A Random sparsification

We illustrate the random-s sparsification here. More examples of unbiased compressors can be found in literature [51].

Example 1 (RANDOM-s SPARSIFICATION)**.** *For any $x \in \mathbb{R}^d$, the random-s sparsification is defined by $C(x) := \frac{d}{s}(\xi \odot x)$ where $\odot$ denotes the entry-wise product and $\xi \in \{0,1\}^d$ is a uniformly random binary vector with s non-zero entries. This random-s sparsification operator C satisfies Assumption 2 with $\omega = d/s - 1$. When each entry of the input x is represented with r bits, random-s sparsification compressor takes rs bits to transmit s entries and $\log_2 \binom{d}{s}$ bits to transmit the indices of s transmitted entries, resulting in a total $\frac{rd}{1+\omega} + \log_2 \binom{d}{s}$ bits in each communication round, see [51, Table 1].*

B Proof of Proposition 1

We first recall a result proved by [51].

Lemma 2 ([51], Theorem 2)**.** *Let $C : \mathbb{R}^d \to \mathbb{R}^d$ be any unbiased compressors satisfying 2 and b be the total number of bits needed to encode the compressed vector $C(x)$ for any $x \in \mathbb{R}^d$. If each entry of the input x is represented with r bits, it holds that $\max\{\frac{\omega}{1+\omega}, 4^{-r}\}4^{b/d} \geq 1$.*

Using Lemma 2, when $\omega/(1+\omega) \leq 4^{-r}$, *i*.e., $\omega \leq (4^r - 1)^{-1} \leq 1/3$, we have $(1+\omega) = \Theta(1)$ and $b \geq rd = \Omega_r(d/(1+\omega))$, where r is regarded as a constant in $\Omega_r(\cdot)$. When $\omega/(1+\omega) \geq 4^{-r}$, we have

$$b \geq d \log_4(1+\omega^{-1}) = d \ln(1+\omega^{-1})/\ln(4) \geq d \frac{\omega^{-1}}{\ln(4)(1+\omega^{-1})} = \Omega_r\left(\frac{d}{1+\omega}\right),$$

where we use the inequality $\ln(1+t) \geq t/(1+t)$ with $t = \omega^{-1} \geq 0$.

C Proof of Theorem 2

Following [5, 9] , we denote the k-th coordinate of a vector $x \in \mathbb{R}^d$ by $[x]_k$ for $k = 1, \ldots, d$, and let $\mathrm{prog}(x)$ be

$$\mathrm{prog}(x) := \begin{cases} 0, & \text{if } x = 0, \\ \max_{1 \leq k \leq d}\{k : [x]_k \neq 0\}, & \text{otherwise.} \end{cases}$$

Similarly, for a set of multiple points $\mathcal{X} = \{x_1, x_2, \ldots\}$, we define $\mathrm{prog}(\mathcal{X}) := \max_{x \in \mathcal{X}} \mathrm{prog}(x)$. We call a function f zero-chain if it satisfies

$$\mathrm{prog}(\nabla f(x)) \leq \mathrm{prog}(x) + 1, \quad \forall\, x \in \mathbb{R}^d,$$

which implies that starting from $x^0 = 0$, a single gradient evaluation can only earn at most one more non-zero coordinate for the model parameters.

Let us now illustrate the setup of distributed optimization with communication compression. For any $t \geq 1$, we consider the t-th communication round, which begins with the server broadcasting a vector denoted as u^t to all workers. We initialize u^1 as x^0. Upon receiving the vector u^t from the server, each worker performs necessary algorithmic operations, and the round concludes with each worker sending a compressed message back to the server.

We denote v_i^t as the vector that worker i aims to send in the t-th communication round before compression, and $\hat{v}_i^t$ as the compressed vector that will be received by the server, *i*.e., $\hat{v}_i^t = C_i(v_i^t)$. While we require communication to be synchronous among workers, we do not impose restrictions on the number of gradient queries made by each worker within a communication round. We use $\mathcal{Y}_i^t$ to represent the set of vectors at which worker i makes gradient queries in the t-th communication round, after receiving u^t but before sending $\hat{v}_i^t$.

Following the above description, we now formally state the linear spanning property in the setting of centralized distributed optimization with communication compression.

Definition 2 (LINEAR-SPANNING ALGORITHMS)**.** *We say a distributed algorithm A is linear-spanning if, for any $t \geq 1$, the following conditions hold:*

1. *The server can only send a vector in the linear manifold spanned by all the past received messages, sent messages,* i.*e.,* $u^t \in \operatorname{span}\left(\{u^r\}_{r=1}^{t-1} \cup \{\hat{v}_i^r : 1 \le i \le n\}_{r=1}^{t-1}\right)$.
2. *Worker* i *can only query at vectors in the linear manifold spanned by its past received messages, compressed messages, and gradient queries,* i.*e.,* $\mathcal{Y}_i^t \subseteq \operatorname{span}\left(\{u^r\}_{r=1}^{t} \cup \{\nabla f_i(y) : y \in \mathcal{Y}_i^r\}_{r=1}^{t-1} \cup \{\hat{v}_i^r\}_{r=1}^{t-1}\right)$.
3. *Worker* i *can only send a vector in the linear manifold spanned by its past received messages, compressed messages, and local gradient queries,* i.*e.,* $v_i^t \in \operatorname{span}\left(\{u^r\}_{r=1}^{t} \cup \{\nabla f_i(y) : y \in \mathcal{Y}_i^r\}_{r=1}^{t} \cup \{\hat{v}_i^r\}_{r=1}^{t-1}\right)$.
4. *After* t *communication rounds, the server can only output a model in the linear manifold spanned by all the past received messages, sent messages,* i.*e.,* $\hat{x}^t \in \operatorname{span}\left(\{u^r\}_{r=1}^{t} \cup \{\hat{v}_i^r : 1 \le i \le n\}_{r=1}^{t}\right)$.

In essence, when starting from $x^0 = 0$, the above linear-spanning property requires that any expansion of non-zero coordinates in vectors held by worker i (*e*.g., $\mathcal{Y}_i^t$, v_i^t) are attributed to its past local gradient updates, local compression, or synchronization with the server. Meanwhile, it also requires that any expansion of non-zero coordinate in vectors held, including the final algorithmic output, in the server is due to the received compressed messages from workers.

Without loss of generality, we assume algorithms to start from $x^0 = 0$ throughout the proofs. When $\{f_i\}_{i=1}^n$ are further assumed to be zero-chain, following Definition 2, one can easily establish by induction that for any $t \ge 1$,

$$\begin{aligned}
&\max_{1\le r\le t} \operatorname{prog}(u^r) \le \max_{1\le r<t} \max_{1\le i\le n} \operatorname{prog}(\hat{v}_i^r) \qquad (10)\\
&\max_{1\le r\le t} \operatorname{prog}(v_i^t) \le \max_{1\le r<t} \max\left\{ \max_{1\le i\le n} \operatorname{prog}(\hat{v}_i^r), \operatorname{prog}(\mathcal{Y}_i^r) \right\} \le \max_{1\le r<t} \max_{1\le i\le n} \operatorname{prog}(\hat{v}_i^r) + 1\\
&\operatorname{prog}(\hat{x}^t) \le \max_{1\le r\le t} \max_{1\le i\le n} \operatorname{prog}(\hat{v}_i^r)
\end{aligned}$$

Next, we outline the proofs for the lower bounds presented in Theorem 2. For each case, we provide separate proofs for terms in the lower bound by constructing different hard-to-optimize examples, respectively. The construction of these proofs follows four steps:

- Constructing a set of zero-chain local functions $\{f_i\}_{i=1}^n$.
- Constructing a set of independent unbiased compressors $\{C_i\}_{i=1}^n \subseteq \mathcal{U}_\omega^{\mathrm{ind}}$. These compressors are delicately designed to impede algorithms from expanding the non-zero coordinates of model parameters.
- Establishing a limitation on zero-respecting algorithms that utilize the predefined compressor with t rounds of compressed communication on each worker. This limitation is based on the non-zero coordinates of model parameters.
- Translating the above limitation into the lower bound of the complexity measure defined in equation (3).

While the overall proof structure is similar to that of [19], our novel construction of functions and compressors enable us to derive lower bounds for independent compressors. These lower bounds clarify the unique properties and benefits of independent compressors.

We will use the following lemma in the analysis of the third step.

Lemma 3 ([19], Lemma 3)**.** *Given a constant* $p \in [0, 1]$ *and random variables* $\{B^t\}_{t=0}^\infty$ *such that* $B^t \le B^{(t-1)} + 1$ *and* $\mathbb{P}(B^t \le B^{t-1} \mid \{B^r\}_{r=0}^{t-1}) \ge 1 - p$ *for any* $t \ge 1$*, it holds for* $t \ge 1/p$*, with probability at least* $1 - e^{-1}$*, that* $B^t \le B^0 + ept$.

C.1 Strongly-convex case

Below, we present two examples, each of which corresponding to a lower bound LB_m for T_ϵ. We integrate the two lower bounds together and use the inequality

$$T_\epsilon \ge \max_{1\le m\le 2} \{LB_m\} = \Omega\left(LB_1 + LB_2\right)$$

to accomplish the lower bound for strongly-convex problems in Theorem 2.

Example 1. In this example, we prove the lower bound $\Omega((1+\omega)(1+\sqrt{\kappa/n})\ln(\mu\Delta/\epsilon))$.

(Step 1.) We assume the variable $x \in \ell_2 \triangleq \{([x]_1, [x]_2, \dots,) : \sum_{r=1}^{\infty}[x]_r^2 < \infty\}$ to be infinitely dimensional and square-summable for simplicity. It is easy to adapt the argument for finitely dimensional variables as long as the dimension is proportionally larger than t. Let M be

$$M = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{\infty\times\infty},$$

then it is easy to see $0 \preceq M \preceq 4I$. Let $\{f_i\}_{i=1}^n$ be as follows

$$f_i(x) = \begin{cases} \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{4}\sum_{r\geq 0}([x]_{nr+i} - [x]_{nr+i+1})^2, & \text{if } 1 \leq i \leq n-1, \\ \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{4}\left([x]_1^2 + \sum_{r\geq 1}([x]_{nr} - [x]_{nr+1})^2 - 2\lambda[x]_1\right), & \text{if } i = n. \end{cases}$$

where $\lambda \in \mathbb{R}\backslash\{0\}$ is to be specified. It is easy to see that $\sum_{r\geq 0}([x]_{nr+i} - [x]_{nr+i+1})^2$ and $[x]_1^2 + \sum_{r\geq 0}([x]_{nr} - [x]_{nr+1})^2 - 2\lambda[x]_1$ are convex and 4-smooth. Consequently, all f_is are L-smooth and μ-strongly convex. More importantly, it is easy to verify that all f_is defined above are zero-chain functions and satisfy

$$\text{prog}(\nabla f_i(x)) \begin{cases} = \text{prog}(x) + 1, & \text{if } \text{prog}(x) \equiv i \mod n, \\ \leq \text{prog}(x), & \text{otherwise.} \end{cases} \tag{11}$$

We further have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{4n}\left(x^\top M x - 2\lambda[x]_1\right)$. For the functions defined above, we also establish that

Lemma 4. *Let $\kappa \triangleq L/\mu \geq 1$, it holds for any x that,*

$$f(x) - \min_x f(x) \geq \frac{\mu}{2}\left(1 - 2\left(1 + \sqrt{1 + \frac{2(\kappa-1)}{n}}\right)^{-1}\right)^{2\text{prog}(x)} \|x^0 - x^\star\|^2.$$

Proof. The minimum $x^\star$ of function f satisfies $\left(\frac{L-\mu}{2n}M + \mu\right)x^\star - \lambda\frac{L-\mu}{2}e_1 = 0$, which is equivalent to

$$\begin{aligned} \frac{2\kappa + 2n - 2}{\kappa - 1}[x^\star]_1 - [x^\star]_2 &= \lambda, \\ -[x^\star]_{j-1} + \frac{2\kappa + 2n - 2}{\kappa - 1}[x^\star]_j - [x^\star]_{j+1} &= 0, \quad \forall\, j \geq 2. \end{aligned} \tag{12}$$

Note that

$$q = \frac{\kappa + n - 1 - \sqrt{n(2\kappa + n - 2)}}{\kappa - 1} = 1 - \frac{2}{1 + \sqrt{1 + \frac{2(\kappa-1)}{n}}}$$

is the only root of the equation $q^2 - \frac{2\kappa+2n-2}{\kappa-1}q + 1 = 0$ that is smaller than 1. Then it is straight forward to check $x^\star = \left([x^\star]_j = \lambda q^j\right)_{j\geq 1}$ satisfies (12). By the strong convexity of f, $x^\star$ is the unique solution. Therefore, we have that

$$\|x - x^\star\|^2 \geq \sum_{j=\text{prog}(x)+1}^{\infty} \lambda^2 q^{2j} = \lambda^2 \frac{q^{2(r+1)}}{1-q^2} = q^{2r}\|x^0 - x^\star\|^2.$$

Finally, using the strong convexity of f leads to the conclusion. □

Following the proof of Lemma 4, we have

$$\|x^0 - x^\star\|^2 = \lambda^2 \sum_{j=1}^{\infty} q^{2j} = \lambda^2 \frac{q^2}{1-q^2}$$

Therefore, for any given $\Delta > 0$, letting $\lambda = \sqrt{((1-q^2)\Delta)/q^2}$ results in $\|x^0 - x^\star\|^2 = \Delta$. Consequently, our construction ensures $\{f_i\}_{i=1}^n \in \mathcal{F}_{L,\mu}^{\Delta}$.

(Step 2.) For the construction of ω-unbiased compressors, we consider $\{C_i\}_{i=1}^n$ to be independent random sparsification compressors. Building upon Example 1, we make a slight modification: during a round of communication on any worker, each coordinate is independetly chosen with a probability of $(1+\omega)^{-1}$ to be transmitted, and if selected, its value is scaled by $(1+\omega)$ and then the scaled value is transmitted. Notably, the indices of chosen coordinates are not identical across all workers due to the independence of compressors. It can be easily verified that this construction ensures that $\{C_i\}_{i=1}^n \subseteq \mathcal{U}_\omega^{\mathrm{ind}}$.

(Step 3.) Since the algorithmic output $\hat{x}^t$ calculated by the server lies in the linear manifold spanned by received messages, we can use (10) to obtain the following expression:

$$\mathrm{prog}(\hat{x}^t) \leq \max_{1\leq r\leq t}\max_{1\leq i\leq n}\max\{\mathrm{prog}(u^r), \mathrm{prog}(\hat{v}_i^r)\} = \max_{1\leq r\leq t}\max_{1\leq i\leq n}\mathrm{prog}(\hat{v}_i^r) \triangleq B^t. \tag{13}$$

We next bound B^t with $B^0 := 0$ by showing that $\{B^t\}_{t=0}^\infty$ satisfies Lemma 3 with $p = (1+\omega)^{-1}$.

For any linear-spanning algorithm A, according to (11), the worker i can only attain one additional non-zero coordinate through local gradient-based updates when $\mathrm{prog}(\mathcal{Y}_i^t) \equiv i \mod n$. In other words, upon receiving messages $\{u_i^r\}_{r=1}^t$ from the server, we have

$$\mathrm{prog}(v_i^t) \leq \begin{cases} \max_{1\leq r\leq t}\mathrm{prog}(u_i^r) + 1 \leq B^{t-1}+1, & \text{if } \mathrm{prog}(\mathcal{Y}_i^t) \equiv i \mod n, \\ \max_{1\leq r\leq t}\mathrm{prog}(u_i^r) \leq B^{t-1}, & \text{otherwise.} \end{cases}$$

Consequently, we have

$$\max_{1\leq r\leq t}\mathrm{prog}(v_i^r) \leq \max_{1\leq r\leq t} B^{r-1}+1 = B^{t-1}+1.$$

It then follows from the definition of the constructed C_i in Step 2 that $\max_{1\leq i\leq n}\mathrm{prog}(\hat{v}_i^t) \leq \max_{1\leq i\leq n}\mathrm{prog}(v_i^t)$, and therefore we have:

$$B^t \leq \max_{1\leq r\leq t}\max_{1\leq i\leq n}\mathrm{prog}(v_i^r) \leq B^{t-1}+1.$$

Next, we aim to prove that $B^t \leq B^{t-1}+1$ with a probability of at least $\omega/(1+\omega)$. For any $t \geq 1$, let $i \in \{1,\ldots,n\}$ be such that $B^{t-1} \equiv i \mod n$. Due to the property in equation (11), during the t-th communication round, if $\mathrm{prog}(\mathcal{Y}_i^t) = B^{t-1}$, worker i can push the number of non-zero entries forward by 1, resulting in $\mathrm{prog}(v_i^t) = B^{t-1}+1$, using local gradient updates. Note that any other worker j cannot achieve this even if $\mathrm{prog}(\mathcal{Y}_j^t) = B^{t-1}$ due to equation (11).

Therefore, to achieve $B^t = B^{t-1}+1$, it is necessary for worker i to transmit a non-zero value at the $(B^{t-1}+1)$-th entry to the server. Otherwise, we have $B^t \leq B^{t-1}$. However, since the compressor C_i associated with worker i has a probability $\omega/(1+\omega)$ to zero out the $(B^{t-1}+1)$-th entry in the t-th communication round, we have

$$\mathbb{P}\left(B^t \leq B^{t-1} \mid \{B^r\}_{r=0}^{t-1}\right) \geq \omega/(1+\omega).$$

In summary, we have shown that $B^t \leq B^{t-1}+1$ and $\mathbb{P}(B^t \leq B^{t-1} \mid \{B^r\}_{r=0}^{t-1}) \geq \omega/(1+\omega)$.

By applying Lemma 3, we can conclude that for any $t \geq (1+\omega)^{-1}$, with a probability of at least $1-e^{-1}$, it holds that $B^t \leq et/(1+\omega)$ and hence $\mathrm{prog}(\hat{x}^t) \leq et/(1+\omega)$ due to (13).

(Step 4.) Using Lemma 4 and that $\mathrm{prog}(\hat{x}^t) \leq et/(1+\omega)$ with probability at least $1-e^{-1}$, we obtain

$$\begin{aligned}\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \geq & \frac{(1-e^{-1})\mu\Delta}{2}\left(1-2\left(1+\sqrt{1+\frac{2(\kappa-1)}{n}}\right)^{-1}\right)^{2et/(1+\omega)} \\ = & \Omega\left(\mu\Delta\exp\left(-\frac{4et}{(\sqrt{\kappa/n}+1)(1+\omega)}\right)\right).\end{aligned} \tag{14}$$

Therefore, to ensure $\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \leq \epsilon$, relation (14) implies the lower bound $T_\epsilon = \Omega((1+\omega)(1+\sqrt{\kappa/n})\ln(\mu\Delta/\epsilon))$.

Example 2. Considering $f_1 = f$ to be homogeneous and $C_i = I$ to be a loss-less compressor for all $1 \leq i \leq n$, the problem reduces to single-node convex optimization. In this case, the lower bound of $\Omega(\sqrt{\kappa}\ln(\mu\Delta/\epsilon))$ is well-known in the literature, as shown in [45, 44].

With the two lower bounds achieved in Examples 1 and 2, we have

$$\begin{aligned} T_\epsilon &= \Omega\Big((1+\omega)(1+\sqrt{\kappa/n})\ln(\mu\Delta/\epsilon) + \sqrt{\kappa}\ln(\mu\Delta/\epsilon)\Big) \\ &= \Omega\Big((1+\omega+\sqrt{\kappa/n}+\omega\sqrt{\kappa/n}+\sqrt{\kappa})\ln(\mu\Delta/\epsilon)\Big) \\ &= \Omega\Big((\omega+\omega\sqrt{\kappa/n}+\sqrt{\kappa})\ln(\mu\Delta/\epsilon)\Big) \end{aligned}$$

which is the result for the strongly-convex case in Theorem 2.

C.2 Generally-convex case

Below, we present three examples, each of which corresponding to a lower bound LB_m for T_ϵ. We integrate the three lower bounds together and use the inequality

$$T_\epsilon \geq \max_{1\leq m\leq 3}\{LB_m\} = \Omega\left(LB_1 + LB_2 + LB_3\right)$$

to accomplish the lower bound for the generally-convex case in Theorem 2.

Example 1. In this example, we prove the lower bound $\Omega((1+\omega)(L\Delta/\epsilon)^{1/2})$.

(Step 1.) We assume variable $x \in \mathbb{R}^d$, where d can be sufficiently large and will be determined later. Let M denote

$$M = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{d\times d},$$

it is easy to verify $0 \preceq M \preceq 4I$. Similar to example 1 of the strongly-convex case, we consider

$$f_i(x) = \begin{cases} \frac{L}{4}\sum_{r\geq 0}([x]_{nr+i} - [x]_{nr+i+1})^2, & \text{if } 1 \leq i \leq n-1, \\ \frac{L}{4}\left([x]_1^2 + \sum_{r\geq 1}([x]_{nr} - [x]_{nr+1})^2 - 2\lambda[x]_1\right), & \text{if } i = n. \end{cases}$$

where $\lambda \in \mathbb{R}\backslash\{0\}$ is to be specified. It is easy to see that all f_is are L-smooth. We further have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{L}{4n}\left(x^\top Mx - 2\lambda[x]_1\right)$. The f_i functions defined above are also zero-chain functions satisfying (11).

Following [44], it is easy to verify that the optimum of f satisfies

$$x^\star = \left(\lambda\left(1 - \frac{k}{d+1}\right)\right)_{1\leq k\leq d} \quad \text{and} \quad f(x^\star) = \min_x f(x) = -\frac{\lambda^2 Ld}{4n(d+1)}.$$

More generally, it holds for any $0 \leq k \leq d$ that

$$\min_{x:\,\mathrm{prog}(x)\leq k} f(x) = -\frac{\lambda^2 Lk}{4n(k+1)}. \tag{15}$$

Since $\|x^0 - x^\star\|^2 = \frac{\lambda^2}{(d+1)^2}\sum_{k=1}^d k^2 = \frac{\lambda^2 d(2d+1)}{6(d+1)} \leq \frac{\lambda^2 d}{3}$, letting $\lambda = \sqrt{3\Delta/d}$, we have $\{f_i\}_{i=1}^n \in \mathcal{F}_{L,0}^\Delta$.

(Step 2.) Same as Step 2 of Example 1 of the strongly-convex case, we consider $\{C_i\}_{i=1}^n$ to be independent random sparsification operators.

(Step 3.) Following the same argument as step 3 of example 1 of the strongly-convex case, we have that for any $t \geq (1+\omega)^{-1}$, it holds with probability at least $1-e^{-1}$ that $\mathrm{prog}(\hat{x}^t) \leq et/(1+\omega)$.

(Step 4.) Thus, combining (15), we have

$$\begin{aligned}\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \geq & (1-e^{-1})\frac{\lambda^2 L}{4n}\left(\frac{d}{d+1} - \frac{et/(1+\omega)}{1+et/(1+\omega)}\right) \\ = & (1-e^{-1})\frac{3L\Delta}{4nd}\left(\frac{d}{d+1} - \frac{et/(1+\omega)}{1+et/(1+\omega)}\right)\end{aligned}$$

Letting $d = 1 + et/(1+\omega)$, we further have

$$\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \geq \frac{3(1-e^{-1})L\Delta}{8net(1+\omega)^{-1}(1+2et(1+\omega)^{-1})} = \Omega\left(\frac{(1+\omega)^2 L\Delta}{nt^2}\right).$$

Therefore, to ensure $\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \leq \epsilon$, the above inequality implies the lower bound to be $T = \Omega((1+\omega)(L\Delta/(n\epsilon))^{\frac{1}{2}})$.

Example 2. Considering $f_1 = f$ to be homogeneous and $C_i = I$ to be a loss-less compressor for all $1 \leq i \leq n$. The problem reduces to the single-node convex optimization. The lower bound $\Omega(\sqrt{L\Delta/\epsilon})$ is well-known in literature, see, *e*.g., [45, 44].

Example 3. In this example, we prove the lower bound $\Omega(\omega \ln(L\Delta/\epsilon))$.

(Step 1.) We consider $f_1 = \cdots = f_{n-1} = L\|x\|^2/2$ and $f_n = L\|x\|^2/2 + n\lambda\langle \mathbb{1}_d, x\rangle$ where $\mathbb{1}_d \in \mathbb{R}^d$ is the vector with all enries being 1 and $\lambda \in \mathbb{R}$ is to be determined. By definition, $\{f_i\}_{i=1}^n$ are μ-strongly-convex and L-smooth and the solution $x^\star = -\frac{\lambda}{L}\mathbb{1}_d$. Letting $\lambda = L\sqrt{\Delta}/\sqrt{n}$, we have $\|x^\star - x^0\|^2 = \Delta$. Thus, the construction ensures $\{f_i\}_{i=1}^n \in \mathcal{F}_{L,\mu}^{\Delta}$.

(Step 2.) Same as in Example 1, we consider $\{C_i\}_{i=1}^n$ to be independent random sparsification operators.

(Step 3.) By the construction of $\{f_i\}_{i=1}^n$, we observe that the optimization process relies solely on transmitting the information of $\mathbb{1}_d$ from worker n to the server. Let E^t denote the set of entries at which the server has received a non-zero value from worker n in the first t communication rounds. Note that for each entry, due to the construction of $\{C_i\}_{i=1}^n$, the server has a probability of at least $(\omega/(1+\omega))^t$ of not receiving a non-zero value at that entry from worker n. Consequently, $|(E^t)^c|$ is lower bounded by the sum of n independent $\mathrm{Bernoulli}(\omega^t/(1+\omega)^t)$ random variables. Therefore, we have $\mathbb{E}[|(E^t)^c|] \geq d\omega^t/(1+\omega)^t$.

(Step 4.) Given $|E^t|$, due to the linear-spanning property, we have $\hat{x}^t \in \mathrm{span}\{e_j : j \in E^t\}$ where e_j is the j-th canonical vector. As a result, we have

$$\begin{aligned}&\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \\ \geq & \mathbb{E}[\min_{x \in \mathrm{span}\{e_j : j \in E^t\}} f(x)] - \min_x f(x) = \frac{L\Delta}{2}\frac{\mathbb{E}[|(E^t)^c|]}{d} \geq \frac{L\Delta}{2}\frac{\omega^t}{(1+\omega)^t}.\end{aligned} \tag{16}$$

Therefore, to ensure $\mathbb{E}[f(\hat{x}^t)] - \min_x f(x) \leq \epsilon$, (16) implies the lower bound $T_\epsilon = \Omega(\omega \ln(L\Delta/\epsilon))$.

With the three lower bounds achieved in Examples 1, 2, and 3, we have

$$\begin{aligned}T_\epsilon &= \Omega\Big(\sqrt{\frac{L\Delta}{\epsilon}} + (1+\omega)\sqrt{\frac{L\Delta}{n\epsilon}} + \omega\ln(L\Delta/\epsilon)\Big) \\ &= \Omega\Big(\sqrt{\frac{L\Delta}{\epsilon}} + \omega\sqrt{\frac{L\Delta}{n\epsilon}} + \omega\ln(L\Delta/\epsilon)\Big)\end{aligned}$$

which is the result for the generally-convex case in Theorem 2.

D Proof of Theorem 3

D.1 Strongly-convex case

We first present several important lemmas, followed by the definition of a Lyapunov function with delicately chosen coefficients for each term. Finally, we prove Theorem 3 by utilizing these lemmas. Throughout the convergence analysis, we use the following notations:

$$\begin{aligned}
\mathcal{W}^k =& f(w^k) - f^\star, \quad \mathcal{Y}^k = f(y^k) - f^\star, \quad \mathcal{Z}^k = \|z^k - x^\star\|^2, \\
\mathcal{H}^k =& \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(w^k)\|^2, \quad \mathcal{G}^k = \|g^k - \nabla f(x^k)\|^2, \\
\mathcal{G}_w^k =& \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(w^k) - \nabla f_i(x^k)\|^2, \quad \mathcal{G}_y^k = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(y^k) - \nabla f_i(x^k)\|^2.
\end{aligned}$$

We use $\mathbb{E}_k$ or $\mathbb{E}$ indicate the expectation with respect to the randomness in the k-th iteration or all historical randomness, respectively.

Lemma 5. *If $0 \le \beta \le 1$, it holds for $\forall\, k \ge 0$ that,*

$$\begin{aligned}
\mathcal{Z}^{k+1} \le& 2\gamma_k\langle g^k, x^\star - x^k\rangle + \frac{2\gamma_k\beta\theta_2}{\theta_{1,k}}\langle g^k, w^k - x^k\rangle + \frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}}\langle g^k, y^k - x^k\rangle \\
&+ \beta\mathcal{Z}^k + (1-\beta)\|x^k - x^\star\|^2 + \gamma_k^2\|g^k\|^2. \qquad (17)
\end{aligned}$$

Proof. Following the update rules in Algorithm 1, we have

$$\begin{aligned}
\mathcal{Z}^{k+1} =& \left\|\beta z^k + (1-\beta)x^k - x^\star + \frac{\gamma_k}{\eta_k}(y^{k+1} - x^k)\right\|^2 \\
=& \|\beta(z^k - x^\star) + (1-\beta)(x^k - x^\star)\|^2 + \gamma_k^2\|g^k\|^2 \\
&+ \langle 2\gamma_k g^k, \beta z^k + (1-\beta)x^k - x^\star\rangle. \qquad (18)
\end{aligned}$$

Since $x^k = \theta_{1,k}z^k + \theta_2 w^k + (1-\theta_{1,k}-\theta_2)y^k$, we have

$$\beta z^k + (1-\beta)x^k - x^\star = (x^k - x^\star) + \frac{\beta\theta_2}{\theta_{1,k}}(x^k - w^k) + \frac{\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}}(x^k - y^k). \qquad (19)$$

Plugging (19) into (18), using

$$\|\beta(z^k - x^\star) + (1-\beta)(x^k - x^\star)\|^2 \le \beta\|z^k - x^\star\|^2 + (1-\beta)\|x^k - x^\star\|^2,$$

we obtain (17). □

Lemma 6. *Under Assumption 1, if parameters satisfy $\theta_{1,k}, \theta_2, 1-\theta_{1,k}-\theta_2 \in (0,1)$, $\eta_k \in \left(0, \frac{1}{2L}\right]$, $\gamma_k = \frac{\eta_k}{2\theta_{1,k}+\eta_k\mu}$ and $\beta = 1 - \gamma\mu = \frac{2\theta_{1,k}}{2\theta_{1,k}+\eta_k\mu}$, then we have for any iteration $k \ge 0$ that*

$$\begin{aligned}
\frac{2\gamma_k\beta}{\theta_{1,k}}\mathbb{E}_k[\mathcal{Y}^{k+1}] + \mathbb{E}_k[\mathcal{Z}^{k+1}] \le& \frac{2\gamma_k\beta\theta_2}{\theta_{1,k}}\mathcal{W}^k + \frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}}\mathcal{Y}^k + \beta\mathcal{Z}^k + \frac{5\gamma_k\beta\eta_k}{4\theta_{1,k}}\mathcal{G}^k \\
&- \frac{\gamma_k\beta\theta_2}{L\theta_{1,k}}\mathcal{G}_w^k - \frac{\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{L\theta_{1,k}}\mathcal{G}_y^k. \qquad (20)
\end{aligned}$$

Proof. By Assumption 1 and update rules in Algorithm 1, we have

$$\begin{aligned}
f(y^{k+1}) \le& f(x^k) + \langle\nabla f(x^k), y^{k+1} - x^k\rangle + \frac{L}{2}\|y^{k+1} - x^k\|^2 \\
=& f(x^k) - \langle\nabla f(x^k), \eta_k g^k\rangle + \frac{L}{2}\eta_k^2\|g^k\|^2 \\
=& f(x^k) - \eta_k\langle\nabla f(x^k) - g^k, g^k\rangle + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\|g^k\|^2. \qquad (21)
\end{aligned}$$

By L-smoothness and μ-strongly convexity, we have for $\forall u \in \mathbb{R}^d$ that

$$f(u) \geq f(x^k) + \langle \nabla f(x^k), u - x^k \rangle + \frac{\mu}{2}\|u - x^k\|^2,$$

and that

$$f_i(u) \geq f_i(x^k) + \langle \nabla f_i(x^k), u - x^k \rangle + \frac{1}{2L}\|\nabla f_i(u) - \nabla f_i(x^k)\|^2,$$

thus we obtain for $\forall u \in \mathbb{R}^d$,

$$\begin{aligned} f(x^k) \leq & f(u) - \langle \nabla f(x^k), u - x^k \rangle \\ & - \max\left\{ \frac{\mu}{2}\|u - x^k\|^2, \frac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(u) - \nabla f_i(x^k)\|^2 \right\}. \end{aligned} \tag{22}$$

Applying Young's inequality to (21) and using $\eta_k \leq 1/(2L)$, we reach

$$\begin{aligned} f(y^{k+1}) \leq & f(x^k) + \frac{\eta_k}{2}\mathcal{G}^k - \frac{\eta_k}{2}(1 - L\eta_k)\|g^k\|^2 \\ \leq & f(x^k) + \frac{\eta_k}{2}\mathcal{G}^k - \frac{\eta_k}{4}\|g^k\|^2. \end{aligned} \tag{23}$$

Adding (17) in Lemma 5 to $\left(\frac{2\gamma_k\beta}{\theta_{1,k}} + 2\gamma_k(1-\beta)\right)\times$(23) + $2\gamma_k\times$(22) (where $u = x^\star$) + $\frac{2\gamma_k\beta\theta_2}{\theta_{1,k}}\times$(22) (where $u = w^k$) + $\frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}}\times$(22) (where $u = y^k$) and using the unbiasedness of g^k, we obtain

$$\begin{aligned} & \frac{2\gamma_k\beta}{\theta_{1,k}}\mathbb{E}_k[\mathcal{Y}^{k+1}] + \mathbb{E}_k[\mathcal{Z}^{k+1}] \\ \leq & \beta\mathcal{Z}^k + (1 - \beta - \mu\gamma_k)\|x^k - x^\star\|^2 + \left(\gamma_k^2 - \frac{\eta_k\gamma_k\beta}{2\theta_{1,k}}\right)\mathbb{E}_k[\|g^k\|^2] + \eta_k\left(\frac{\gamma_k\beta}{\theta_{1,k}} + \gamma_k(1-\beta)\right)\mathcal{G}^k \\ & - \frac{\gamma_k\beta\theta_2}{L\theta_{1,k}}\mathcal{G}_w^k - \frac{\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{L\theta_{1,k}}\mathcal{G}_y^k + \frac{2\gamma_k\beta\theta_2}{\theta_{1,k}}\mathcal{W}^k + \frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}}\mathcal{Y}^k \\ & - 2\gamma_k(1-\beta)\mathbb{E}_k[\mathcal{Y}^{k+1}] - \frac{\eta_k\gamma_k(1-\beta)}{2}\mathbb{E}_k[\|g^k\|^2] \end{aligned}$$

On top of that, by applying our choice of the parameters, it can be easily verified that $1-\beta-\mu\gamma_k = 0$, $\gamma_k^2 - \frac{\eta_k\gamma_k\beta}{2\theta_{1,k}} = 0$, $1 - \beta \leq \frac{\beta}{4\theta_{1,k}}$, which leads to (20). □

Lemma 7 ([33], Lemma 3, 4, 5)**.** *Under Assumptions 1, 2, and 3, the iterates of Algorithm 1 satisfy the following inequalities:*

$$\mathbb{E}[\mathcal{W}^{k+1}] = (1-p)\mathbb{E}[\mathcal{W}^k] + p\mathbb{E}[\mathcal{Y}^k], \tag{24}$$

$$\mathbb{E}[\mathcal{G}^k] \leq \frac{2\omega}{n}\mathbb{E}[\mathcal{G}_w^k] + \frac{2\omega}{n}\mathbb{E}[\mathcal{H}^k], \tag{25}$$

$$\mathbb{E}[\mathcal{H}^{k+1}] \leq \left(1 - \frac{\alpha}{2}\right)\mathbb{E}[\mathcal{H}^k] + 2p\left(1 + \frac{2p}{\alpha}\right)(\mathbb{E}[\mathcal{G}_w^k] + \mathbb{E}[\mathcal{G}_y^k]). \tag{26}$$

Now we define a Lyapunov function Ψ^k for $k \geq 1$ as

$$\Psi^k = \lambda_{k-1}\mathcal{W}^k + \frac{2\gamma_{k-1}\beta}{\theta_{1,k-1}}\mathcal{Y}^k + \mathcal{Z}^k + \frac{10\eta_{k-1}\omega(1+\omega)\gamma_{k-1}\beta}{\theta_{1,k-1}n}\mathcal{H}^k, \quad \forall k \geq 1, \tag{27}$$

where $\lambda_k = \frac{\gamma_k\beta}{p\theta_{1,k}}(\theta_{1,k} + \theta_2 - p + \sqrt{(p - \theta_{1,k} - \theta_2)^2 + 4p\theta_2})$. Furthermore, it is straightforward to verify that

$$\frac{2\gamma_k\beta\theta_2}{p\theta_{1,k}} \leq \lambda_k \leq \frac{2\gamma_k\beta(\theta_{1,k} + \theta_2)}{p\theta_{1,k}}.$$

Now we restate the convergence result in the strongly-convex case in Theorem 3 and prove it using Lemma 6, 7 and the Lyapunov function.

Theorem 4. *If $\mu > 0$ and parameters satisfy $\eta_k \equiv \eta = n\theta_2/(120\omega L)$, $\theta_{1,k} \equiv \theta_1 = 1/(3\sqrt{\kappa})$, $\alpha = p = 1/(1+\omega)$, $\gamma_k \equiv \gamma = \eta/(2\theta_1 + \eta\mu)$, $\beta = 2\theta_1/(2\theta_1 + \eta\mu)$, and $\theta_2 = 1/(3\sqrt{n} + 3n/\omega)$, then the number of communication rounds performed by ADIANA to find an ϵ-accurate solution such that $\mathbb{E}[f(\hat{x})] - \min_x f(x) \leq \epsilon$ is at most $\mathcal{O}((\omega + (1+\omega/\sqrt{n})\sqrt{\kappa})\ln(L\Delta/\epsilon))$.*

Proof. In the strongly-convex case, parameters $\{\gamma_k\}_{k\geq 1}$ and $\{\theta_{1,k}\}_{k\geq 1}$ are constants, then so is λ_k. Thus, we simply write $\gamma \triangleq \gamma_k$, $\theta_1 \triangleq \theta_{1,k}$, and $\lambda \triangleq \lambda_k$ for all $k \geq 1$. Considering (20)+λ(24)+$\frac{5\gamma\beta\eta}{4\theta_1}$(25)+$\frac{10\eta\omega(1+\omega)\gamma\beta}{n\theta_1}$(26), we have

$$\begin{aligned}\mathbb{E}[\Psi^{k+1}] \leq& \left(\frac{2\gamma\beta\theta_2}{\theta_1} + (1-p)\lambda\right)\mathcal{W}^k + \left(\frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1} + p\lambda\right)\mathcal{Y}^k + \beta\mathcal{Z}^k \\ &+ \left(1 - \frac{1}{4(1+\omega)}\right)\frac{10\eta\omega(1+\omega)\gamma\beta}{\theta_1 n}\mathcal{H}^k - \left(\frac{\gamma\beta\theta_2}{L\theta_1} - \frac{125\gamma\beta\eta\omega}{2n\theta_1}\right)\mathcal{G}_w^k \\ &- \left(\frac{\gamma\beta(1-\theta_1-\theta_2)}{L\theta_1} - \frac{60\eta\omega\gamma\beta}{n\theta_1}\right)\mathcal{G}_y^k. \end{aligned} \tag{28}$$

By the definition of λ, we have

$$\begin{aligned}\frac{2\gamma\beta\theta_2}{\theta_1} + (1-p)\lambda =& \lambda\left(1 - p + \frac{2p\theta_2}{\sqrt{(p-\theta_1-\theta_2)^2 + 4p\theta_2} + \theta_1 + \theta_2 - p}\right) \\ =& \lambda\left(1 - p + \frac{2p\theta_2}{2\theta_2 + \frac{4\theta_1\theta_2}{\sqrt{(p-\theta_1-\theta_2)^2+4p\theta_2} - \theta_1 + \theta_2 + p}}\right) \\ \leq& \lambda\left(1 - p + \frac{p}{1 + \frac{2\theta_1}{(p+\theta_1+\theta_2) - \theta_1 + \theta_2 + p}}\right) = \left(1 - \frac{p\theta_1}{p + \theta_1 + \theta_2}\right)\lambda, \end{aligned} \tag{29}$$

and

$$\begin{aligned}\frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1} + p\lambda =& \frac{2\gamma\beta}{\theta_1}\left[1 - \theta_1 - \theta_2 + \frac{1}{2}\left(\theta_1 + \theta_2 - p + \sqrt{(p-\theta_1-\theta_2)^2 + 4p\theta_2}\right)\right] \\ =& \frac{2\gamma\beta}{\theta_1}\left(1 - \frac{2p\theta_1}{p + \theta_1 + \theta_2 + \sqrt{(p-\theta_1-\theta_2)^2 + 4p\theta_2}}\right) \\ \leq& \left(1 - \frac{p\theta_1}{p + \theta_1 + \theta_2}\right)\frac{2\gamma\beta}{\theta_1}. \end{aligned} \tag{30}$$

From the choice of η, it is easy to verify that

$$\frac{\gamma\beta\theta_2}{L\theta_1} - \frac{5\gamma\beta\eta\omega}{2n\theta_1} - \frac{60\eta\omega\gamma\beta}{n\theta_1} \geq 0, \tag{31}$$

and further noting $1 - \theta_1 - \theta_2 \geq \theta_2$,

$$\frac{\gamma\beta(1-\theta_1-\theta_2)}{L\theta_1} - \frac{60\eta\omega\gamma\beta}{n\theta_1} \geq 0. \tag{32}$$

Plugging (29), (30), (31), and (32) into (28), we obtain

$$\begin{aligned}\mathbb{E}[\Psi^{k+1}] \leq& \left(1 - \min\left\{\frac{p\theta_1}{p+\theta_1+\theta_2}, \frac{\eta\mu}{2\theta_1+\eta\mu}, \frac{1}{4(1+\omega)}\right\}\right)\Psi^k \\ \leq& \left(1 - \frac{1}{\frac{p+\theta_1+\theta_2}{p\theta_1} + \frac{2\theta_1+\eta\mu}{\eta\mu} + 4(1+\omega)}\right)\Psi^k \\ \leq& \left(1 - \frac{1}{250\left(\omega + \left(1 + \frac{\omega}{\sqrt{n}}\right)\sqrt{\kappa}\right)}\right)\Psi^k, \quad \forall k \geq 0, \end{aligned} \tag{33}$$

where $\Psi^0 := \lambda\mathcal{W}^0 + \frac{2\gamma\beta}{\theta_1}\mathcal{Y}^0 + \mathcal{Z}^0 + \frac{10\eta\omega(1+\omega)\gamma\beta}{\theta_1 n}\mathcal{H}^0$. Note that since we use initialization $y^0 = z^0 = w^0 = h_i^0 = h^0$, $\forall 1 \leq i \leq n$, we have $\mathcal{W}^0 = \mathcal{Y}^0 \leq (L\Delta)/2$, $\mathcal{Z}^0 \leq \Delta$, $\mathcal{H}^0 \leq L^2\Delta$, which indicates that

$$\Psi^0 \leq \frac{L}{2} \cdot (\lambda_W + \lambda_Y + \lambda_Z + \lambda_H)\Delta,$$

where $\lambda_W = \lambda \geq \frac{2\gamma\beta\theta_2}{\theta_1 p}$, $\lambda_Y = \frac{2\gamma\beta}{\theta_1}$, $\lambda_Z = \frac{2}{L}$, $\lambda_H = \frac{20\eta\omega(1+\omega)\gamma\beta L}{\theta_1 n}$. These coefficients have the following inequalities:

$$\begin{aligned}\lambda_W + \lambda_Y \geq & \frac{4\eta(\theta_2+p)}{p(2\theta_1+\eta\mu)^2} = \frac{n\theta_2(\theta_2+p)}{30\omega Lp(2/3\sqrt{\kappa}+n\theta_2/120\omega\kappa)^2} \geq \frac{n\theta_2(\theta_2+p)\kappa}{15\omega Lp} \\ \geq & \frac{\kappa}{135L} \geq \frac{1}{270}\lambda_Z,\end{aligned}$$

and

$$\frac{3}{32}(\lambda_W + \lambda_Y) \geq \frac{\kappa}{1440L} \geq \frac{(1+\omega)n\theta_2^2\kappa}{160\omega L} \geq \frac{40\eta^2\omega(1+\omega)L}{(2\theta_1+\eta\mu)^2 n} = \lambda_H.$$

Consequently, the initial value of the Lyapunov function can be bounded as

$$\Psi^0 \leq 136L(\lambda_W + \lambda_Y)\Delta,$$

which together with (33) further implies that

$$\begin{aligned}&\min\{\mathbb{E}[f(w^T)], \mathbb{E}[f(y^T)]\} - f^\star \\ \leq & \min\left\{\frac{1}{\lambda_W}, \frac{1}{\lambda_Y}\right\}\left(1 - \frac{1}{250\left(\omega + \left(1+\frac{\omega}{\sqrt{n}}\right)\sqrt{\kappa}\right)}\right)^T \Psi^0 \\ \leq & 272L\Delta\left(1 - \frac{1}{250\left(\omega + \left(1+\frac{\omega}{\sqrt{n}}\right)\sqrt{\kappa}\right)}\right)^T.\end{aligned}$$

Thus, $\mathcal{O}\left(\left(\omega + \left(1+\frac{\omega}{\sqrt{n}}\right)\sqrt{\kappa}\right)\ln\left(\frac{L\Delta}{\epsilon}\right)\right)$ iterations are sufficient to guarantee an ϵ-solution. □

D.2 Generally-convex case

In this subsection, we restate the convergence result in the generally-convex case as in Theorem 3 and prove it using Lemma 6, 7 and the Lyapunov function defined in (27).

Theorem 5. *If $\mu = 0$ and parameters satisfy $\alpha = 1/(1+\omega)$, $\beta = 1$, $p = \theta_2 = 1/(3(1+\omega))$, $\theta_{1,k} = 9/(k+27(1+\omega))$, $\gamma_k = \eta_k/(2\theta_{1,k})$, and*

$$\eta_k = \min\left\{\frac{k+1+27(1+\omega)}{9(1+\omega)^2(1+27(1+\omega))L}, \frac{3n}{200\omega(1+\omega)L}, \frac{1}{2L}\right\},$$

then the number of communication rounds performed by ADIANA to find an ϵ-accurate solution such that $\mathbb{E}[f(\hat{x})] - \min_x f(x) \leq \epsilon$ is provided by $\mathcal{O}((1+\omega/\sqrt{n})\sqrt{L\Delta/\epsilon} + (1+\omega)\sqrt[3]{L\Delta/\epsilon})$.

Proof. Considering (20) + λ_k(24) + $\frac{5\gamma_k\beta\eta_k}{4\theta_{1,k}}$(25) + $\frac{10\eta_k\omega(1+\omega)\gamma_k\beta}{n\theta_{1,k}}$(26) and applying the choice of θ_2, p and α, we have

$$\begin{aligned}&\mathbb{E}_k[\Psi^{k+1}] \\ \leq & \left(\frac{2\gamma_k\beta\theta_2}{\theta_{1,k}} + (1-p)\lambda_k\right)\mathcal{W}^k + \left(\frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}} + p\lambda_k\right)\mathcal{Y}^k + \beta\mathcal{Z}^k \\ & + \left(1 - \frac{1}{4(1+\omega)}\right)\frac{10\eta_k\omega(1+\omega)\gamma_k\beta}{n\theta_{1,k}}\mathcal{H}^k - \left(\frac{\gamma_k\beta\theta_2}{L\theta_{1,k}} - \frac{5\omega\gamma_k\beta\eta_k}{2n\theta_{1,k}} - \frac{100\eta_k\omega\gamma_k\beta}{9n\theta_{1,k}}\right)\mathcal{G}_w^k \\ & - \left(\frac{\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{L\theta_{1,k}} - \frac{100\eta_k\omega\gamma_k\beta}{9n\theta_{1,k}}\right)\mathcal{G}_y^k. \end{aligned} \tag{34}$$

Similar to the proof of Theorem 4, we can simplify (34) by validating

$$\begin{cases} \frac{2\gamma_k\beta\theta_2}{\theta_{1,k}} + (1-p)\lambda_k \le \left(1 - \frac{p\theta_{1,k}}{p+\theta_{1,k}+\theta_2}\right)\lambda_k \le \left(1 - \frac{\theta_{1,k}}{3}\right)\lambda_k, \\ \frac{2\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{\theta_{1,k}} + p\lambda_k \le \left(1 - \frac{p\theta_{1,k}}{p+\theta_{1,k}+\theta_2}\right)\frac{2\gamma_k\beta}{\theta_{1,k}} \le \left(1 - \frac{\theta_{1,k}}{3}\right)\frac{2\gamma_k\beta}{\theta_{1,k}}, \\ \frac{\gamma_k\beta\theta_2}{L\theta_{1,k}} - \frac{5\omega\gamma_k\beta\eta_k}{2n\theta_{1,k}} - \frac{100\eta_k\omega\gamma_k\beta}{9n\theta_{1,k}} \ge 0, \\ \frac{\gamma_k\beta(1-\theta_{1,k}-\theta_2)}{L\theta_{1,k}} - \frac{100\eta_k\omega\gamma_k\beta}{9n\theta_{1,k}} \ge 0, \end{cases}$$

and then obtain

$$\begin{aligned} \mathbb{E}_k[\Psi^{k+1}] \le & \left(1 - \frac{\theta_{1,k}}{3}\right)\lambda_k\mathcal{W}^k + \left(1 - \frac{\theta_{1,k}}{3}\right)\frac{2\gamma_k}{\theta_{1,k}}\mathcal{Y}^k + \mathcal{Z}^k \\ & + \left(1 - \frac{1}{4(1+\omega)}\right)\frac{10\eta_k\omega(1+\omega)\gamma_k}{\theta_{1,k}n}\mathcal{H}^k. \end{aligned} \tag{35}$$

For $\forall k \ge 1$, we have $\theta_{1,k} \le \theta_{1,k-1}$ and thus

$$\begin{aligned} \left(1 - \frac{\theta_{1,k}}{3}\right)\lambda_k =& \left(1 - \frac{3}{k+27(1+\omega)}\right)\frac{\eta_k}{2p\theta_{1,k}^2}\left(\theta_{1,k} + \sqrt{\theta_{1,k}^2 + 4p\theta_2}\right) \\ \le& \left(1 - \frac{3}{k+27(1+\omega)}\right)\frac{\eta_k}{2p\theta_{1,k}^2}(\theta_{1,k-1} + \sqrt{\theta_{1,k-1}^2 + 4p\theta_2}) \\ =& \left(1 - \frac{3}{k+27(1+\omega)}\right)\left(\frac{k+27(1+\omega)}{k-1+27(1+\omega)}\right)^2\frac{\eta_k}{\eta_{k-1}}\lambda_{k-1}. \end{aligned}$$

Further noting $\frac{\eta_k}{\eta_{k-1}} \le 1 + \frac{1}{k+27(1+\omega)}$, we obtain

$$\begin{aligned} \left(1 - \frac{\theta_{1,k}}{3}\right)\lambda_k \le & \left(1 - \frac{3}{k+27(1+\omega)}\right)\left(1 - \frac{1}{k+27(1+\omega)}\right)^{-2}\left(1 + \frac{1}{k+27(1+\omega)}\right)\lambda_{k-1} \\ \le & \lambda_{k-1}. \end{aligned} \tag{36}$$

Similarly,

$$\begin{aligned} & \left(1 - \frac{\theta_{1,k}}{3}\right)\frac{2\gamma_k}{\theta_{1,k}} \\ =& \left(1 - \frac{3}{k+27(1+\omega)}\right)\left(\frac{k+27(1+\omega)}{k-1+27(1+\omega)}\right)^2\frac{\eta_k}{\eta_{k-1}}\frac{2\gamma_{k-1}}{\theta_{1,k-1}} \\ \le& \left(1 - \frac{3}{k+27(1+\omega)}\right)\left(1 - \frac{1}{k+27(1+\omega)}\right)^{-2}\left(1 + \frac{1}{k+27(1+\omega)}\right)\frac{2\gamma_{k-1}}{\theta_{1,k-1}} \\ \le& \frac{2\gamma_{k-1}}{\theta_{1,k-1}}, \end{aligned} \tag{37}$$

and

$$\begin{aligned} & \left(1 - \frac{1}{4(1+\omega)}\right)\frac{10\eta_k\omega(1+\omega)\gamma_k\beta}{n\theta_{1,k}} \\ =& \left(1 - \frac{1}{4(1+\omega)}\right)\left(\frac{k+27(1+\omega)}{k-1+27(1+\omega)}\right)^2\left(\frac{\eta_k}{\eta_{k-1}}\right)^2\frac{10\eta_{k-1}\omega(1+\omega)\gamma_{k-1}\beta}{n\theta_{1,k-1}} \\ \le& \frac{\left(1 - \frac{5}{k+27(1+\omega)}\right)\left(1 + \frac{3}{k+27(1+\omega)}\right)}{\left(1 - \frac{1}{k+27(1+\omega)}\right)^2}\frac{10\eta_{k-1}\omega(1+\omega)\gamma_{k-1}\beta}{n\theta_{1,k-1}} \\ \le& \frac{10\eta_{k-1}\omega(1+\omega)\gamma_{k-1}\beta}{\theta_{1,k-1}n}. \end{aligned} \tag{38}$$

Combining (35),(36),(37), and (38), we have for $\forall\, k \geq 1$ that

$$\mathbb{E}_k[\Psi^{k+1}] \leq \Psi^k. \tag{39}$$

By applying (39) with $k = T-1, T-2, \cdots, 1$ and (35) with $k = 0$, we obtain

$$\begin{aligned}
\mathbb{E}[\Psi^T] \leq & \left(1 - \frac{\theta_{1,0}}{3}\right) \frac{2\gamma_0(\theta_{1,0}+\theta_2)}{p\theta_{1,0}} \mathcal{W}^0 + \left(1 - \frac{\theta_{1,0}}{3}\right) \frac{2\gamma_0}{\theta_{1,0}} \mathcal{Y}^0 + \mathcal{Z}^0 \\
& + \left(1 - \frac{1}{4(1+\omega)}\right) \frac{10\eta_0\omega(1+\omega)\gamma_0}{\theta_{1,0} n} \mathcal{H}^0 \\
\leq & \frac{2}{L}\mathcal{W}^0 + \frac{1}{L}\mathcal{Y}^0 + \mathcal{Z}^0 + \frac{3}{40L^2}\mathcal{H}^0 \leq \left(1 + \frac{1}{2} + 1 + \frac{3}{40}\right)\Delta \leq 3\Delta.
\end{aligned}$$

Note that

$$\Psi^T \geq \lambda_{T-1}\mathcal{W}^T + \frac{2\gamma_{T-1}\beta}{\theta_{1,T-1}}\mathcal{Y}^T \geq \frac{2\gamma_{T-1}\beta\theta_2}{\theta_{1,T-1}p}\mathcal{W}^T + \frac{2\gamma_{T-1}\beta}{\theta_{1,T-1}}\mathcal{Y}^T = \frac{\eta_{T-1}}{\theta_{1,T-1}^2}(\mathcal{W}^T + \mathcal{Y}^T),$$

thus

$$\begin{aligned}
& \max\{\mathbb{E}[f(w^T)], \mathbb{E}[f(y^T)]\} - f^\star \\
\leq & \frac{\theta_{1,T-1}^2}{\eta_{T-1}} \mathbb{E}[\Psi^T] \\
\leq & \frac{243\Delta}{(T-1+27(1+\omega))^2} \cdot \max\left\{\frac{9(1+\omega)^2(1+27(1+\omega))L}{T+27(1+\omega)}, \frac{200\omega(1+\omega)L}{3n}, 2L\right\} \\
= & \mathcal{O}\left(\frac{(1+\omega^2/n)L\Delta}{T^2} + \frac{(1+\omega^3)L\Delta}{T^3}\right),
\end{aligned}$$

thus it suffices to achieve an ϵ-solution with $\mathcal{O}\left(\left(1 + \frac{\omega}{\sqrt{n}}\right)\sqrt{\frac{L\Delta}{\epsilon}} + (1+\omega)\sqrt[3]{\frac{L\Delta}{\epsilon}}\right)$ iterations. $\square$

E Correction on CANITA [34]

We observe that when $\omega \gg n$, the original convergence rate of CANITA [34] contradicts the lower bounds presented in our Theorem 2. This discrepancy may stem from errors in the derivation of equations (35) and (36) in [34], or from the omission of certain conditions such as $\omega = \Omega(n)$. To address this issue, we provide a corrected proof and the corresponding convergence rate. Here we modify the choice of β_0 in ([34], Theorem 2) to $9(1+b+\omega)^2/(2(1+b))$, while keeping all other choices consistent with the original proof, *i*.e., $b = \min\{\omega, \sqrt{\omega(1+\omega)^2/n}\}$, $p_t \equiv 1/(1+b)$, $\alpha_t \equiv 1/(1+\omega)$, $\theta_t = 3(1+b)/(t+9(1+b+\omega))$, $\beta = 48\omega(1+\omega)(1+b+2(1+\omega))/(n(1+b)^2)$ and

$$\eta_t = \begin{cases} \frac{1}{L(\beta_0+3/2)}, & \text{for } t = 0, \\ \min\left\{\left(1 + \frac{1}{t+9(1+b+\omega)}\right)\eta_{t-1}, \frac{1}{L(\beta+3/2)}\right\}, & \text{for } t \geq 1. \end{cases}$$

By definition we have

$$\begin{aligned}
\eta_T = & \min\left\{\frac{T+1+9(1+b+\omega)}{1+9(1+b+\omega)}\eta_0, \frac{1}{L(\beta+3/2)}\right\} \\
= & \min\left\{\frac{T+1+9(1+b+\omega)}{1+9(1+b+\omega)}\frac{1}{L(\beta_0+3/2)}, \frac{1}{L(\beta+3/2)}\right\} \\
\geq & \min\left\{\frac{(T+9(1+b+\omega))(1+b)}{60L(1+b+\omega)^3}, \frac{1}{L(\beta+3/2)}\right\}
\end{aligned} \tag{40}$$

Plugging (40) and ([34],34) into ([34],33), we obtain

$$\begin{aligned}
\mathbb{E}[F^{T+1}] = & \mathcal{O}\left(\frac{(1+b+\omega)^3 L\Delta}{(T+9(1+b+\omega))^3} + \frac{(1+b)(\beta+3/2)L\Delta}{(T+9(1+b+\omega))^2}\right) \\
= & \mathcal{O}\left(\frac{(1+b+\omega)^3 L\Delta}{T^3} + \frac{(1+b)(\beta+3/2)L\Delta}{T^2}\right).
\end{aligned} \tag{41}$$

Using $b = \min\{\omega, \sqrt{\omega(1+\omega)^2/n}\}$, we have

$$(1+b+\omega)^3 = \Theta\left((1+\omega)^3\right),$$

and

$$\begin{aligned}(1+b)(\beta+3/2) =& \Theta\left((1+b) + \frac{\omega(1+\omega)(1+b+\omega)}{n(1+b)}\right)\\ =& \Theta\left(1 + \frac{\omega^{3/2}}{n^{1/2}} + \frac{\omega^2}{n}\right),\end{aligned}$$

thus (41) can be simplified as

$$\mathbb{E}[F^{T+1}] = \mathcal{O}\left(\frac{(1+\omega)^3 L\Delta}{T^3} + \frac{(1+\omega^{3/2}/n^{1/2}+\omega^2/n)L\Delta}{T^2}\right).$$

Consequently, for $\epsilon < L\Delta/2$ (*i*.e., a precision that the initial point does not satisfy), the communication rounds to achieve precision ϵ is given by $\mathcal{O}\left(\omega\frac{\sqrt[3]{L\Delta}}{\sqrt[3]{\epsilon}} + \left(1 + \frac{\omega^{3/4}}{n^{1/4}} + \frac{\omega}{\sqrt{n}}\right)\frac{\sqrt{L\Delta}}{\sqrt{\epsilon}}\right)$.

F Experimental details and additional results

This section provides more details of the experiments listed in Sec. 6, as well as a few new experiments to validate our theories.

F.1 Experimental details

This section offers a comprehensive and detailed description of the experiments listed in Sec. 6, including problem formulation, data generation, cost calculation, and algorithm implementation.

Least squares. The local objective function of node i is defined as $f_i(x) := \frac{1}{2}\|A_i x - b_i\|^2$, where $A_i \in \mathbb{R}^{M\times d}$, $b_i \in \mathbb{R}^M$. We set $d = 20$, $M = 25$, and the number of nodes $n = 400$. To generate A_i's, we first randomly generate a Gaussian matrix $G \in \mathbb{R}^{nM\times d}$; we then apply the SVD decomposition $G = U\Sigma V^\top$ and replace the singular values in Σ by an arithmetic sequence starting from 1 and ending at 100 to get $\tilde{\Sigma}$ and the resulted data matrix $\tilde{G} = U\tilde{\Sigma}V^\top$; we finally allocate the submatrix of $\tilde{G}$ composed of the $((i-1)M+1)$-th row to the (iM)-th row to be A_i for all $1 \le i \le n$.

Logistic regression. The local objective function of node i is defined as $f_i(x) := \frac{1}{M}\sum_{m=1}^M \ln(1 + \exp(-b_{i,m}a_{i,m}^\top x)$, where number of nodes $n = 400$, $a_{i,m}$ stands for the feature of the m-th datapoint in the node i's dataset, and $b_{i,m}$ stands for the corresponding label. In `a9a` dataset, node i owns the $(81(i-1)+1)$-th to the $(81i)$-th datapoint with feature dimension $d = 123$. In `w8a` dataset, node i owns the $(120(i-1)+1)$-th to the $(120i)$-th datapoint with feature dimension $d = 300$.

Constructed problem. The local objective function of node i is defined as

$$f_i(x) := \begin{cases}\frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{4}([x]_1^2 + \sum_{1\le r\le d/2-1}([x]_{2r} - [x]_{2r+1})^2 + [x]_d^2 - 2[x]_1), & \text{if } i \le n/2,\\ \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{4}(\sum_{1\le r\le d/2}([x]_{2r-1} - [x]_{2r})^2), & \text{if } i > n/2,\end{cases}$$

where $[x]_l$ denotes the l-th entry of vector $x \in \mathbb{R}^d$. We set $\mu = 1$, $L = 10^4$, $d = 20$ and number of nodes $n = 400$.

Compressors. We apply various compressors to the algorithms with communication compression through our experiments. In the constructed quadratic problem, we consider ADIANA algorithm with random-s compressors (see Example 1 in Appendix A) in six different settings, *i*.e., three choices of s ($s = 1, 2, 4$), with two different (shared or independent) randomness settings. In the least squares and logistic regression problems, we apply the independent random-$\lfloor d/20\rfloor$ compressor to ADIANA, CANITA and DIANA algorithm. In particular, we use the unscaled version of the independent random-$\lfloor d/20\rfloor$ compressor for EF21 to guarantee convergence, where the values of selected entries are transmitted directly to the server without being scaled by d/s times. In Appendix F.2, we further apply independent *natural compression* [20] and *random quantization* [3] with $s = \lceil\sqrt{d}\rceil$ in the above algorithms.

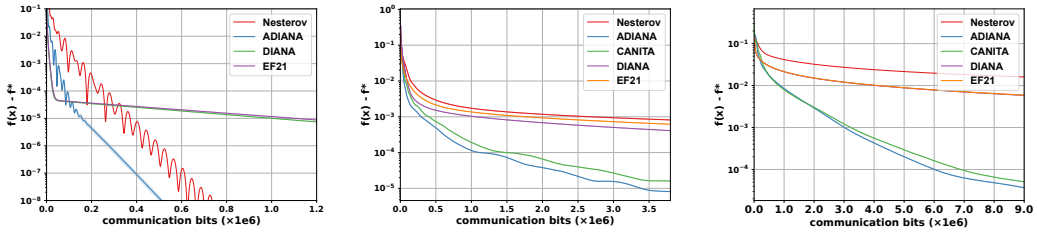

Figure 3: Convergence results of various distributed algorithms on a synthetic least squares problem (left), logistic regression problems with dataset `a9a` (middle) and `w8a` (right). The y-axis represents $f(\hat{x}) - f^\star$ and the x-axis indicates the total communicated bits sent by per worker. All compressors used are independent natural compression.

Total communicated bits. For non-compression algorithms and algorithms with a fixed-length compressor, such as random-s and natural compression, the total communication bits can be calculated using the following formula: *total communication bits = number of iterations* $\times$ *communication rounds per iteration* $\times$ *communicated bits per round*. Among the algorithms we compare, ADIANA and CANITA communicate twice per iteration, while the other algorithms communicate only once. The communicated bits per round for non-compression algorithms amount to $64d$ for d `float64` entries. In the case of the random-s compressor, the communicated bits per communication are calculated as $64s + \lceil \log_2 \binom{d}{s} \rceil$. Similarly, for natural compression, the communicated bits per round are fixed at $12d$, with 1 sign bit and 11 exponential bits allocated for each entry. In the case of adaptive-length random quantization, the communication cost is evaluated using *Elias* integer encoding [16]. This cost is then averaged among n nodes, providing a more representative estimate.

Algorithm implementation. We implement ADIANA, CANITA, DIANA, EF21 algorithms following the formulation in Algorithm 1, [34], [27], and [50], respectively. We implement Nesterov's accelerated algorithm with the following recursions:

$$\begin{cases} y^k = (1-\theta_t)x^k + \theta_t z^k, \\ x^{k+1} = y^k - \eta_k \nabla f(y^k), \\ z^{k+1} = x^k + \frac{1}{\theta_k}(x^{k+1} - x^k). \end{cases}$$

The value of α in ADIANA, CANITA and DIANA are all set to $1/(1+\omega)$, and we set γ_k, β of ADIANA as in Theorem 3. Other parameters are all selected through running Bayesian Optimization [46] for the first 20% iterations with 5 initial points and 20 trials. The exact value of the selected parameters are listed in Appendix F.3. Each curve (except for Nesterov's accelerated algorithm which does not involve randomness) is averaged through 20 trials, with the range of standard deviation depicted.

Computational resource. All experiments are run on an NVIDIA A100 server. Each trial consumes up to 10 minutes of running time.

F.2 Additional experiments

Addtional compressors. In addition to the experiments in Sec. 6, we consider applying different compressors in the algorithms with communication compression. Fig. 3 and Fig. 4 show results of using natural compression and random quantization, respectively. These results are consistent with the results in Sec. 6.

CIFAR-10 dataset. We also consider binominal logistic regression with CIFAR-10 dataset, where labels of each datum are categorized by whether they equal to 3, *i*.e., the corresponding figures belong to the `cat` category. The full training set with 50000 images, are devided equally to $n = 250$ nodes. The compressor choices follow the same strategies as in Appendix F.1, where dimension $d = 3072$. Fig. 5 compares convergence results between Nesterov method and ADIANA with different compressors. It can be observed that ADIANA equipped with more aggressive compressors, *i*.e., those with bigger ω, benefits more from the compression, which is consistent with our theoretical results.

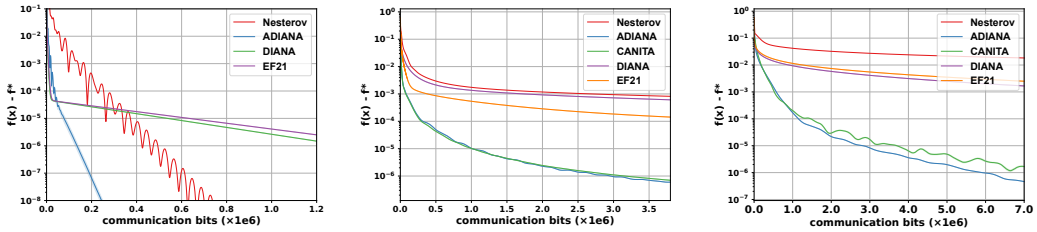

Figure 4: Convergence results of various distributed algorithms on a synthetic least squares problem (left), logistic regression problems with dataset `a9a` (middle) and `w8a` (right). The y-axis represents $f(\hat{x}) - f^\star$ and the x-axis indicates the total communicated bits sent by per worker. All compressors used are independent random quantization.

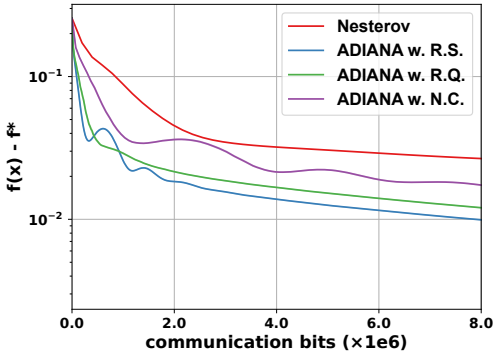

Figure 5: Experimental results of logistic regression problem on the CIFAR-10 dataset. The objective function is constructed by relabeling the 10 classes into 2 classes, namely `cat` (corresponding to the original cat class) and `non-cat` (corresponding to the rest classes). ADIANA w. R.S. / R.Q. / N.C. represents ADIANA algorithm with random-$\lfloor d/20 \rfloor$ compressor / random quantization compressor with $s = \lceil \sqrt{d} \rceil$ / natural compression compressor, where $d = 3072$ is the dimension of gradient vectors as well as the number of features in CIFAR-10 dataset. The experiments are conducted under the same setting as in the "Algorithm implementation" part in Appendix F.1.

F.3 Parameter values

In this subsection, we list all the parameter values that are selected by applying Bayesian Optimization. Table 2, 3, 4, 5, 6 list the parameters chosen in the least squares problem, logistic regression using `a9a` dataset, logistic regression using `w8a` dataset, the constructed problem, and logistic regression using CIFAR-10 dataset, respectively.

Table 2: Parameters for algorithms in the least squares problem. Notation R.S. stands for independent random sparsification, N.C. stands for independent natural compression, R.Q. stands for independent random quantization.

Algorithm	Parameters
Nesterov	$\eta = 3.0 \times 10^{-2}, \theta = 1.4 \times 10^{-2}$.
ADIANA R.S.	$\eta = 4.8 \times 10^{-2}, \theta_1 = 2.2 \times 10^{-2}, \theta_2 = 7.6 \times 10^{-2}, p = 4.1 \times 10^{-2}$.
ADIANA N.C.	$\eta = 3.9 \times 10^{-2}, \theta_1 = 1.0 \times 10^{-2}, \theta_2 = 2.9 \times 10^{-1}, p = 9.9 \times 10^{-1}$.
ADIANA R.Q.	$\eta = 6.5 \times 10^{-2}, \theta_1 = 1.4 \times 10^{-2}, \theta_2 = 2.7 \times 10^{-1}, p = 5.5 \times 10^{-1}$.
DIANA R.S.	$\gamma = 7.9 \times 10^{-2}$.
DIANA N.C.	$\gamma = 7.4 \times 10^{-2}$.
DIANA R.Q.	$\gamma = 7.6 \times 10^{-2}$.
EF21 R.S.	$\gamma = 6.2 \times 10^{-2}$.
EF21 N.C.	$\gamma = 6.8 \times 10^{-2}$.
EF21 R.Q.	$\gamma = 7.4 \times 10^{-2}$.

Table 3: Parameters for algorithms in the logistic regression problem with `a9a` dataset. Notation k stands for the index of iteration. Other notations are as in Table 2.

Algorithm	Parameters
Nesterov	$\eta = 9.4 \times 10^{-1}, \theta = 1.7 \times 10^{-1}$.
ADIANA R.S.	$\eta = 2.1, \theta_1 = \frac{1.3 \times 10^1}{k + 5.2 \times 10^2}, \theta_2 = 2.1 \times 10^{-1}, p = 7.7 \times 10^{-1}$.
ADIANA N.C.	$\eta = 2.1, \theta_1 = \frac{1.0}{k+4.3}, \theta_2 = 8.0 \times 10^{-3}, p = 8.0 \times 10^{-1}$.
ADIANA R.Q.	$\eta = 2.2, \theta_1 = \frac{1.3}{k+1.3}, \theta_2 = 1.5 \times 10^{-1}, p = 8.5 \times 10^{-1}$.
CANITA R.S.	$\eta = \min\{\frac{k + 2.1 \times 10^2}{2.1 \times 10^2}, 1.4\}, \theta = \frac{2.0 \times 10^1}{k + 2.3 \times 10^2}, p = 7.8 \times 10^{-1}$.
CANITA N.C.	$\eta = 1.2, \theta = \frac{2.1}{k + 1.2 \times 10^1}, p = 5.2 \times 10^{-1}$.
CANITA R.Q.	$\eta = 2.0, \theta = \frac{3.0}{k+3.0}, p = 7.2 \times 10^{-1}$.
DIANA N.C.	$\gamma = 2.6$.
DIANA R.S.	$\gamma = 9.4 \times 10^{-1}$.
DIANA R.Q.	$\gamma = 4.7 \times 10^{-1}$
EF21 R.S.	$\gamma = 1.3$.
EF21 N.C.	$\gamma = 1.6$.
EF21 R.Q.	$\gamma = 2.7$.

Table 4: Parameters for algorithms in logistic regression with `w8a` dataset. Notations are as in Table 3.

Algorithm	Parameters
Nesterov	$\eta = 1.5 \times 10^{1}, \theta = 9.4 \times 10^{-1}$.
ADIANA R.S.	$\eta = \min\{\frac{k+4.1\times 10^{2}}{1.2\times 10^{2}}, 15\}, \theta_1 = \frac{8.8}{k+4.8\times 10^{2}}, \theta_2 = 2.4 \times 10^{-2}, p = 3.6 \times 10^{-1}$.
ADIANA N.C.	$\eta = 1.5 \times 10^{1}, \theta_1 = \frac{2.5}{k+1.1\times 10^{1}}, \theta_2 = 6.7 \times 10^{-1}, p = 8.3 \times 10^{-1}$.
ADIANA R.Q.	$\eta = 1.5 \times 10^{1}, \theta_1 = \frac{1.9}{k+7.4}, \theta_2 = 4.2 \times 10^{-1}, p = 9.9 \times 10^{-1}$.
CANITA R.S.	$\eta = \min\{\frac{k+2.0\times 10^{2}}{2.2\times 10^{2}}, 7.7\}, \theta = \frac{1.1\times 10^{1}}{k+2.3\times 10^{2}}, p = 4.3 \times 10^{-1}$.
CANITA N.C.	$\eta = \min\{\frac{k+1.1\times 10^{1}}{2.7}, 1.1 \times 10^{1}\}, \theta = \frac{5.4}{k+7.4\times 10^{1}}, p = 4.9 \times 10^{1}$.
CANITA R.Q.	$\eta = \min\{\frac{k+1.6\times 10^{1}}{7.3}, 1.5 \times 10^{1}\}, \theta = \frac{2.2}{k+2.2\times 10^{1}}, p = 4.6 \times 10^{-1}$.
DIANA R.S.	$\gamma = 1.5 \times 10^{1}$.
DIANA N.C.	$\gamma = 1.6 \times 10^{1}$.
DIANA R.Q.	$\gamma = 1.5 \times 10^{1}$.
EF21 R.S.	$\gamma = 2.0 \times 10^{1}$.
EF21 N.C.	$\gamma = 1.5 \times 10^{1}$.
EF21 R.Q.	$\gamma = 1.5 \times 10^{1}$.

Table 5: Parameters for algorithms in the constructed problem. Notation i.d.rand-s denotes independent random-s compressor, s.d.rand-s denotes random-s compressor with shared randomness.

Algorithm	Parameters
Nesterov	$\eta = 1.4 \times 10^{-1}, \theta = 1.2 \times 10^{-4}$.
ADIANA i.d.rand-1	$\eta = 1.5 \times 10^{-4}, \theta_1 = 1.8 \times 10^{-1}, \theta_2 = 1.3 \times 10^{-1}, p = 1.5 \times 10^{-1}$.
ADIANA i.d.rand-2	$\eta = 1.5 \times 10^{-4}, \theta_1 = 1.5 \times 10^{-4}, \theta_2 = 5.0 \times 10^{-2}, p = 1.9 \times 10^{-1}$.
ADIANA i.d.rand-4	$\eta = 1.3 \times 10^{-4}, \theta_1 = 9.2 \times 10^{-2}, \theta_2 = 5.0 \times 10^{-2}, p = 2.3 \times 10^{-1}$.
ADIANA s.d.rand-1	$\eta = 1.4 \times 10^{-6}, \theta_1 = 2.0 \times 10^{-2}, \theta_2 = 1.6 \times 10^{-1}, p = 2.7 \times 10^{-2}$.
ADIANA s.d.rand-2	$\eta = 9.6 \times 10^{-6}, \theta_1 = 7.0 \times 10^{-2}, \theta_2 = 4.3 \times 10^{-1}, p = 1.8 \times 10^{-1}$.
ADIANA s.d.rand-4	$\eta = 1.6 \times 10^{-5}, \theta_1 = 6.0 \times 10^{-2}, \theta_2 = 2.1 \times 10^{-1}, p = 1.6 \times 10^{-1}$.

Table 6: Parameters for algorithms in logistic regression with CIFAR-10 dataset. Notations are as in Table 4.

Algorithm	Parameters
Nesterov	$\eta = 1.1 \times 10^{-1}, \theta = 1.5 \times 10^{-1}$.
ADIANA R.S.	$\eta = 1.4 \times 10^{-1}, \theta_1 = \frac{12}{k+3.3\times 10^{2}}, \theta_2 = 9.0 \times 10^{-2}, p = 4.3 \times 10^{-1}$.
ADIANA N.C.	$\eta = 1.4 \times 10^{-1}, \theta_1 = \frac{1.0\times 10^{-2}}{k+7.0}, \theta_2 = 7.0 \times 10^{-1}, p = 8.5 \times 10^{-1}$.
ADIANA R.Q.	$\eta = 1.2 \times 10^{1}, \theta_1 = \frac{8.2}{k+59}, \theta_2 = 8.0 \times 10^{-1}, p = 6.0 \times 10^{-1}$.