\|_p}s - l \\ l, & \text{otherwise} \end{cases},$$

where the level l satisfies $\frac{|x_i|}{\|x\|_p} \in [\frac{l}{s}, \frac{l+1}{s}]$. The probability is chosen so that $\mathbb{E}[\xi_s(i)] = \frac{|x_i|}{\|x\|_p}s$. This (p,s)-quantization operator $\mathcal{C}$ satisfies (5) with $\omega = 2 + \frac{d^{1/p}+d^{1/2}}{s}$. In particular, QSGD (Alistarh et al., 2017) used $p=2$ (i.e., $(2,s)$-quantization) and proved that the expected sparsity of $\mathcal{C}(x)$ is $\mathbb{E}[\|\mathcal{C}(x)\|_0] = O\left(s(s+\sqrt{d})\right)$.

4 Accelerated CGD: Single Machine

In this section, we study the special case of problem (1) with a single machine ($n=1$) and no regularizer ($\psi(x)\equiv 0$), i.e., the problem

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

4.1 The CGD algorithm

First, we recall the update step in compressed gradient descent (CGD) method, i.e.,

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

where $\mathcal{C}$ is a (fresh sample of) ω-compression operator defined in Definition 1.

As mentioned earlier, convergence results of CGD are $O\left((1+\omega)\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$ for strongly convex problems and $O\left((1+\omega)\frac{L}{\epsilon}\right)$ for convex problems (see Table 1). The convergence proof for strongly convex problems (i.e., $O\left((1+\omega)\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$) can be found in (Khirirat et al., 2018). For completeness, we now establish a convergence result for convex functions.

Theorem 1 *Suppose f is convex with L-Lipschitz continuous gradient and the compression operator $\mathcal{C}$ satisfies* (5). *Fixing the step size $\eta = \frac{1}{(1+\omega)L}$, the number of iterations performed by CGD to find an ϵ-solution such that*

$$\mathbb{E}[f(x^k)] - f(x^*) \leq \epsilon$$

is at most

$$k = O\left(\frac{(1+\omega)L}{\epsilon}\right).$$

4.2 The ACGD algorithm

Algorithm 1 Accelerated CGD (ACGD)

Input: initial point x^0, $\{\eta_k\}, \{\theta_k\}, \{\beta_k\}, \{\gamma_k\}$, p

$z^0 = y^0 = x^0$

for $k = 0, 1, 2, \ldots$ **do**

  $x^k = \theta_k y^k + (1-\theta_k) z^k$

  Compress gradient $g^k = \mathcal{C}(\nabla f(x^k))$

  $y^{k+1} = x^k - \frac{\eta_k}{p} g^k$

  $z^{k+1} = \frac{1}{\gamma_k} y^{k+1} + \left(\frac{1}{p} - \frac{1}{\gamma_k}\right) y^k + \left(1 - \frac{1}{p}\right)(1-\beta_k) z^k + \left(1 - \frac{1}{p}\right)\beta_k x^k$

end for

Note that in the non-compressed case $\omega = 0$ (i.e., CGD is reduced to standard GD), there exists methods for obtaining accelerated convergence rates of $O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ and $O\left(\sqrt{\frac{L}{\epsilon}}\right)$ for strongly convex and convex problems, respectively. However, no accelerated convergence results exist for CGD methods. Inspired by Nesterov's accelerated gradient descent (AGD) method (Nesterov, 2004) and FISTA (Beck & Teboulle, 2009), we propose an accelerated compressed gradient descent (ACGD) method, described in Algorithm 1.

4.3 Convergence theory

Our accelerated convergence results for ACGD are stated in Theorems 2 and 3, formulated next.

Theorem 2 (ACGD: convex case) *Let f be convex with L-Lipschitz continuous gradient and let the compression operator $\mathcal{C}$ satisfy* (5). *Choose the parameters in ACGD (Algorithm 1) as follows:*

$$\eta_k \equiv \frac{1}{L}, \qquad \theta_k = 1 - \frac{2}{k+2}, \qquad \beta_k \equiv 0, \qquad \gamma_k = \frac{2p}{k+2}, \qquad p = 1+\omega.$$

Then the number of iterations performed by ACGD to find an ϵ-solution such that

$$\mathbb{E}[f(x^k)] - f(x^*) \leq \epsilon$$

is at most

$$k = O\left((1+\omega)\sqrt{\frac{L}{\epsilon}}\right).$$

Theorem 3 (ACGD: strongly convex case) *Let f be μ-strongly convex with L-Lipschitz continuous gradient and let the compression operator $\mathcal{C}$ satisfy* (5)*.Choose the parameters in ACGD (Algorithm 1) as follows:*

$$\eta_k \equiv \frac{1}{L}, \qquad \theta_k \equiv \frac{p}{p+\sqrt{\frac{\mu}{L}}}, \qquad \beta_k \equiv \frac{\sqrt{\frac{\mu}{L}}}{p}, \qquad \gamma_k \equiv \sqrt{\frac{\mu}{L}}, \qquad p = 1+\omega.$$

Then the number of iterations performed by ACGD to find an ϵ-solution such that

$$\mathbb{E}[f(x^k)] - f(x^*) \leq \epsilon$$

(or $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon$) is at most

$$k = O\left((1+\omega)\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right).$$

In the non-compressed case $\omega = 0$ (i.e., $\mathcal{C}(x) \equiv x$), our results recover the standard optimal rates of accelerated gradient descent. Further, if we consider the random-k sparsification operator, ACGD can be seen as a variant of accelerated randomized coordinate descent (Nesterov, 2012). Our results recover the optimal results of accelerated randomized coordinate descent method (Allen-Zhu et al., 2016; Hanzely & Richtárik, 2019a) under standard smoothness assumptions.

4.4 Proof sketch

The following lemma which demonstrates improvement in one iteration plays a key role in our analysis.

Lemma 1 *If parameters $\{\eta_k\}, \{\theta_k\}, \{\beta_k\}$, $\{\gamma_k\}$ and p satisfy $\theta_k = \frac{1-\gamma_k/p}{1-\beta_k\gamma_k/p}$ $\beta_k \leq \min\{\frac{\mu\eta_k}{\gamma_k p}, 1\}$, $p \geq \frac{(1+L\eta_k)(1+\omega)}{2}$ and the compression operator $\mathcal{C}^k$ satisfies* (5)*, then we have for any iteration k, and for all $x \in \mathbb{R}^d$, we have*

$$\frac{2\eta_k}{\gamma_k^2}\mathbb{E}[f(y^{k+1}) - f(x)] + \mathbb{E}[\|z^{k+1} - x\|^2] \leq \left(1 - \frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k) - f(x)\right) + (1-\beta_k)\|z^k - x\|^2,$$

where the expectation is with respect to the randomness of compression operator sampled at iteration k.

The proof of Theorems 2 and 3 can be derived (i.e., plug into the specified parameters and collect all iterations) from Lemma 1. The detailed proof can be found in the Supplementary Material.

5 Accelerated CGD: Distributed Setting

We now turn our attention to the general distributed case, i.e., problem (1):

$$\min_{x\in\mathbb{R}^d}\Big\{P(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) + \psi(x)\Big\}.$$

In this case, each machine/agent computes its local gradient (e.g., $\nabla f_i(x^k)$) and a perturbed version thereof is compressed and sent to the server. The server subsequently aggregates all received messages, to form a stochastic gradient estimator g^k of $\frac{1}{n}\sum_i \nabla f(x^k)$, and then performs a prox step. The perturbations h_i^k are adaptively changing throughout the iterative process, and have the role of reducing the variance introduced by compression. If no compression is used, we may simply set the perturbations to be $h_i^k = 0$ for all i, k. Variance reduction for compression operators was first develop by Hanzely et al. (2018) in the $n = 1$ case and for compression operators based on sketching. Our execution was inspired by Mishchenko et al. (2019a), who first studied variance reduction for CGD methods (for a specific ternary compression operator) in the $n > 1$ case, and Horváth et al. (2019b) who studied the distributed case for the general class of ω-compression operators we also study here. However, we had to make certain modifications to make variance reduction work in the accelerated case.

Algorithm 2 Accelerated DIANA (ADIANA)

Input: initial point x^0, $\{h_i^0\}_{i=1}^n$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$, parameters $\alpha, \eta, \theta_1, \theta_2, \beta, \gamma, p$

$z^0 = y^0 = w^0 = x^0$

for $k = 0, 1, 2, \ldots$ **do**

$x^k = \theta_1 z^k + \theta_2 w^k + (1 - \theta_1 - \theta_2) y^k$

for all machines $i = 1, 2, \ldots, n$ **do in parallel**

Compress shifted local gradient $\mathcal{C}_i^k(\nabla f_i(x^k) - h_i^k)$ and send to the server

Update local shift $h_i^{k+1} = h_i^k + \alpha \mathcal{C}_i^k(\nabla f_i(w^k) - h_i^k)$

end for

Aggregate received compressed gradient information

$$g^k = \frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(\nabla f_i(x^k) - h_i^k) + h^k$$

$$h^{k+1} = h^k + \alpha \frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(\nabla f_i(w^k) - h_i^k)$$

$y^{k+1} = \mathrm{prox}_{\eta\psi}(x^k - \eta g^k)$

$z^{k+1} = \beta z^k + (1-\beta)x^k + \frac{\gamma}{\eta}(y^{k+1} - x^k)$

$w^{k+1} = \begin{cases} y^k, & \text{with probability } p \\ w^k, & \text{with probability } 1-p \end{cases}$

end for

5.1 The ADIANA algorithm

We now propose an *accelerated* algorithm for solving problem (1). Our method combines both acceleration and variance reduction, and hence can be seen as an accelerated version of DIANA. hence, we call our method ADIANA (Algorithm 2).

Our method relies on a randomized update rule for the auxiliary vectors w^k, resembling the workings of the loopless SVRG method proposed by Kovalev et al. (2020).

5.2 Convergence theory

Our main convergence result for ADIANA (Algorithm 2) is formulated in Theorem 4. We focus on the strongly convex setting.

Theorem 4 *Suppose f is μ-strongly convex and that the functions f_i have L-Lipschitz continuous gradient for all i. Further, let the compression operator $\mathcal{C}$ satisfy* (5)*. Choose the ADIANA (Algorithm 2) parameters as follows:*

$$\alpha = \frac{1}{\omega+1}, \qquad \eta = \min\left\{\frac{1}{2L}, \frac{n}{64\omega(2p(\omega+1)+1)^2 L}\right\}, \qquad \theta_1 = \min\left\{\frac{1}{4}, \sqrt{\frac{\eta\mu}{p}}\right\}, \qquad \theta_2 = \frac{1}{2}$$

and

$$\gamma = \frac{\eta}{2(\theta_1 + \eta\mu)}, \qquad \beta = 1 - \gamma\mu, \qquad p = \min\left\{1, \frac{\max\{1, \sqrt{\frac{n}{32\omega}} - 1\}}{2(1+\omega)}\right\}.$$

Then the number of iterations performed by ADIANA to find an ϵ-solution such that

$$\mathbb{E}[\|z^k - x^*\|^2] \leq \epsilon$$

is at most

$$k = \begin{cases} O\left(\left[\omega + \omega\sqrt{\frac{L}{n\mu}}\right]\log\frac{1}{\epsilon}\right), & n \leq \omega, \\ O\left(\left[\omega + \sqrt{\frac{L}{\mu}} + \sqrt{\sqrt{\frac{\omega}{n}}\frac{\omega L}{\mu}}\right]\log\frac{1}{\epsilon}\right), & n > \omega. \end{cases}$$

As we have explained in the introduction, the above rate is vastly superior to that of non-accelerated distributed CGD methods, including that of DIANA.

5.3 Proof sketch

In the proof, we use the following notation:

$$\mathcal{Z}^k := \left\|z^k - x^*\right\|^2, \tag{6}$$

$$\mathcal{Y}^k := P(y^k) - P(x^*), \tag{7}$$

$$\mathcal{W}^k := P(w^k) - P(x^*), \tag{8}$$

$$\mathcal{H}^k := \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(w^k)\right\|^2. \tag{9}$$

We first present a key technical lemma which plays a similar role to that of Lemma 1.

Lemma 2 *If the parameters satisfy $\eta \leq \frac{1}{2L}, \theta_1 \leq \frac{1}{4}, \theta_2 = \frac{1}{2}, \gamma = \frac{\eta}{2(\theta_1+\eta\mu)}$ and $\beta = 1-\gamma\mu$, then we have for any iteration k,*

$$\frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[\mathcal{Y}^{k+1}\right] + \mathbb{E}\left[\mathcal{Z}^{k+1}\right] \leq (1-\theta_1-\theta_2)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \beta\mathcal{Z}^k$$

$$+ 2\gamma\beta\frac{\theta_2}{\theta_1}\mathcal{W}^k + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \tag{10}$$

$$- \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 \tag{11}$$

$$- \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2. \tag{12}$$

Theorem 4 can be proved by combing the above lemma with three additional Lemmas: Lemma 3, 4 and 5, which we present next. In view of the presence of $\mathcal{W}^k$ in (10), the following result is useful as it allows us to add $\mathcal{W}^{k+1}$ into the Lyapunov function.

Lemma 3 *According to Line 11 of Algorithm 2 and Definition* (7)–(8)*, we have*

$$\mathbb{E}\left[\mathcal{W}^{k+1}\right] = (1-p)\mathcal{W}^k + p\mathcal{Y}^k.$$

To cancel the term $\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$ in (10), we use the defining property of compression operator (i.e., (5)) :

Lemma 4 *If the compression operator $\mathcal{C}($ satisfies* (5)*, we have*

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq \frac{2\omega}{n^2}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \frac{2\omega}{n}\mathcal{H}^k. \tag{13}$$

Note that the bound on variance obtained above introduces an additional term $\mathcal{H}^k$ (see (13)). We will therefore add the terms $\mathcal{H}^{k+1}$ into the Lyapunov function as well.

Lemma 5 *If $\alpha \leq \frac{1}{\omega+1}$, we have*

$$\mathbb{E}\left[\mathcal{H}^{k+1}\right] \leq \left(1-\frac{\alpha}{2}\right)\mathcal{H}^k + \left(1+\frac{2p}{\alpha}\right)\frac{2p}{n}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \left(1+\frac{2p}{\alpha}\right)\frac{2p}{n}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2.$$

Note that the terms $\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2$ and $\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2$ in Lemma 5 and (13) can be cancelled by (11) and (12) by choosing the parameters appropriately.

Finally, it is not hard to obtain the following key inequality for the Lyapunov function by plugging Lemmas 3-5 into our key Lemma 2:

$$\mathbb{E}\left[c_1\mathcal{Y}^{k+1} + c_2\mathcal{Z}^{k+1} + c_3\mathcal{W}^{k+1} + c_4\mathcal{H}^{k+1}\right] \leq (1-c_5)\left(c_1\mathcal{Y}^k + c_2\mathcal{Z}^k + c_3\mathcal{W}^k + c_4\mathcal{H}^k\right). \tag{14}$$

Above, the constants $c_1, \ldots, c_5$ are related to the algorithm parameters $\alpha, \eta, \theta_1, \theta_2, \beta, \gamma$ and p. Finally, the proof of Theorem 4 can be derived (i.e., plug into the specified parameters) from inequality (14). The detailed proof can be found in the Supplementary Material.

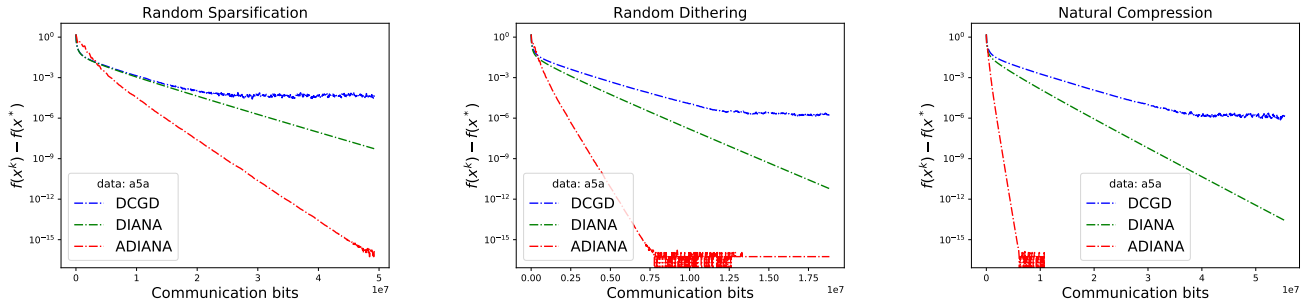

Figure 1: The communication complexity performance of DCGD vs DIANA vs ADIANA for three different compressors (random sparsification, random dithering and natural compression) on the `a5a` dataset with regularization constant $\lambda = 10^{-3}$.

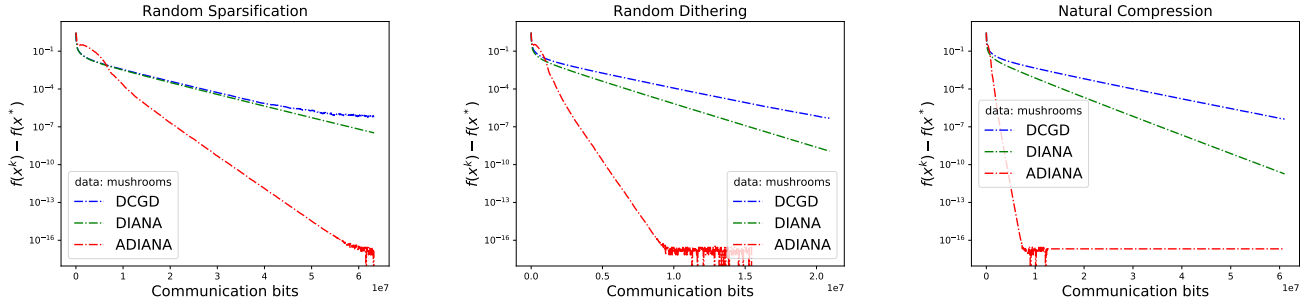

Figure 2: The communication complexity performance of DCGD vs DIANA vs ADIANA for three different compressors on the `mushrooms` dataset and regularization constant $\lambda = 10^{-3}$.

6 Experiments

We illustrate the performance of our accelerated DIANA method (ADIANA) with different compression operators on the logistic regression problem for binary classification,

$$f_i(x) := \log(1 + \exp(-b_i A_i^\top x)) + \frac{\lambda}{2}\|x\|^2,$$

where b_i and A_i are data points. In our experiments, we plot the relation of the optimality gap $P(x^k) - P(x^*)$ and the number of accumulated transmitted bits. The number of nodes is 20. The numerical results for different number of nodes can be found in the Appendix. The optimal solution for each case is obtained by getting the minimum of the three uncompressed versions of ADIANA, DIANA, and DCGD for 100000 iterations.

Data sets. In our experiments we use two datasets, namely, `a5a` and `mushrooms` from the LibSVM collection. Extensive experiments are provided in the Appendix.

Compression operators. We use three different compression operators: random sparsification (Alistarh et al., 2017), random dithering (Alistarh et al., 2017; Horváth et al., 2019a), and natural compression (Horváth et al., 2019a). For random-r sparsification, the number of communicated bits per iteration is $32r$, and we choose $r = d/4$. For random dithering, we choose $s = \sqrt{d}$, which means the number of communicated bits per iteration is $2.8d + 32$ (Alistarh et al., 2017). For natural compression, the number of communicated bits per iteration is $9d$ bits (Horváth et al., 2019a). In the following experiments, we use the theoretical stepsize and parameters for all the three algorithms: distributed compressed gradient descent (DCGD), DIANA, and ADIANA.

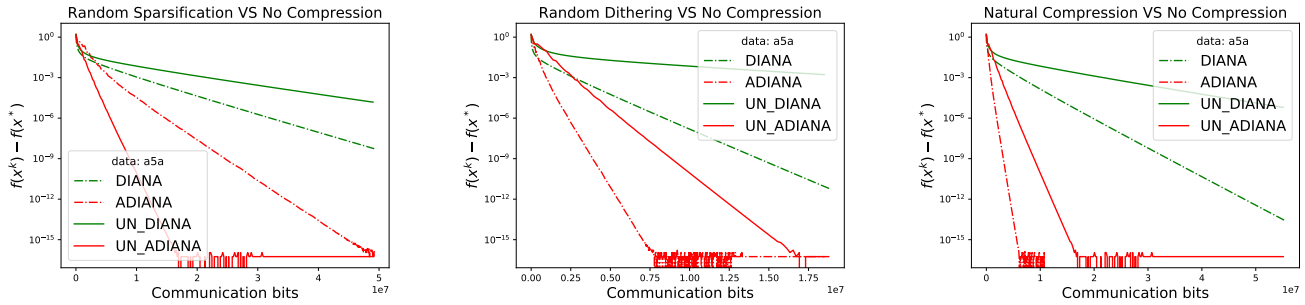

Figure 3: The communication complexity performance of DIANA and ADIANA with and without compression on `a5a` dataset and with regularization parameter $\lambda = 10^{-3}$.

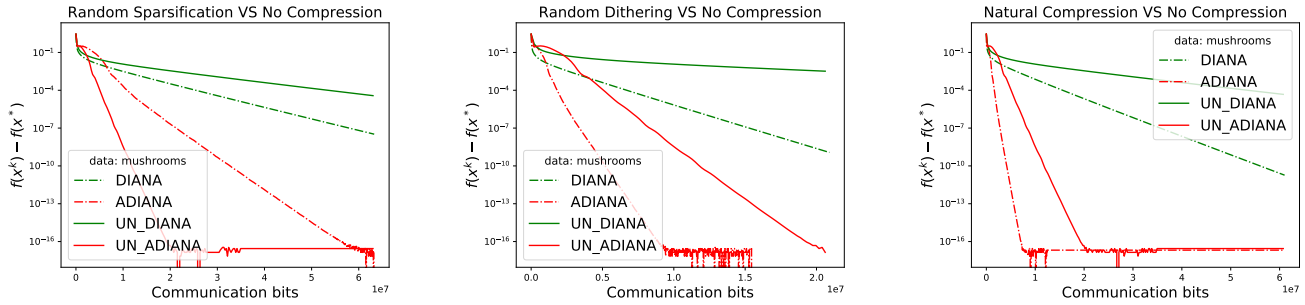

Figure 4: The communication complexity performance of DIANA and ADIANA with and without compression on `mushrooms` dataset and with regularization parameter $\lambda = 10^{-3}$.

6.1 Comparison with DIANA and DCGD

In this subsection, we compare ADIANA with DIANA and DCGD with three compression operators: random sparsification, random dithering, and natural compression in Figures 1 and 2. Figure 1 shows that, for dataset a5a and three compressions, ADIANA converges fastest except for the very early stage, and DCGD does not converge to the optimal solution because of the nonzero compression error. In Figure 2, we can see the similar performance for dataset mushrooms.

6.2 Communication efficiency

In this subsection, we compare ADIANA, DIANA, with their uncompressed versions to show the communication efficiency of our method. The numerical results for the two dataset are shown in Figures 3 and 4. For random-r sparsification, DIANA is better than unpressed version. However, ADIANA behaves worse than the uncompressed version. For random dithering and natural compression, ADIANA is about twice faster than the uncompressed version, and is much faster than DIANA. These numerical results indicate that ADIANA could be a more communication efficiency method, especially for random dithering and natural compression.

References

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

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017.

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In *The 33rd International Conference on Machine Learning*, pp. 1110–1119, 2016.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019a.

Hanzely, F. and Richtárik, P. One method to rule them all: variance reduction for data, parameters and many new methods. *arXiv preprint arXiv:1905.11266*, 2019b.

Hanzely, F., Mishchenko, K., and Richtárik, P. SEGA: variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems 31*, pp. 2082–2093, 2018.

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

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

Karimireddy, S., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local GD on heterogeneous data. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pp. 1–11, 2019.

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

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

Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. In *The 3rd International Conference on Learning Representations*, 2014. URL `https://arxiv.org/pdf/1412.6980.pdf`.

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

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pp. 10462–10472, 2019.

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

Ma, C., Konečný, J., Jaggi, M., Smith, V., Jordan, M. I., Richtárik, P., and Takáč, M. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017.

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

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019a.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019b.

Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pp. 543–547, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Qian, X., Qu, Z., and Richtárik, P. L-SVRG and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019.

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.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

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

Appendix

A Missing Proofs

In this appendix, we provide the detailed proofs for our Theorems 1–4.

A.1 Proof of Theorem 1

We first restate our Theorem 1 here and then provide the detailed proof.

Theorem 1 *Suppose $f(x)$ is convex with L-Lipschitz continuous gradient and the compression operator $\mathcal{C}(\cdot)$ satisfies (5). Let step size $\eta = \frac{1}{(1+\omega)L}$, then the number of iterations performed by CGD to find an ϵ-solution such that $\mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$ is at most $k = O(\frac{(1+\omega)L}{\epsilon})$.*

Proof: According to CGD update $x^{k+1} = x^k - \eta g^k$, we have

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] &= \mathbb{E}\left[\|x^k - \eta g^k - x^*\|^2\right] \\
&= \mathbb{E}\left[\|x^k - \eta \mathcal{C}(\nabla f(x^k)) - x^*\|^2\right] \\
&= \mathbb{E}\left[\|x^k - x^*\|^2 - 2\eta\langle \mathcal{C}(\nabla f(x^k)), x^k - x^*\rangle + \eta^2\|\mathcal{C}(\nabla f(x^k))\|^2\right] \\
&\overset{(5)}{=} \|x^k - x^*\|^2 - 2\eta\langle \nabla f(x^k), x^k - x^*\rangle + \eta^2\mathbb{E}\left[\|\mathcal{C}(\nabla f(x^k))\|^2\right] \\
&\overset{(5)}{=} \|x^k - x^*\|^2 - 2\eta\langle \nabla f(x^k), x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|^2 + \mathbb{E}\left[\|\mathcal{C}(\nabla f(x^k)) - \nabla f(x^k)\|^2\right] \\
&\overset{(5)}{\leq} \|x^k - x^*\|^2 - 2\eta\langle \nabla f(x^k), x^k - x^*\rangle + \eta^2(1+\omega)\|\nabla f(x^k)\|^2 \\
&\leq \|x^k - x^*\|^2 - 2\eta(f(x^k) - f(x^*)) + \eta^2(1+\omega)\|\nabla f(x^k)\|^2, \qquad (15)
\end{aligned}$$

where the last inequality holds due to convexity of f. Besides, according to L-smoothness of f (see (2)), we have

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] &\leq \mathbb{E}\left[f(x^k) - f(x^*) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2\right] \\
&= \mathbb{E}\left[f(x^k) - f(x^*) + \langle \nabla f(x^k), -\eta\mathcal{C}(\nabla f(x^k))\rangle + \frac{L\eta^2}{2}\|\mathcal{C}(\nabla f(x^k))\|^2\right] \\
&\overset{(5)}{=} \mathbb{E}\left[f(x^k) - f(x^*) - \eta\|\nabla f(x^k)\|^2 + \frac{L\eta^2}{2}\|\mathcal{C}(\nabla f(x^k))\|^2\right] \\
&\overset{(5)}{\leq} \mathbb{E}\left[f(x^k) - f(x^*) - \eta\left(1 - \frac{L\eta(1+\omega)}{2}\right)\|\nabla f(x^k)\|^2\right]. \qquad (16)
\end{aligned}$$

Adding (15) and $\frac{\eta(1+\omega)}{1-\frac{L\eta(1+\omega)}{2}}$ times (16) to cancel the term $\|\nabla f(x^k)\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\eta(1+\omega)}{1-\frac{L\eta(1+\omega)}{2}}\left(f(x^{k+1}) - f(x^*)\right) + \|x^{k+1} - x^*\|^2 + 2\eta\left(f(x^k) - f(x^*)\right)\right] \\
&\leq \mathbb{E}\left[\frac{\eta(1+\omega)}{1-\frac{L\eta(1+\omega)}{2}}\left(f(x^k) - f(x^*)\right) + \|x^k - x^*\|^2\right]. \qquad (17)
\end{aligned}$$

Summing up the above inequality from iteration 0 to k, we have

$$\begin{aligned}
\mathbb{E}\left[2\eta\sum_{i=0}^{k}\left(f(x^i) - f(x^*)\right)\right] &\leq \frac{\eta(1+\omega)}{1-\frac{L\eta(1+\omega)}{2}}\left(f(x^0) - f(x^*)\right) + \|x^0 - x^*\|^2 \\
&\leq \frac{\eta(1+\omega)}{1-\frac{L\eta(1+\omega)}{2}}\left(\frac{L}{2}\|x^0 - x^*\|^2\right) + \|x^0 - x^*\|^2 \\
&= \frac{2\|x^0 - x^*\|^2}{2 - L\eta(1+\omega)}, \qquad (18)
\end{aligned}$$

where the last inequality uses the L-smoothness of f. Finally, noting that $\mathbb{E}[f(x^{i+1}) - f(x^*)] \leq \mathbb{E}[f(x^i) - f(x^*)]$ for all $i = 0, \dots k$ according to (16), then (18) turns to be

$$\begin{aligned}\mathbb{E}\left[2\eta k\left(f(x^k) - f(x^*)\right)\right] &\leq \frac{2\|x^0 - x^*\|^2}{2 - L\eta(1+\omega)} \\ \mathbb{E}\left[f(x^k) - f(x^*)\right] &\leq \frac{\|x^0 - x^*\|^2}{(2 - L\eta(1+\omega))\,\eta k} = \frac{(1+\omega)L\|x^0 - x^*\|^2}{k},\end{aligned} \tag{19}$$

where the last equality uses the choice of step size $\eta = \frac{1}{(1+\omega)L}$. Now the proof of Theorem 1 is finished by letting the number of iteration $k = \frac{(1+\omega)L\|x^0 - x^*\|^2}{\epsilon}$, i.e., obtain the ϵ-solution x^k such that $\mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$ within $O(\frac{(1+\omega)L}{\epsilon})$ iterations. □

A.2 Proof of Theorems 2 and 3

First, we restate our key Lemma 1 here. Then we use it to prove Theorems 2 and 3. Finally, we provide the proof for this key Lemma 1.

Lemma 1 *If parameters $\{\eta_k\}, \{\theta_k\}, \{\beta_k\}$, $\{\gamma_k\}$ and p satisfy $\theta_k = \frac{1-\gamma_k/p}{1-\beta_k\gamma_k/p}$ $\beta_k \leq \min\{\frac{\mu\eta_k}{\gamma_k p}, 1\}$, $p \geq \frac{(1+L\eta_k)(1+\omega)}{2}$ and the compression operator $\mathcal{C}(\cdot)$ satisfies (5), then we have for any iteration k, $\forall x \in \mathbb{R}^d$,*

$$\frac{2\eta_k}{\gamma_k^2}\mathbb{E}[f(y^{k+1}) - f(x)] + \mathbb{E}[\|z^{k+1} - x\|^2] \leq \left(1 - \frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k) - f(x)\right) + (1-\beta_k)\|z^k - x\|^2, \tag{20}$$

where the expectation is with respect to the randomness of compression $\mathcal{C}(\cdot)$ at iteration k.

Now, we recall our Theorem 2 here and are ready to prove it using Lemma 1.

Theorem 2 *Suppose $f(x)$ is convex with L-Lipschitz continuous gradient and the compression operator $\mathcal{C}(\cdot)$ satisfies (5). Let parameters $\eta_k \equiv \frac{1}{L}$, $\theta_k = 1 - \frac{2}{k+2}$, $\beta_k \equiv 0$, $\gamma_k = \frac{2p}{k+2}$ and $p = 1+\omega$, then the number of iterations performed by ACGD (Algorithm 1) to find an ϵ-solution such that $\mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$ is at most $k = O\left((1+\omega)\sqrt{\frac{L}{\epsilon}}\right)$.*

Proof of Theorem 2. First, we know $\mu = 0$ in this general convex case. By choosing step size $\eta_k \equiv \frac{1}{L}$, $p = 1+\omega$, $\beta_k \equiv 0$, $\gamma_k = \frac{2p}{k+2}$ and $\theta_k = 1 - \frac{2}{k+2}$, Lemma 1 turns to be

$$\frac{(k+2)^2}{2L(1+\omega)^2}\mathbb{E}[f(y^{k+1}) - f(x)] + \mathbb{E}[\|z^{k+1} - x\|^2] \leq \frac{(k+1)^2 - 1}{2L(1+\omega)^2}\left(f(y^k) - f(x)\right) + \|z^k - x\|^2.$$

Summing up the above inequality from iteration 0 to $k-1$ and letting $x = x^*$, we have

$$\begin{aligned}\frac{(k+1)^2}{2L(1+\omega)^2}\mathbb{E}[f(y^k) - f(x^*)] &\leq \sum_{i=0}^{k-1}\frac{-1}{2L(1+\omega)^2}\left(f(y^i) - f(x^*)\right) + \|y^0 - x^*\|^2 \leq \|x^0 - x^*\|^2, \\ \mathbb{E}[f(y^k) - f(x^*)] &\leq \frac{2L(1+\omega)^2\|x^0 - x^*\|^2}{(k+1)^2},\end{aligned} \tag{21}$$

where the second inequality uses $f(y^i) - f(x^*) \geq 0$ and $y^0 = x^0$. Now the proof of Theorem 2 is finished, i.e., we obtain the ϵ-solution y^k such that $\mathbb{E}[f(y^k) - f(x^*)] \leq \epsilon$ within $k = O\left((1+\omega)\sqrt{\frac{L}{\epsilon}}\right)$ iterations. □

Similarly, we recall our Theorem 3 here and also prove it using Lemma 1.

Theorem 3 *Suppose $f(x)$ is μ-strongly convex with L-Lipschitz continuous gradient and the compression operator $\mathcal{C}(\cdot)$ satisfies (5). Let parameters $\eta_k \equiv \frac{1}{L}$, $\theta_k \equiv \frac{p}{p+\sqrt{\mu/L}}$, $\beta_k \equiv \frac{\sqrt{\mu/L}}{p}$, $\gamma_k \equiv \sqrt{\frac{\mu}{L}}$ and $p = 1+\omega$, then the number of iterations performed by ACGD (Algorithm 1) to find an ϵ-solution such that $\mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$ (or $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon$) is at most $k = O\left((1+\omega)\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$.*

Proof of Theorem 3. By choosing step size $\eta_k \equiv \frac{1}{L}$, $p = 1+\omega$, $\gamma_k \equiv \sqrt{\frac{\mu}{L}}$, $\beta_k \equiv \frac{\sqrt{\mu/L}}{p}$ and $\theta_k \equiv \frac{1}{1+\sqrt{\mu/p^2L}}$, Lemma 1 turns to be

$$\begin{aligned}\frac{2}{\mu}\mathbb{E}[f(y^{k+1}) - f(x)] + \mathbb{E}[\|z^{k+1} - x\|^2] &\leq \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)\frac{2}{\mu}\left(f(y^k) - f(x)\right) + \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)\|z^k - x\|^2 \\ &= \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)\left(\frac{2}{\mu}\left(f(y^k) - f(x)\right) + \|z^k - x\|^2\right).\end{aligned}$$

Telescoping the above inequality from iteration 0 to k and letting $x = x^*$, we have

$$\begin{aligned}\frac{2}{\mu}\mathbb{E}[f(y^k) - f(x^*)] + \mathbb{E}[\|z^k - x^*\|^2] &= \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)^k\left(\frac{2}{\mu}\left(f(y^0) - f(x^*)\right)\right) + \|z^0 - x^*\|^2\Big) \\ &\leq \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)^k\left(\frac{4}{\mu}\left(f(x^0) - f(x^*)\right)\right), \qquad (22)\\ \mathbb{E}[f(y^k) - f(x^*)] &\leq \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)^k\left(f(x^0) - f(x^*)\right), \\ \mathbb{E}[\|z^k - x^*\|^2] &\leq \left(1 - \frac{\sqrt{\mu/L}}{1+\omega}\right)^k\left(\frac{4}{\mu}\left(f(x^0) - f(x^*)\right)\right),\end{aligned}$$

where the first inequality uses μ-strongly convex of f (see (3)) and $y^0 = z^0 = x^0$. Now the proof of Theorem 3 is finished, i.e., we obtain the ϵ-solution y^k (or z^k) such that $\mathbb{E}[f(y^k) - f(x^*)] \leq \epsilon$ (or $\mathbb{E}[\|z^k - x^*\|^2] \leq \epsilon$) within $k = O\left((1+\omega)\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ iterations. □

Key Lemma: Now, the only remaining thing is to prove the key Lemma 1.

Proof of Lemma 1. First, we get the following equality (its proof is deferred to the end):

$$\begin{aligned}\forall x \in \mathbb{R}^d, \quad \mathbb{E}[\|z^{k+1} - x\|^2] &= (1-\beta_k)\|z^k - x\|^2 - \beta_k(1-\beta_k)\|x^k - z^k\|^2 + \beta_k\|x^k - x\|^2 + \frac{\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g^k\|^2 \\ &\quad + \frac{2\eta_k}{\gamma_k^2}\left\langle \nabla f(x^k), \left(1 - \frac{\gamma_k}{p}\right)(y^k - x^k) + \frac{\gamma_k}{p}(x - x^k)\right\rangle. \qquad (23)\end{aligned}$$

Then, to cancel the last inner product in (23), we use the property (smoothness and/or strong convexity) of f:

$$\begin{aligned}\mathbb{E}[f(y^{k+1})] &\leq \mathbb{E}\left[f(x^k) + \langle \nabla f(x^k), y^{k+1} - x^k\rangle + \frac{L}{2}\|y^{k+1} - x^k\|^2\right] \qquad (24)\\ &= \mathbb{E}\left[f(x^k) - \frac{\eta_k}{p}\langle \nabla f(x^k), g^k\rangle + \frac{L\eta_k^2}{2p^2}\|g^k\|^2\right] \\ &= f(x^k) - \frac{\eta_k}{p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^2}{2p^2}\mathbb{E}\|g^k\|^2 \\ &\leq f(u) - \langle \nabla f(x^k), u - x^k\rangle - \frac{\mu}{2}\|u - x^k\|^2 - \frac{\eta_k}{p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^2}{2p^2}\mathbb{E}\|g^k\|^2, \quad \forall u \in \mathbb{R}^d \qquad (25)\end{aligned}$$

where (24) and (25) uses L-smoothness (see (2)) and μ-strong convexity (see (3)), respectively. By adding $\left(1-\frac{\gamma_k}{p}\right)$ times (25) (where $u=y^k$) and $\frac{\gamma_k}{p}$ times (25) (where $u=x$), we have

$$\begin{aligned}\mathbb{E}[f(y^{k+1})] \leq \left(1-\frac{\gamma_k}{p}\right) f(y^k) + \frac{\gamma_k}{p} f(x) - \left\langle \nabla f(x^k), \left(1-\frac{\gamma_k}{p}\right)(y^k - x^k) + \frac{\gamma_k}{p}(x-x^k)\right\rangle \\ - \left(1-\frac{\gamma_k}{p}\right)\frac{\mu}{2}\|y^k-x^k\|^2 - \frac{\gamma_k\mu}{2p}\|x-x^k\|^2 - \frac{\eta_k}{p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^2}{2p^2}\mathbb{E}\|g^k\|^2. \end{aligned} \tag{26}$$

Now, this key lemma (i.e., (20)) is proved as follows by adding $\frac{2\eta_k}{\gamma_k^2}$ times (26) and (23):

$$\begin{aligned}&\frac{2\eta_k}{\gamma_k^2}\mathbb{E}[f(y^{k+1})-f(x)] + \mathbb{E}[\|z^{k+1}-x\|^2] \\ &\leq \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2 - \beta_k(1-\beta_k)\|x^k-z^k\|^2 - \left(\frac{\mu\eta_k}{\gamma_k p}-\beta_k\right)\|x^k-x\|^2 \\ &\quad - \frac{2\eta_k^2}{\gamma_k^2 p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^3+\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g_k\|^2 \\ &\leq \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2 - \frac{2\eta_k^2}{\gamma_k^2 p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^3+\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g_k\|^2 \qquad (27) \\ &= \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2 - \frac{2\eta_k^2}{\gamma_k^2 p}\|\nabla f(x^k)\|^2 + \frac{L\eta_k^3+\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|\mathcal{C}(\nabla f(x^k)\|^2 \\ &= \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2 - \frac{2\eta_k^2 p - L\eta_k^3-\eta_k^2}{\gamma_k^2 p^2}\|\nabla f(x^k)\|^2 \\ &\quad + \frac{L\eta_k^3+\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}[\|\mathcal{C}(\nabla f(x^k)) - \nabla f(x^k)\|^2] \qquad (28) \\ &\leq \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2 - \frac{2\eta_k^2 p - (1+\omega)(L\eta_k^3+\eta_k^2)}{\gamma_k^2 p^2}\|\nabla f(x^k)\|^2 \qquad (29) \\ &\leq \left(1-\frac{\gamma_k}{p}\right)\frac{2\eta_k}{\gamma_k^2}\left(f(y^k)-f(x)\right) + (1-\beta_k)\|z^k-x\|^2, \qquad (30)\end{aligned}$$

where (27) holds due to condition $\beta_k \leq \min\{\frac{\mu\eta_k}{\gamma_k p}, 1\}$, (28) and (29) use the property of compression (5), and (30) uses the condition $p \geq \frac{(1+L\eta_k)(1+\omega)}{2}$. Now, the only remaining thing is to prove (23). For any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\mathbb{E}[\|z^{k+1}-x\|^2] &= \mathbb{E}\left[\left\|\frac{1}{\gamma_k}y^{k+1} + \left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k + \left(1-\frac{1}{p}\right)(1-\beta_k)z^k + \left(1-\frac{1}{p}\right)\beta_k x^k - x\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k + \left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k + \left(1-\frac{1}{p}\right)(1-\beta_k)z^k - x - \frac{\eta_k}{\gamma_k p}g^k\right\|^2\right] \\ &= \left\|\left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k + \left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k + \left(1-\frac{1}{p}\right)(1-\beta_k)z^k - x\right\|^2 + \frac{\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g^k\|^2 \\ &\quad + \frac{2\eta_k}{\gamma_k p}\left\langle \mathbb{E}[g^k], x - \left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k - \left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k - \left(1-\frac{1}{p}\right)(1-\beta_k)z^k\right\rangle \\ &= (1-\beta_k)\|z^k-x\|^2 - \beta_k(1-\beta_k)\|x^k-z^k\|^2 + \beta_k\|x^k-x\|^2 + \frac{\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g^k\|^2 \qquad (31) \\ &\quad + \frac{2\eta_k}{\gamma_k p}\left\langle \mathbb{E}[g^k], x - \left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k - \left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k - \left(1-\frac{1}{p}\right)(1-\beta_k)z^k\right\rangle \\ &= (1-\beta_k)\|z^k-x\|^2 - \beta_k(1-\beta_k)\|x^k-z^k\|^2 + \beta_k\|x^k-x\|^2 + \frac{\eta_k^2}{\gamma_k^2 p^2}\mathbb{E}\|g^k\|^2 \\ &\quad + \frac{2\eta_k}{\gamma_k^2}\left\langle \nabla f(x^k), \left(1-\frac{\gamma_k}{p}\right)(y^k-x^k) + \frac{\gamma_k}{p}(x-x^k)\right\rangle, \qquad (32)\end{aligned}$$

where (31) and (32) use equalities (36) and (40), respectively. Further,

$$
\begin{aligned}
&\left\|\left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k+\left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2\\
&=\left\|\left(\frac{1}{\gamma_k}-\left(1-\frac{1}{p}\right)\beta_k\right)x^k-\frac{1}{\gamma_k}\left(1-\frac{\gamma_k}{p}\right)y^k+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2\\
&=\left\|\left(\frac{1}{\gamma_k}-\left(1-\frac{1}{p}\right)\beta_k\right)x^k-\frac{1}{\gamma_k}\left(1-\frac{\beta_k\gamma_k}{p}\right)\theta_k y^k+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2 &(33)\\
&=\left\|\left(\frac{1}{\gamma_k}-\left(1-\frac{1}{p}\right)\beta_k\right)x^k-\frac{1}{\gamma_k}\left(1-\frac{\beta_k\gamma_k}{p}\right)(x^k-(1-\theta_k)z^k)+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2 &(34)\\
&=\left\|\beta_k x^k+\frac{1}{\gamma_k}\left(1-\frac{\beta_k\gamma_k}{p}\right)(1-\theta_k)z^k+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2\\
&=\left\|\beta_k x^k+\frac{1}{p}(1-\beta_k)z^k+\left(1-\frac{1}{p}\right)(1-\beta_k)z^k-x\right\|^2 &(35)\\
&=\|z^k-x+\beta_k(x^k-z^k)\|^2\\
&=\|z^k-x\|^2+\beta_k^2\|x^k-z^k\|^2+2\beta_k\langle x^k-z^k,z^k-x\rangle\\
&=(1-\beta_k)\|z^k-x\|^2-\beta_k(1-\beta_k)\|x^k-z^k\|^2+\beta_k\|x^k-x\|^2, &(36)
\end{aligned}
$$

where (33) and (35) hold due to the condition $\theta_k=\frac{1-\gamma_k/p}{1-\beta_k\gamma_k/p}$, and (34) holds due to the relation $x^k=\theta_k y^k+(1-\theta_k)z^k$ (Line 3 in Algorithm 1). Now, we finish the proof for the inner product term.

$$
\begin{aligned}
&\frac{2\eta_k}{\gamma_k p}\left\langle\nabla f(x^k),x-\left(\frac{1}{\gamma_k}+\left(1-\frac{1}{p}\right)\beta_k\right)x^k-\left(\frac{1}{p}-\frac{1}{\gamma_k}\right)y^k-\left(1-\frac{1}{p}\right)(1-\beta_k)z^k\right\rangle\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\frac{\gamma_k}{p}x-\left(\frac{1}{p}+\left(1-\frac{1}{p}\right)\frac{\beta_k\gamma_k}{p}\right)x^k+\frac{1}{p}\left(1-\frac{\gamma_k}{p}\right)y^k-\left(1-\frac{1}{p}\right)(1-\beta_k)\frac{\gamma_k}{p}z^k\right\rangle\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\frac{\gamma_k}{p}x-\left(\frac{1}{p}+\left(1-\frac{1}{p}\right)\frac{\beta_k\gamma_k}{p}\right)x^k+\frac{1}{p}\left(1-\frac{\gamma_k}{p}\right)y^k-\left(1-\frac{1}{p}\right)\left(1-\frac{\beta_k\gamma_k}{p}\right)(1-\theta_k)z^k\right\rangle &(37)\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\frac{\gamma_k}{p}x-\left(\frac{1}{p}+\left(1-\frac{1}{p}\right)\frac{\beta_k\gamma_k}{p}\right)x^k+\frac{1}{p}\left(1-\frac{\gamma_k}{p}\right)y^k-\left(1-\frac{1}{p}\right)\left(1-\frac{\beta_k\gamma_k}{p}\right)(x^k-\theta_k y^k)\right\rangle &(38)\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\frac{\gamma_k}{p}x-x^k+\frac{1}{p}\left(1-\frac{\gamma_k}{p}\right)y^k+\left(1-\frac{1}{p}\right)\left(1-\frac{\beta_k\gamma_k}{p}\right)\theta_k y^k\right\rangle\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\frac{\gamma_k}{p}x-x^k+\frac{1}{p}\left(1-\frac{\gamma_k}{p}\right)y^k+\left(1-\frac{1}{p}\right)\left(1-\frac{\gamma_k}{p}\right)y^k\right\rangle &(39)\\
&=\frac{2\eta_k}{\gamma_k^2}\left\langle\nabla f(x^k),\left(1-\frac{\gamma_k}{p}\right)(y^k-x^k)+\frac{\gamma_k}{p}(x-x^k)\right\rangle, &(40)
\end{aligned}
$$

where (37) and (39) hold due to the condition $\theta_k=\frac{1-\gamma_k/p}{1-\beta_k\gamma_k/p}$, and (38) holds due to the relation $x^k=\theta_k y^k+(1-\theta_k)z^k$ (Line 3 in Algorithm 1). □

A.3 Proof of Theorem 4

In this section, we provide the detailed proof for accelerated result in the distributed case. Similar to previous Section A.2, we first restate our Lemmas 2–5 here. Then we use them to prove Theorem 4. Finally, we provide the proof for these Lemmas. Before restating lemmas, we recall the following notation:

$$\mathcal{Z}^k := \left\|z^k - x^*\right\|^2, \tag{41}$$

$$\mathcal{Y}^k := P(y^k) - P(x^*), \tag{42}$$

$$\mathcal{W}^k := P(w^k) - P(x^*), \tag{43}$$

$$\mathcal{H}^k := \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(w^k)\right\|^2. \tag{44}$$

Lemma 2 *If parameters satisfy $\eta \le \frac{1}{2L}, \theta_1 \le \frac{1}{4}, \theta_2 = \frac{1}{2}, \gamma = \frac{\eta}{2(\theta_1+\eta\mu)}$ and $\beta = 1-\gamma\mu$, then we have for any iteration k,*

$$\begin{aligned}\frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[\mathcal{Y}^{k+1}\right] + \mathbb{E}\left[\mathcal{Z}^{k+1}\right] \le (1-\theta_1-\theta_2)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \beta\mathcal{Z}^k + 2\gamma\beta\frac{\theta_2}{\theta_1}\mathcal{W}^k + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\ - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2.\end{aligned} \tag{45}$$

Lemma 3 *According to Line 11 of Algorithm 2 and Definition (42)–(43), we have*

$$\mathbb{E}\left[\mathcal{W}^{k+1}\right] = (1-p)\mathcal{W}^k + p\mathcal{Y}^k. \tag{46}$$

Lemma 4 *If the compression operator $\mathcal{C}(\cdot)$ satisfies (5), we have*

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \le \frac{2\omega}{n^2}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \frac{2\omega}{n}\mathcal{H}^k \tag{47}$$

Lemma 5 *If $\alpha \le 1/(1+\omega)$, we have*

$$\mathbb{E}\left[\mathcal{H}^{k+1}\right] \le (1-\frac{\alpha}{2})\mathcal{H}^k + (1+\frac{2p}{\alpha})\frac{2p}{n}\left(\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2\right). \tag{48}$$

Now, we recall our Theorem 4 here and are ready to prove it using Lemmas 2–5.

Theorem 4 *Suppose $f(x)$ is μ-strongly convex and all f_is have L-Lipschitz continuous gradients, and the compression operator $\mathcal{C}(\cdot)$ satisfies (5). Let parameters $\alpha = \frac{1}{\omega+1}$, $\eta = \min\{\frac{1}{2L}, \frac{n}{64\omega(2p(\omega+1)+1)^2L}\}$, $\theta_1 = \min\left\{\frac{1}{4}, \sqrt{\frac{\eta\mu}{p}}\right\}$, $\theta_2 = \frac{1}{2}$, $\gamma = \frac{\eta}{2(\theta_1+\eta\mu)}$, $\beta = 1-\gamma\mu$, and $p = \min\left\{1, \frac{\max\{1,\sqrt{n/32\omega}-1\}}{2(1+\omega)}\right\}$, then the number of iterations performed by ADIANA (Algorithm 2) to find an ϵ-solution such that $\mathbb{E}[\|z^k - x^*\|^2] \le \epsilon$ is at most*

$$k = \begin{cases} O\left(\left[\omega + \omega\sqrt{\frac{L}{n\mu}}\right]\log\frac{1}{\epsilon}\right), & n \le \omega, \\ O\left(\left[\omega + \sqrt{\frac{L}{\mu}} + \sqrt{\sqrt{\frac{\omega}{n}}\frac{\omega L}{\mu}}\right]\log\frac{1}{\epsilon}\right), & n > \omega. \end{cases}$$

Proof of Theorem 4. We define the following Lyapunov function Ψ and induce it as follows: $-$smooth :

$$\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] &:= \mathbb{E}\left[\mathcal{Z}^{k+1} + \frac{2\gamma\beta}{\theta_1}\mathcal{Y}^{k+1} + 2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^{k+1} + \frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^{k+1}\right] \\
&\leq \beta\mathcal{Z}^k + (1-\theta_1-\theta_2)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + 2\gamma\beta\frac{\theta_2}{\theta_1}\mathcal{W}^k + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&\quad - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad + \mathbb{E}\left[2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^{k+1} + \frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^{k+1}\right] \qquad (49) \\
&= \beta\mathcal{Z}^k + (1-\theta_1-\theta_2)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + 2\gamma\beta\frac{\theta_2}{\theta_1}\mathcal{W}^k + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&\quad - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad + 2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}(1-p)\mathcal{W}^k + 2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1}\mathcal{Y}^k + \mathbb{E}\left[\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^{k+1}\right] \qquad (50) \\
&\leq \beta\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k \\
&\quad - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] + \mathbb{E}\left[\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^{k+1}\right] \qquad (51) \\
&\leq \beta\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k \\
&\quad - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad + \frac{2\gamma\eta\omega}{\theta_1 n^2}\left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \frac{2\gamma\eta\omega}{\theta_1 n}\mathcal{H}^k + \mathbb{E}\left[\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^{k+1}\right] \qquad (52) \\
&\leq \beta\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k \\
&\quad - \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad + \frac{2\gamma\eta\omega}{\theta_1 n^2}\left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \frac{2\gamma\eta\omega}{\theta_1 n}\mathcal{H}^k + \frac{8\gamma\eta\omega}{\alpha\theta_1 n}\left(1-\frac{\alpha}{2}\right)\mathcal{H}^k \\
&\quad + \left(1+\frac{2p}{\alpha}\right)\frac{16\gamma\eta\omega p}{\alpha\theta_1 n^2}\left(\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2\right) \qquad (53) \\
&= \beta\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k + \left(1-\frac{\alpha}{4}\right)\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^k \\
&\quad - \frac{\gamma}{n\theta_1}\left(\frac{1}{8L} - \frac{2\eta\omega}{n}\right)\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 \\
&\quad - \frac{\gamma}{n\theta_1}\left(\frac{1}{8L} - (1+\frac{2p}{\alpha})\frac{16\eta\omega p}{\alpha n}\right)\left(\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2\right) \\
&\leq \beta\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k + \left(1-\frac{\alpha}{4}\right)\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^k \qquad (54) \\
&\leq (1-\frac{\eta\mu}{4\theta_1})\mathcal{Z}^k + \left(1-\frac{\theta_1}{2}\right)\frac{2\gamma\beta}{\theta_1}\mathcal{Y}^k + \left(1-\frac{\theta_1 p}{2}\right)2\gamma\beta\frac{\theta_2(1+\theta_1)}{\theta_1 p}\mathcal{W}^k + \left(1-\frac{\alpha}{4}\right)\frac{8\gamma\eta\omega}{\alpha\theta_1 n}\mathcal{H}^k \qquad (55) \\
&\leq \left(1 - \min\left\{\frac{\alpha}{4}, \frac{p}{8}, \frac{\sqrt{\eta\mu p}}{4}\right\}\right)\Psi^k, \qquad (56)
\end{aligned}$$

where (49) uses Lemma 2, (50) uses Lemma 3, (51) uses $\theta_1 \leq 1/4$ and $\theta_2 = 1/2$, (52) uses Lemma 4, (53) uses Lemma 5, (54) uses $\eta = \min\{\frac{1}{2L}, \frac{n}{64\omega(2p(\omega+1)+1)^2 L}\}$, (55) uses $\gamma = \frac{\eta}{2(\theta_1+\eta\mu)}$, $\beta = 1-\gamma\mu \leq 1 - \frac{\eta\mu}{4\theta_1}$ due to $\eta\mu \leq \theta_1$, and (56) uses $\theta_1 = \min\left\{\frac{1}{4}, \sqrt{\frac{\eta\mu}{p}}\right\}$.

Telescoping the above inequality (56) from iteration 0 to k, we have $\mathbb{E}[\Psi^k] \leq \left(1 - \min\left\{\frac{\alpha}{4}, \frac{p}{8}, \frac{\sqrt{\eta\mu p}}{4}\right\}\right)^k \Psi^0$. To obtain an ϵ-solution z^k such that $\mathbb{E}[\|z^k - x^*\|^2] \leq \epsilon$, the number of iterations is at most

$$\begin{aligned} k &= \max\left\{\frac{4}{\alpha}, \frac{8}{p}, \frac{4}{\sqrt{\eta\mu p}}\right\} \log\frac{\Psi^0}{\epsilon} \\ &= \max\left\{4(1+\omega), \frac{8}{p}, 4\sqrt{\frac{L}{\mu}\max\left\{\frac{2}{p}, \frac{64\omega(2p(\omega+1)+1)^2}{np}\right\}}\right\} \log\frac{\Psi^0}{\epsilon} \end{aligned} \tag{57}$$

By letting $p = \min\left\{1, \frac{\max\{1, \sqrt{n/32\omega}-1\}}{2(1+\omega)}\right\}$, it is not hard to verify that the number of iterations performed by ADIANA (Algorithm 2) to find an ϵ-solution such that $\mathbb{E}[\|z^k - x^*\|^2] \leq \epsilon$ is at most

$$k = \begin{cases} O\left(\left[\omega + \omega\sqrt{\frac{L}{n\mu}}\right] \log\frac{1}{\epsilon}\right), & n \leq \omega, \\ O\left(\left[\omega + \sqrt{\frac{L}{\mu}} + \sqrt{\sqrt{\frac{\omega}{n}}\frac{\omega L}{\mu}}\right] \log\frac{1}{\epsilon}\right), & n > \omega, \end{cases}$$

where n is the number of parallel machines. □

Key Lemmas: Now, the remaining thing is to prove Lemmas 2–5. We first prove the relatively simple Lemmas 3–5 and then prove the key Lemma 2.

Proof of Lemma 3. According to Line 11 of Algorithm 2, i.e., $w^{k+1} = \begin{cases} y^k, & \text{with probability } p \\ w^k, & \text{with probability } 1-p \end{cases}$, and definitions $\mathcal{Y}^k := P(y^k) - P(x^*)$ and $\mathcal{W}^k := P(w^k) - P(x^*)$, this lemma is directly obtained, i.e.,

$$\mathbb{E}\left[\mathcal{W}^{k+1}\right] = (1-p)\mathcal{W}^k + p\mathcal{Y}^k. \tag{58}$$

□

Proof of Lemma 4. This lemma is proved as follows:

$$\begin{aligned} \mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{C}(\nabla f_i(x^k) - h_i^k) + h_i^k - \nabla f_i(x^k)\right\|^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathcal{C}(\nabla f_i(x^k) - h_i^k) + h_i^k - \nabla f_i(x^k)\right\|^2\right] \\ &\leq \frac{\omega}{n^2}\sum_{i=1}^n \left\|\nabla f_i(x^k) - h_i^k\right\|^2 \\ &\leq \frac{2\omega}{n^2}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 + \frac{2\omega}{n}\mathcal{H}^k, \end{aligned}$$

where first inequality uses the property of ω-compression operator (i.e., (5)) and the last inequality uses Cauchy-Schwarz inequality and definition $\mathcal{H}^k := \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(w^k)\right\|^2$. □

Proof of Lemma 5. This lemma is proved as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{H}^{k+1}\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_i^{k+1}-\nabla f_i(w^{k+1})\right\|^2\right] \\
&= \frac{p}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_i^{k+1}-\nabla f_i(y^k)\right\|^2\right]+\frac{1-p}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_i^{k+1}-\nabla f_i(w^k)\right\|^2\right] &(59)\\
&\le \left(1+\frac{2p}{\alpha}\right)\frac{p}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w^k)-\nabla f_i(y^k)\right\|^2+\left(\frac{1-p}{n}+\left(1+\frac{\alpha}{2p}\right)\frac{p}{n}\right)\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_i^{k+1}-\nabla f_i(w^k)\right\|^2\right] &(60)\\
&\le \left(1+\frac{2p}{\alpha}\right)\frac{p}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w^k)-\nabla f_i(y^k)\right\|^2+\left(\frac{1+\alpha/2}{n}\right)(1-\alpha)\sum_{i=1}^{n}\left\|h_i^k-\nabla f_i(w^k)\right\|^2 &(61)\\
&\le \left(1+\frac{2p}{\alpha}\right)\frac{p}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w^k)-\nabla f_i(y^k)\right\|^2+\left(1-\frac{\alpha}{2}\right)\mathcal{H}^k &(62)\\
&\le \left(1-\frac{\alpha}{2}\right)\mathcal{H}^k+\left(1+\frac{2p}{\alpha}\right)\frac{2p}{n}\left(\sum_{i=1}^{n}\left\|\nabla f_i(w^k)-\nabla f_i(x^k)\right\|^2+\sum_{i=1}^{n}\left\|\nabla f_i(y^k)-\nabla f_i(x^k)\right\|^2\right), &(63)
\end{aligned}$$

where (59) uses the definition of w^{k+1} (see Line 11 of Algorithm 2), (60) uses Cauchy-Schwarz inequality, (61) uses $h_i^{k+1}=h_i^k+\alpha\mathcal{C}(\nabla f_i(w^k)-h_i^k)$, the property of ω-compression operator (i.e., (5)) and $\alpha\le 1/(1+\omega)$, and (63) uses Cauchy-Schwarz inequality. □

Proof of Lemma 2. Similar to the proof of key Lemma 1 in the last section, we first get the following equality:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^{k+1}\right] &= \mathbb{E}\left[\left\|\beta z^k+(1-\beta)x^k-x^*+\frac{\gamma}{\eta}(y^{k+1}-x^k)\right\|^2\right] \\
&= \left\|\beta(z^k-x^*)+(1-\beta)(x^k-x^*)\right\|^2+\mathbb{E}\left[\frac{2\gamma}{\eta}\langle y^{k+1}-x^k,\beta(z^k-x^k)+x^k-x^*\rangle\right]+\frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\le \beta\mathcal{Z}^k+(1-\beta)\left\|x^k-x^*\right\|^2-\beta(1-\beta)\|x^k-z^k\|^2+\frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad+\mathbb{E}\left[\frac{2\gamma}{\eta}\langle y^{k+1}-x^k,x^k-x^*\rangle+\frac{2\gamma\beta}{\eta}\langle y^{k+1}-x^k,z^k-x^k\rangle\right] \\
&= \beta\mathcal{Z}^k+(1-\beta)\left\|x^k-x^*\right\|^2-\beta(1-\beta)\|x^k-z^k\|^2+\frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad+\mathbb{E}\left[\frac{2\gamma}{\eta}\langle x^k-y^{k+1},x^*-x^k\rangle+\frac{2\gamma\beta\theta_2}{\eta\theta_1}\langle x^k-y^{k+1},w^k-x^k\rangle+\frac{2\gamma\beta(1-\theta_1-\theta_2)}{\eta\theta_1}\langle x^k-y^{k+1},y^k-x^k\rangle\right], &(64)
\end{aligned}$$

where (64) uses $x^k=\theta_1 z^k+\theta_2 w^k+(1-\theta_1-\theta_2)y^k$ (see Line 3 in Algorithm 2). To cancel the inner products in (64), we first use the property of f:

$$\begin{aligned}
\mathbb{E}[f(y^{k+1})] &\le \mathbb{E}\left[f(x^k)+\langle\nabla f(x^k),y^{k+1}-x^k\rangle+\frac{L}{2}\|y^{k+1}-x^k\|^2\right] \\
&\le \mathbb{E}\left[\langle\nabla f(x^k),y^{k+1}-x^k\rangle+\frac{L}{2}\|y^{k+1}-x^k\|^2\right] \\
&\qquad+f(u)-\langle\nabla f(x^k),u-x^k\rangle-\max\left\{\frac{\mu}{2}\|u-x^k\|^2,\frac{1}{2Ln}\sum_{i=1}^{n}\|\nabla f_i(u)-\nabla f_i(x^k)\|^2\right\},\quad\forall u\in\mathbb{R}^d &(65)
\end{aligned}$$

where the inequalities hold uses the L-smoothness and μ-strong convexity of f and $f(x):=\frac{1}{n}\sum_{i=1}^{n}f_i(x)$.

Then, according the definition of y^{k+1} (see Line 9 of Algorithm 2), we have

$$y^{k+1}=x^k-\eta g^k-\eta\Delta, \tag{66}$$

where $\Delta \in \partial\psi(y^{k+1})$. Besides, according to convexity of ψ, we have

$$\mathbb{E}[\psi(y^{k+1})] \leq \mathbb{E}\left[\psi(u) - \langle \Delta, u - y^{k+1}\rangle\right] = \mathbb{E}\left[\psi(u) - \langle \Delta, u - x^k\rangle + \langle \Delta, y^{k+1} - x^k\rangle\right], \quad \forall u \in \mathbb{R}^d. \tag{67}$$

Adding (65) and (67), we have

$$\begin{aligned}
\forall u \in \mathbb{R}^d, \quad \mathbb{E}[P(y^{k+1})] &\leq \mathbb{E}\left[P(u) - \langle \Delta + g^k, u - x^k\rangle + \langle \Delta + \nabla f(x^k), y^{k+1} - x^k\rangle + \frac{L}{2}\|y^{k+1} - x^k\|^2\right] \\
&\quad - \max\left\{\frac{\mu}{2}\|u - x^k\|^2, \frac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(u) - \nabla f_i(x^k)\|^2\right\} \\
&= \mathbb{E}\left[P(u) - \frac{1}{\eta}\langle x^k - y^{k+1}, u - x^k\rangle + \langle \nabla f(x^k) - g^k, y^{k+1} - x^k\rangle - \frac{1}{\eta}\|y^{k+1} - x^k\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{L}{2}\|y^{k+1} - x^k\|^2\right] - \max\left\{\frac{\mu}{2}\|u - x^k\|^2, \frac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(u) - \nabla f_i(x^k)\|^2\right\} \qquad (68) \\
&\leq \mathbb{E}\left[P(u) - \frac{1}{\eta}\langle x^k - y^{k+1}, u - x^k\rangle + \frac{\eta}{2}\|\nabla f(x^k) - g^k\|^2 - \frac{1}{2\eta}\|y^{k+1} - x^k\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{L}{2}\|y^{k+1} - x^k\|^2\right] - \max\left\{\frac{\mu}{2}\|u - x^k\|^2, \frac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(u) - \nabla f_i(x^k)\|^2\right\} \qquad (69) \\
&\leq \mathbb{E}\left[P(u) - \frac{1}{\eta}\langle x^k - y^{k+1}, u - x^k\rangle + \frac{\eta}{2}\|\nabla f(x^k) - g^k\|^2 - \frac{1}{4\eta}\|y^{k+1} - x^k\|^2\right] \\
&\quad - \max\left\{\frac{\mu}{2}\|u - x^k\|^2, \frac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(u) - \nabla f_i(x^k)\|^2\right\}, \qquad (70)
\end{aligned}$$

where (68) uses (66), (69) uses Young's inequality, and (70) uses the condition $\eta \leq 1/2L$.

Now, we are ready to prove this lemma by canceling the inner products in (64) using (70). By plugging 2γ times (70) (where $u = x^*$), $\frac{2\gamma\beta\theta_2}{\theta_1}$ times (70) (where $u = w^k$), and $\frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}$ times (70) (where $u = y^k$) into (64), we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^k\right] &\leq \beta\mathcal{Z}^k + (1-\beta)\left\|x^k - x^*\right\|^2 - \beta(1-\beta)\|x^k - z^k\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1} - x^k\right\|^2\right] \\
&\quad + 2\gamma\mathbb{E}\left[P(x^*) - P(y^{k+1}) + \frac{\eta}{2}\left\|\nabla f(x^k) - g^k\right\|^2 - \frac{1}{4\eta}\left\|y^{k+1} - x^k\right\|^2 - \frac{\mu}{2}\left\|x^k - x^*\right\|^2\right] \\
&\quad + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\Big[P(w^k) - P(y^{k+1}) + \frac{\eta}{2}\left\|\nabla f(x^k) - g^k\right\|^2 - \frac{1}{4\eta}\left\|y^{k+1} - x^k\right\|^2 \\
&\qquad\qquad - \frac{1}{2Ln}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2\Big] \\
&\quad + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\Big[P(y^k) - P(y^{k+1}) + \frac{\eta}{2}\left\|\nabla f(x^k) - g^k\right\|^2 - \frac{1}{4\eta}\left\|y^{k+1} - x^k\right\|^2] \\
&\qquad\qquad - \frac{1}{2Ln}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2\Big] \\
&= \beta\mathcal{Z}^k + (1-\beta-\gamma\mu)\left\|x^k - x^*\right\|^2 - \beta(1-\beta)\|x^k - z^k\|^2 + \left(\frac{\gamma^2}{\eta^2} - \frac{\gamma\beta}{2\eta\theta_1}\right)\mathbb{E}\left[\left\|y^{k+1} - x^k\right\|^2\right] \\
&\quad + \frac{2\gamma\beta\theta_2}{\theta_1}\mathcal{W}^k + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathcal{Y}^k - \frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[\mathcal{Y}^{k+1}\right] \\
&\quad - 2\gamma(1-\beta)\mathbb{E}\left[\mathcal{Y}^{k+1}\right] + \left(\gamma\eta + \gamma\beta\eta\frac{1-\theta_1}{\theta_1}\right)\mathbb{E}\left[\left\|\nabla f(x^k) - g^k\right\|^2\right] \\
&\quad - \frac{\gamma\beta\theta_2}{Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma\beta(1-\theta_1-\theta_2)}{Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2
\end{aligned}$$

$$\leq \beta \mathcal{Z}^k + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathcal{Y}^k + \frac{2\gamma\beta\theta_2}{\theta_1}\mathcal{W}^k - \frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[\mathcal{Y}^{k+1}\right] + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|\nabla f(x^k) - g^k\right\|^2\right]$$
$$- \frac{\gamma\theta_2}{2Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma(1-\theta_1-\theta_2)}{2Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2 \tag{71}$$
$$\leq \beta \mathcal{Z}^k + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathcal{Y}^k + \frac{2\gamma\beta\theta_2}{\theta_1}\mathcal{W}^k - \frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[\mathcal{Y}^{k+1}\right] + \frac{\gamma\eta}{\theta_1}\mathbb{E}\left[\left\|\nabla f(x^k) - g^k\right\|^2\right]$$
$$- \frac{\gamma}{4Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^k)\right\|^2 - \frac{\gamma}{8Ln\theta_1}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^k)\right\|^2, \tag{72}$$

where (71) uses $\gamma = \frac{\eta}{2(\theta_1+\eta\mu)}$ and $\beta = 1-\gamma\mu$, and (72) uses $\theta_1 \leq 1/4$ and $\theta_2 = 1/2$. □

B Extra Experiments

In this section, we conduct more experiments on ADIANA with different compression operators for the logistic regression problem. The two tested datasets are a9a and w6a. For random-r sparsification, we choose $r = d/4$. For random dithering, we choose $s = \sqrt{d}$. For all methods, we use the theoretical stepsize.

B.1 Comparison with DIANA and DCGD

In this subsection, we compare ADIANA with DIANA and DCGD with three compression operators: random sparsification, random dithering, and natural compression in Figure 5 and Figure 6. The number of nodes in our experiments is 20. We can see ADIANA is fater than DIANA and DCGD. Furthermore, because the compression error of DCGD is nonzero in general, DCGD can only converge to the neighborhood of the optimal solution. While, DIANA and ADIANA can converge to the optimal solution.

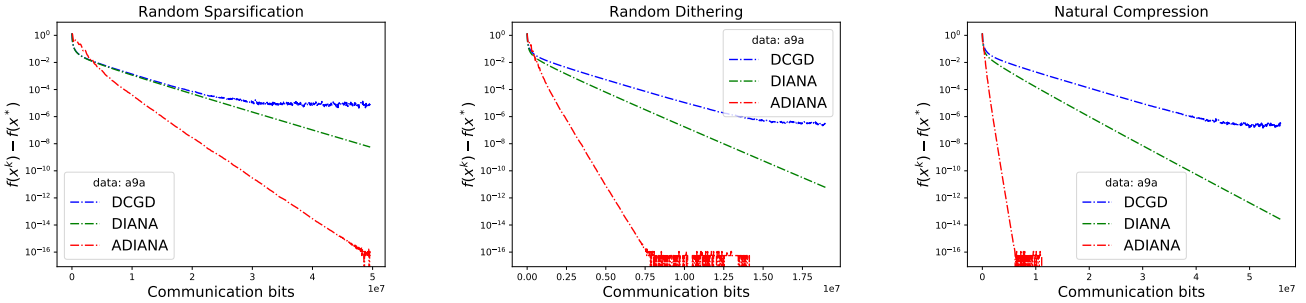

Figure 5: The communication complexity performance of DCGD vs DIANA vs ADIANA for three different compressors (random sparsification, random dithering and natural compression) on the `a9a` dataset with regularization constant $\lambda = 10^{-3}$.

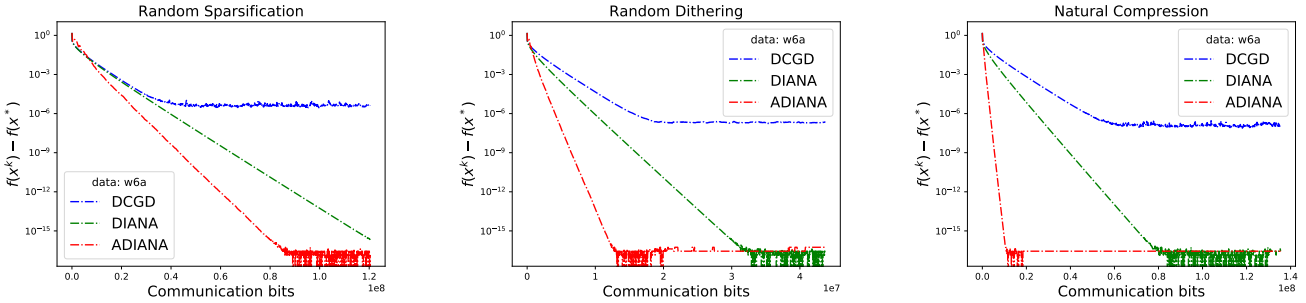

Figure 6: The communication complexity performance of DCGD vs DIANA vs ADIANA for three different compressors (random sparsification, random dithering and natural compression) on the `w6a` dataset with regularization constant $\lambda = 10^{-3}$.

B.2 Communication efficiency

In this subsection, we compare ADIANA, DIANA, and their uncompressed versions to show the communication efficiency of our method in Figure 7 and Figure 8. The number of nodes is 20 as well. From the numerical results, we can see for random sparsification, ADIANA is slower than the uncompressed version. However, for random dithering and natural compression, ADIANA is about two times faster than the uncompressed version.

B.3 Different number of nodes

In this subsection, we show the performance of ADIANA for different number of nodes with random dithering and natural compression for a9a and w6a datasets. In the previous numerical results, the number of communication

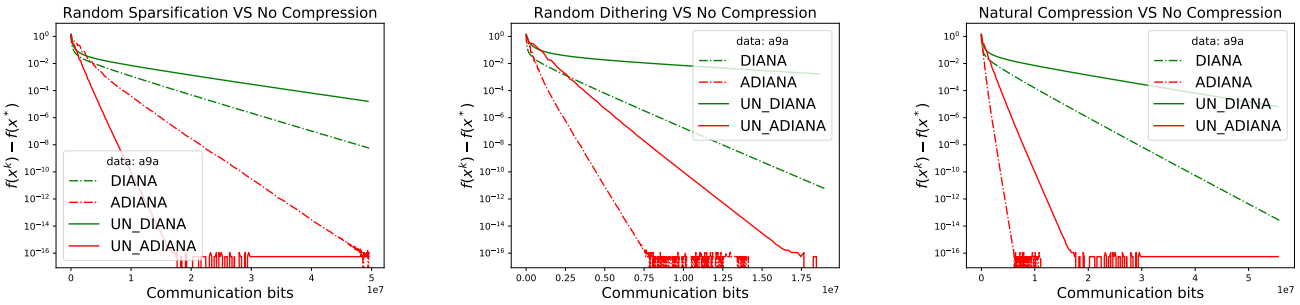

Figure 7: The communication complexity performance of DIANA and ADIANA with and without compression on `a9a` dataset and with regularization parameter $\lambda = 10^{-3}$.

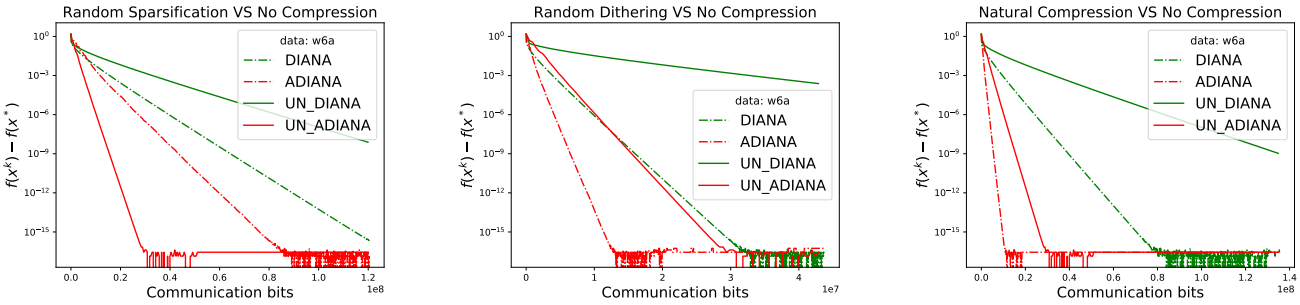

Figure 8: The communication complexity performance of DIANA and ADIANA with and without compression on `w6a` dataset and with regularization parameter $\lambda = 10^{-3}$.

bits is not multiplied by the number of nodes, since the number of nodes is the same for all methods. However, the number of nodes is different in this subsection. Hence, the number of communication bits is multiplied by the number of nodes. Figure 9 and Figure 10 show that it is slower for more nodes with respect to the total communication bits.

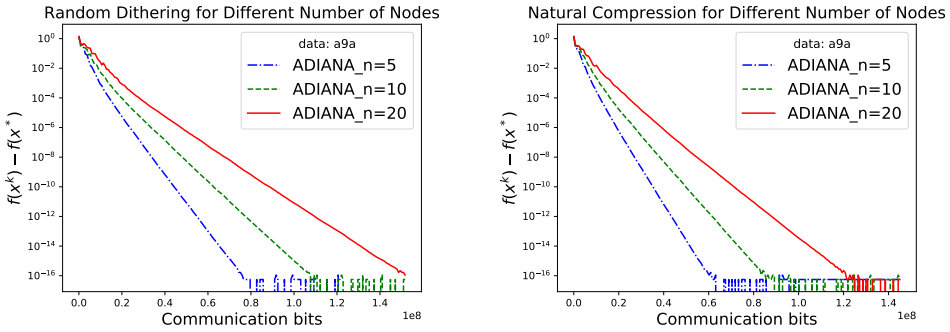

Figure 9: The communication complexity performance of ADIANA with different numbder of nodes on `a9a` dataset and with regularization parameter $\lambda = 10^{-3}$.

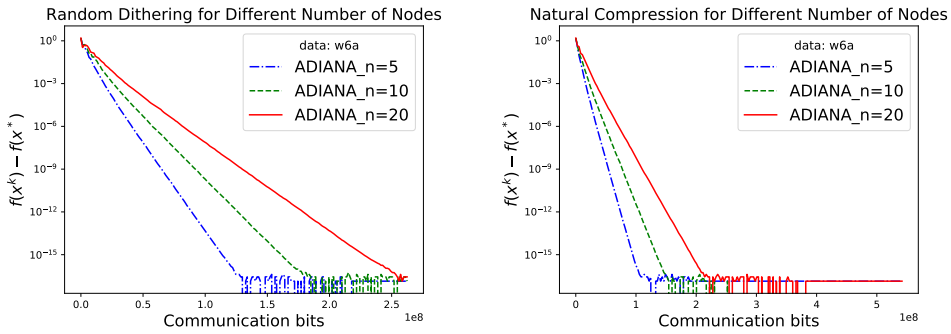

Figure 10: The communication complexity performance of ADIANA with different numbder of nodes on `w6a` dataset and with regularization parameter $\lambda = 10^{-3}$.