

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 PROVING THE LIMITED SCALABILITY OF CENTRALIZED DISTRIBUTED OPTIMIZATION VIA A NEW LOWER BOUND CONSTRUCTION

Anonymous authors

Paper under double-blind review

## ABSTRACT

We consider centralized distributed optimization in the classical federated learning setup, where  $n$  workers jointly find an  $\varepsilon$ -stationary point of an  $L$ -smooth,  $d$ -dimensional nonconvex function  $f$ , having access only to unbiased stochastic gradients with variance  $\sigma^2$ . Each worker requires at most  $h$  seconds to compute a stochastic gradient, and the communication times from the server to the workers and from the workers to the server are  $\tau_s$  and  $\tau_w$  seconds per coordinate, respectively. One of the main motivations for distributed optimization is to achieve scalability with respect to  $n$ . For instance, it is well known that the distributed version of SGD has a variance-dependent runtime term  $h\sigma^2 L\Delta/n\varepsilon^2$ , which improves with the number of workers  $n$ , where  $\Delta := f(x^0) - f^*$ , and  $x^0 \in \mathbb{R}^d$  is the starting point. Similarly, using unbiased sparsification compressors, it is possible to reduce *both* the variance-dependent runtime term and the communication runtime term from  $\tau_w dL\Delta/\varepsilon$  to  $\tau_w dL\Delta/n\varepsilon + \sqrt{\tau_w d h \sigma^2/n\varepsilon} \cdot L\Delta/\varepsilon$ , which also benefits from increasing  $n$ . However, once we account for the communication from the server to the workers  $\tau_s$ , we prove that it becomes infeasible to design a method using unbiased random sparsification compressors that scales both the server-side communication runtime term  $\tau_s dL\Delta/\varepsilon$  and the variance-dependent runtime term  $h\sigma^2 L\Delta/\varepsilon^2$ , better than polylogarithmically in  $n$ , even in the homogeneous (i.i.d.) case, where all workers access the same function or distribution. Indeed, when  $\tau_s \simeq \tau_w$ , our lower bound is  $\tilde{\Omega} \left( \min \left\{ h \left( \frac{\sigma^2}{n\varepsilon} + 1 \right) \frac{L\Delta}{\varepsilon} + \tau_s d \frac{L\Delta}{\varepsilon}, h \frac{L\Delta}{\varepsilon} + h \frac{\sigma^2 L\Delta}{\varepsilon^2} \right\} \right)$ . To establish this result, we construct a new “worst-case” function and develop a new lower bound framework that reduces the analysis to the concentration of a random sum, for which we prove a concentration bound. These results reveal fundamental limitations in scaling distributed optimization, even under the homogeneous (i.i.d.) assumption.

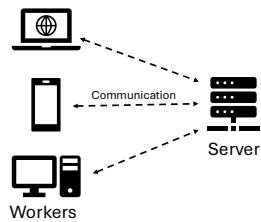
## 1 INTRODUCTION

We focus on the classical federated learning setup, where  $n$  workers, such as CPUs, GPUs, servers, or mobile devices, are connected to a central server via a communication channel (Konečný et al., 2016; McMahan et al., 2017). All workers collaboratively solve a common optimization problem in a distributed fashion by computing stochastic gradients and sharing this information with the server, which then propagates it to other workers. Together, they aim to minimize a smooth nonconvex objective function defined as

$$\min_{x \in \mathbb{R}^d} f(x), \quad (1)$$

where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  and  $d$  is the dimension of  $f$ . We assume that  $d$  is huge, which is indeed the case in modern machine learning and large language model training (Brown et al., 2020; Touvron et al., 2023).

We consider the *homogeneous* (i.i.d.) setting, where all workers have access to stochastic gradients of the same underlying function  $f$ . As the reader will see, the **homogeneous setting assumption**



054 **is a challenge, not a limitation of our work:** all results extend, potentially with even stronger  
 055 implications, to the more general *heterogeneous* (non-i.i.d.) case, when each worker  $i$  works with  
 056  $f_i \neq f$ . We consider the standard assumptions:

057 **Assumption 1.1.**  $f$  is differentiable &  $L$ -smooth, i.e.,  $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \forall x, y \in \mathbb{R}^d$ .

058 **Assumption 1.2.** There exist  $f^* \in \mathbb{R}$  such that  $f(x) \geq f^*$  for all  $x \in \mathbb{R}^d$ . We define  $\Delta :=$   
 059  $f(x^0) - f^*$ , where  $x^0$  is a starting point of methods.

061 For all  $i \in [n]$ , worker  $i$  calculates unbiased stochastic gradients  $\nabla f(x; \xi)$  with  $\sigma^2$ -variance-bounded  
 062 variances, where  $\xi$  is a random variable with some distribution  $\mathcal{D}_\xi$ .

063 **Assumption 1.3 (Homogeneous setting).** For all  $i \in [n]$ , worker  $i$  can only calculate  $\nabla f(x; \xi)$  and  
 064  $\mathbb{E}_\xi[\nabla f(x; \xi)] = \nabla f(x)$  and  $\mathbb{E}_\xi[\|\nabla f(x; \xi) - \nabla f(x)\|^2] \leq \sigma^2$  for all  $x \in \mathbb{R}^d$ , where  $\sigma^2 \geq 0$ .

066 The goal in the nonconvex world is to find an  $\varepsilon$ -stationary point, a (random) point  $\bar{x} \in \mathbb{R}^d$  such that  
 067  $\mathbb{E}[\|\nabla f(\bar{x})\|^2] \leq \varepsilon$  (Nemirovskij & Yudin, 1983; Murty & Kabadi, 1985). We also consider a realistic  
 068 computation and communication scenario:

069 **Assumption 1.4.** Each of the  $n$  workers requires at most  $h$  seconds to compute a stochastic gradient,  
 070 and communication *from the server to any worker* (s2w communication) takes at most  $\tau_s$  seconds per  
 071 coordinate, and communication *from any worker to the server* (w2s communication) takes at most  $\tau_w$   
 072 seconds per coordinate.

074 For instance, under Assumption 1.4, it takes  $d \times \tau_s$  and  $d \times \tau_w$  seconds to send a vector  $v \in \mathbb{R}^d$  from  
 075 the server to any worker and from any worker to the server, respectively. We consider settings with  
 076 bidirectional communication costs, where communication in both directions requires time. Typically,  
 077 especially in the early stages of federated learning algorithm development, most works assume that  
 078 communication *from the server to the workers* is free, i.e.,  $\tau_s = 0$ , which is arguably not true in  
 079 practice: communication over the Internet or 4G/5G networks can be costly in both directions (Huang  
 080 et al., 2012; Narayanan et al., 2021).

## 081 1.1 RELATED WORK

083 **1. Communication is free.** Let us temporarily assume that communication does not take time, i.e.,  
 084  $\tau_s = 0$  and  $\tau_w = 0$ . Then, in this scenario, the theoretically fastest strategy is to run the Synchronized  
 085 SGD method, i.e.,  $x^{k+1} = x^k - \frac{\gamma}{n} \sum_{i=1}^n \nabla f(x^k; \xi_i^k)$ , where  $\gamma = \Theta(\min\{1/L, \varepsilon n / L \sigma^2\})$ ,  $\{\xi_i^k\}$   
 086 are i.i.d., and  $\{\nabla f(x^k; \xi_i^k)\}$  are computed in parallel by the workers, which send to the server  
 087 that aggregates and calculates  $x^{k+1}$ . One can show that the time complexity of this method is  
 088  $\mathcal{O}(h(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L \Delta}{n \varepsilon^2}))$ , because it requires  $\mathcal{O}(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L \Delta}{n \varepsilon^2})$  iterations (Lan, 2020), and each iteration  
 089 takes at most  $h$  seconds due to Assumption 1.4. Moreover, this result is *optimal and can not be*  
 090 *improved* (Arjevani et al., 2022; Tyurin & Richtárik, 2023b).

091 *Observation 1:* One obvious but important observation is that the second “statistical term” in the  
 092 complexity bound scales with  $n$ . The larger the number of workers  $n$ , the smaller the overall time  
 093 complexity of Synchronized SGD, with a linear improvement in  $n$ . This is a theoretical justification  
 094 for the importance of distributed optimization and the use of many workers.

095 **2. Worker-to-server communication can not be ignored.** For now, consider the setup where  
 096 communication from workers to the server takes  $\tau_w > 0$  seconds, while communication from the  
 097 server to the workers is free, i.e.,  $\tau_s = 0$ . In this scenario, the described version of Synchronized  
 098 SGD has a suboptimal  $\mathcal{O}(h(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L \Delta}{n \varepsilon^2}) + \tau_w d(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L \Delta}{n \varepsilon^2}))$  time complexity, because it takes  $h$   
 099 seconds to calculate a stochastic gradient and  $\tau_w d$  seconds to send the stochastic gradients of size  $d$   
 100 to the server, which calculates  $x^{k+1}$ . However, if we slightly modify this method and consider Batch  
 101 Synchronized SGD:

$$102 \quad x^{k+1} = x^k - \frac{\gamma}{n} \sum_{i=1}^n \frac{1}{b} \sum_{j=1}^b \nabla f(x^k; \xi_{ij}^k) \quad (\text{Batch Synchronized SGD})$$

105 with  $b = \Theta(\sigma^2 / \varepsilon n)$  and  $\gamma = \Theta(1/L)$ , then the time complexity becomes

$$106 \quad \mathcal{O}\left(h\left(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L \Delta}{n \varepsilon^2}\right) + \tau_w d \frac{L\Delta}{\varepsilon}\right), \quad (2)$$

108 because the number of iterations reduces to  $\mathcal{O}(L\Delta/\varepsilon)$ . In other words, each worker, instead of immediately sending a gradient, locally aggregates a batch of size  $b$  to reduce the number of communications. It turns out that the last complexity can be improved further with the help of unbiased compressors:

111 **Definition 1.5.** A mapping  $\mathcal{C} : \mathbb{R}^d \times \mathbb{S}_\nu \rightarrow \mathbb{R}^d$  with a distribution  $\mathcal{D}_\nu$  is an *unbiased compressor* if  
 112 there exists  $\omega \geq 0$  such that  $\mathbb{E}_\nu[\mathcal{C}(x; \nu)] = x$  and  $\mathbb{E}_\nu[\|\mathcal{C}(x; \nu) - x\|^2] \leq \omega \|x\|^2$  for all  $x \in \mathbb{R}^d$ . We  
 113  $\mathbb{U}(\omega)$  denote the family of such compressors. The community uses the shorthand  $\mathcal{C}(x; \nu) \equiv \mathcal{C}(x)$ ,  
 114 which we also follow.

116 A standard example of an unbiased compressor is  $\text{Rand}K \in \mathbb{U}(d/K - 1)$ , which selects  $K$  random  
 117 coordinates of the input vector  $x$ , scales them by  $d/K$ , and sets the remaining coordinates to zero (see  
 118 Def. C.1). Numerous other examples of unbiased compressors have been explored in the literature  
 119 (Beznosikov et al., 2020; Xu et al., 2021; Horváth et al., 2022; Szlendak et al., 2021). Using the  
 120 seminal ideas (Seide et al., 2014), we can construct a modified version of QSGD (Alistarh et al., 2017)  
 121 (special case of Shadowheart SGD from (Tyurin et al., 2024)), which we call Batch QSGD:

$$122 \quad x^{k+1} = x^k - \frac{\gamma}{nbm} \sum_{i=1}^n \sum_{k=1}^m \mathcal{C}_{ik} \left( \sum_{j=1}^b \nabla f(x^k; \xi_{ij}^k) \right), \quad (\text{Batch QSGD})$$

125 where worker  $i$  sends  $m$  compressed vectors  $\{\mathcal{C}_{ik}(\cdot)\}_{k \in [m]}$  to the server, which aggregates and  
 126 calculates  $x^{k+1}$ . With  $\text{Rand}K$  and proper parameters<sup>1</sup> (Tyurin et al., 2024), we can improve (2) to

$$128 \quad \mathcal{O} \left( h \left( 1 + \frac{\sigma^2}{n\varepsilon} \right) \frac{L\Delta}{\varepsilon} + \tau_w \left( \frac{d}{n} + 1 \right) \frac{L\Delta}{\varepsilon} + \sqrt{\frac{d\tau_w h\sigma^2}{n\varepsilon} \frac{L\Delta}{\varepsilon}} \right). \quad (3)$$

130 *Observation 2:* As in *Observation 1*, unlike (2), the time complexity (3) scales with the number  
 131 of workers  $n$ , which once again justifies the use of many workers in the optimization of (1). The  
 132 “statistical term”  $h\sigma^2 L\Delta/n\varepsilon^2$  and the “communication term”  $\tau_w dL\Delta/n\varepsilon$  improve linearly with  $n$ , while  
 133 the “coupling term”  $\sqrt{d\tau_w h\sigma^2/n\varepsilon} L\Delta/\varepsilon$  improves with the square root of  $n$ , which can reduce the effect  
 134 of  $d$  and  $\sigma^2/\varepsilon$  for reasonably large  $n$ .

136 A high-level explanation for why the dependence on  $d$  improves with  $n$  is that all workers use i.i.d.  
 137 and unbiased compressors  $\{\mathcal{C}_{ik}\}$ , which allow them to collaboratively explore more coordinates. This  
 138 effect is similar to Synchronized SGD, where the variance  $\mathbb{E}_\xi[\|\frac{1}{n} \sum_{i=1}^n \nabla f(x; \xi_i^k) - \nabla f(x)\|^2] \leq \frac{\sigma^2}{n}$   
 139 also improves with  $n$ . There are many other compressed methods that also improve with  $n$ , including  
 140 DIANA (Mishchenko et al., 2019), Accelerated DIANA (Li et al., 2020), MARINA (Gorbunov et al.,  
 141 2021), DASHA (Tyurin & Richtárik, 2023a), and FRECON (Zhao et al., 2021).

142 **3. Both communications can not be ignored.** Consider a more practical scenario, and our main  
 143 point of interest, where the communication time from the server to the workers is  $\tau_s > 0$ . In this case,  
 144 **Batch QSGD** requires

$$146 \quad \mathcal{O} \left( h \left( 1 + \frac{\sigma^2}{n\varepsilon} \right) \frac{L\Delta}{\varepsilon} + \tau_w \left( \frac{d}{n} + 1 \right) \frac{L\Delta}{\varepsilon} + \sqrt{\frac{d\tau_w h\sigma^2}{n\varepsilon} \frac{L\Delta}{\varepsilon}} + \tau_s d \frac{L\Delta}{\varepsilon} \right) \quad (4)$$

148 seconds because the server has to send  $x^k \in \mathbb{R}^d$  of size  $d$  to the workers in every iteration.

149 *Observation 3:* If  $\tau_s \simeq \tau_w$ , then (4) asymptotically equals  $\mathcal{O}(h(L\Delta/\varepsilon + \sigma^2 L\Delta/n\varepsilon^2) + \tau_s dL\Delta/\varepsilon)$ ,  
 150 reducing to (2), as in the method that does not compress at all! The “communication term”  $\tau_s dL\Delta/\varepsilon$   
 151 **does not** improve with  $n$ .

153 We now arrive at our **main research question**:

154 In the first case (**1. Communication is free**) and the second case (**2. Worker-to-**

155 **server communication can not be ignored**

156 ), it is possible to design a method  
 157 that scales the complexity with the number of workers  $n$ , while improving the  
 158 dependencies on  $d$  and  $\sigma^2/\varepsilon$ .

159 Can we design a similarly efficient method for the third case (**3. Both communica-**  
 160 **tions can not be ignored**) using unbiased compressors, where the communication

161 <sup>1</sup> $b = \Theta(\frac{t^*}{h})$ ,  $m = \Theta(\frac{t^*}{\tau_w})$ ,  $t^* = \Theta \left( \max \left\{ h, \tau_w, \frac{\tau_w d}{n}, \frac{h\sigma^2}{n\varepsilon}, \sqrt{\frac{d\tau_w h\sigma^2}{n\varepsilon}} \right\} \right)$ ,  $\gamma = \Theta(\frac{1}{L})$ ,  $K = 1$  in  $\text{Rand}K$

162 time from the server to the workers cannot be ignored, and where the dependence  
 163 on both  $d$  and  $\sigma^2/\varepsilon$  improves with  $n$ , either linearly or with the square root of  $n$ ?  
 164

165 At least, can this be achieved in the simplest homogeneous setting, where all  
 166 workers have access to the same function, a scenario that arguably represents the  
 167 simplest form of distributed optimization?

168 We know for certain that the answer is “No” in the *heterogeneous case*, due to the result of [Grun-  
 169 tkowska et al. \(2024\)](#), who proved that the iteration complexity does not improve with the number of  
 170 workers  $n$  under Assumptions [1.1](#) and [1.2](#). However, the homogeneous setting is “easier,” giving us  
 171 hope that the workers can exploit the fact that they all have access to the same distribution.

172 **1.2 CONTRIBUTIONS**

173 ♠ **Lower bound.** Surprisingly, the answer is “No” to our **main research question**, even in the  
 174 *homogeneous case*. We prove the following theorem.

175 **Theorem 1.6** (Informal Formulation of Theorems [4.2](#) and [F.1](#)). *Let Assumptions [1.1](#), [1.2](#), [1.3](#), and  
 176 [1.4](#) hold. It is infeasible to find an  $\varepsilon$ -stationary point faster than*

$$177 \Omega \left( \min \left\{ h \left( \frac{\sigma^2}{n\varepsilon} + 1 \right) \frac{L\Delta}{\varepsilon} + \tau_w \left( \frac{d}{n} + 1 \right) \frac{L\Delta}{\varepsilon} + \sqrt{\frac{d\tau_w h \sigma^2}{n\varepsilon} \frac{L\Delta}{\varepsilon}} + \tau_s d \frac{L\Delta}{\varepsilon}, h \frac{L\Delta}{\varepsilon} + h \frac{\sigma^2 L\Delta}{\varepsilon^2} \right\} \right) \quad (5)$$

180 *seconds (up to logarithmic factors), using unbiased compressors (Def. [1.5](#)) based on random sparsifi-  
 181 cation, for all  $L, \Delta, \varepsilon, n, \sigma^2, d, \tau_w, \tau_s, h > 0$  such that  $L\Delta \geq \tilde{\Theta}(\varepsilon)$  and dimension  $d \geq \tilde{\Theta}(L\Delta/\varepsilon)$ .*

185 Because of the min, the bound shows that it is possible to improve either the dependence on  
 186  $d$  or the dependence on  $\sigma^2/\varepsilon$  as the number of workers  $n$  increases, but not both simultane-  
 187 ously. The lower bound is matched either by **Batch QSGD** or by the non-distributed SGD method  
 188 (without any communication or cooperation). Moreover, if  $\tau_s \simeq \tau_w$ , the lower bound becomes  
 189  $\tilde{\Omega} \left( \min \left\{ h \left( \frac{\sigma^2}{n\varepsilon} + 1 \right) \frac{L\Delta}{\varepsilon} + \tau_s d \frac{L\Delta}{\varepsilon}, h \frac{L\Delta}{\varepsilon} + h \frac{\sigma^2 L\Delta}{\varepsilon^2} \right\} \right)$ , which can be matched by **Batch Synchro-  
 190 nized SGD** with the complexity [\(2\)](#) (without compression) or by the non-distributed SGD method. [In  
 191 other words, if  \$\tau\_s \simeq \tau\_w\$ , then using methods with random sparsification compression in the distributed  
 192 centralized setting offers no advantage. However, if  \$\tau\_s \lesssim \tau\_w\$ , the compression techniques can help on the  
 193 workers side in the regimes when  \$\tau\_w d/n + \sqrt{d\tau\_w h \sigma^2/n\varepsilon}\$  is larger than  \$\tau\_s d\$ , due to the former scaling with  \$n\$ .](#)

194 ♠ **New “worst-case” function.** To prove the lower bound, as we explain in Section [2.3](#), we needed a  
 195 new “worst-case” function construction (see Section [3](#)). We designed a new function  $F_{T,K,a}$  in [\(9\)](#),  
 196 which extends the ideas by [Carmon et al. \(2020\)](#). Proving its properties in Lemmas [3.1](#) and [3.2](#), as  
 197 well as designing the function itself, can be an important contribution on its own.

198 ♦ **Proof technique.** Using the new function, we develop a new proof technique and explain how the  
 199 problem of establishing the lower bound reduces to a statistical problem (see Section [4](#)), where we  
 200 need to prove a concentration bound for a special sum [\(13\)](#), which represents the minimal possible  
 201 random time required to find an  $\varepsilon$ -stationary point. Combining this result with the proven properties,  
 202 we obtain our main result [\(11\)](#).

203 ♦ **Improved analysis when  $\tau_w > 0$ .** To obtain the complete lower bound, we extended and improved  
 204 the result by [Tyurin et al. \(2024\)](#), which was limited for our scenario and required additional  
 205 modifications to finally obtain [\(5\)](#) (see Sections [F](#) and [5](#) for details).

208 **2 PRELIMINARIES**

210 For better comprehension of our new idea, we now present arguably one of the most important  
 211 worst-case functions by [Carmon et al. \(2020\)](#), which is widely used to prove lower bounds in  
 212 nonconvex optimization. It has been used by [Arjevani et al. \(2022; 2020a\)](#) to derive lower bounds  
 213 in the stochastic setting, by [Lu & De Sa \(2021\)](#) in the decentralized setting, by [Tyurin & Richtárik  
 214 \(2023b; 2024\)](#); [Tyurin et al. \(2024\)](#) in the asynchronous setting, by [Huang et al. \(2022\)](#) to show the  
 215 lower iteration bound for unidirectional compressed methods, and by [Li et al. \(2021\)](#) in problems  
 with a nonconvex-strongly-concave structure.

216 For any  $T \in \mathbb{N}$ , Carmon et al. (2020) define  $F_T : \mathbb{R}^T \rightarrow \mathbb{R}$  such that <sup>2</sup>

$$218 \quad 219 \quad 220 \quad 221 \quad 222 \quad 223 \quad 224 \quad 225 \quad 226 \quad 227 \quad 228 \quad 229 \quad 230 \quad 231 \quad 232 \quad 233 \quad 234 \quad 235 \quad 236 \quad 237 \quad 238 \quad 239 \quad 240 \quad 241 \quad 242 \quad 243 \quad 244 \quad 245 \quad 246 \quad 247 \quad 248 \quad 249 \quad 250 \quad 251 \quad 252 \quad 253 \quad 254 \quad 255 \quad 256 \quad 257 \quad 258 \quad 259 \quad 260 \quad 261 \quad 262 \quad 263 \quad 264 \quad 265 \quad 266 \quad 267 \quad 268 \quad 269$$

$$F_T(x) := \sum_{i=1}^T [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)], \quad (6)$$

where  $x_0 \equiv 1$ ,  $x_i$  is the  $i^{\text{th}}$  coordinate of  $x \in \mathbb{R}^T$ ,

$$\Psi(x) = \begin{cases} 0, & x \leq 1/2, \\ \exp\left(1 - \frac{1}{(2x-1)^2}\right), & x \geq 1/2, \end{cases} \quad \text{and} \quad \Phi(x) = \sqrt{e} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt. \quad (7)$$

Notice that this function has a “chain-like” structure. If a method starts from  $x^0 = 0$  and computes the gradient of  $F_T$ , then the gradient will have a non-zero value only in the first coordinate (use that  $\Psi(0) = \Psi'(0) = 0$ ). Thus, by computing a single gradient, any “reasonable” method can “discover” at most one coordinate. At the same time, if the method wants to find an  $\varepsilon$ -stationary point, it should eventually discover the  $T^{\text{th}}$  coordinate. These two facts imply that every “reasonable” method should compute the gradient of  $F_T$  at least  $T$  times. In the construction, Carmon et al. (2020) take  $T = \Theta\left(\frac{L\Delta}{\varepsilon}\right)$ . This construction is a “more technical” version of the celebrated quadratic optimization construction from (Nesterov, 2018), which has similar properties. Let us define  $\text{prog}(x) := \max\{i \geq 0 \mid x_i \neq 0\}$  ( $x_0 \equiv 1$ ), then the following lemma is a formalization of the described properties.

**Lemma 2.1** (Carmon et al. (2020)). *The function  $F_T$  satisfies:*

1. For all  $x \in \mathbb{R}^T$ ,  $\text{prog}(\nabla F_T(x)) \leq \text{prog}(x) + 1$ .
2. For all  $x \in \mathbb{R}^T$ , if  $\text{prog}(x) < T$ , then  $\|\nabla F_T(x)\| > 1$ .

Actually, in most proofs, the structure of (6) is not needed, and it is sufficient to work with Lemmas 2.1 and Lemma 2.2 from below, where the latter allows us to show that a scaled version of  $F_T$  satisfies Assumptions 1.1 and 1.2.

**Lemma 2.2** (Carmon et al. (2020)). *The function  $F_T$  satisfies:*

1.  $F_T(0) - \inf_{x \in \mathbb{R}^T} F_T(x) \leq \Delta^0 T$ , where  $\Delta^0 := 12$ .
2. The function  $F_T$  is  $\ell_1$ -smooth, where  $\ell_1 := 152$ .
3. For all  $x \in \mathbb{R}^T$ ,  $\|\nabla F_T(x)\|_\infty \leq \gamma_\infty$ , where  $\gamma_\infty := 23$ .

Hence, one of the main results by Carmon et al. (2020) was to show that it is infeasible to find an  $\varepsilon$ -stationary point without calculating  $\mathcal{O}\left(\frac{L\Delta}{\varepsilon}\right)$  gradients of a function satisfying Assumptions 1.1 and 1.2. In turn, the classical gradient descent (GD) method matches this lower bound.

## 2.1 FAMILY OF DISTRIBUTED METHODS

In our lower bound, we focus on the family of methods described by Protocol 1. This protocol takes an algorithm as input and runs the standard functions of the workers and the server: the workers compute stochastic gradients locally, [send compressed information](#), the server aggregates them asynchronously and in parallel, and sends compressed information back based on the local information. For now, we ignore the communication times from the workers to the server in Protocol 1.

For all  $i \in [n]$ , the algorithm can choose any point, based on the local information  $I_i$ , at which worker  $i$  will start computing a stochastic gradient. It can also choose any point  $s_i^k$ , based on the server’s local information  $I$ , along with the corresponding compressor  $C_i^k$ , which will be sent to worker  $i$ . This protocol captures the behavior of virtually any asynchronous optimization process with workers connected to a server. We work with *zero-respecting* algorithms, as defined below.

**Definition 2.3.** We say that an algorithm  $A$  that follows Protocol 1 is *zero-respecting* if it does not explore or assign non-zero values to any coordinate unless at least one of the available local vectors contains a non-zero value in that coordinate. The family of such algorithms we denote as  $\mathcal{A}_{\text{zr}}$ .

<sup>2</sup>similarly  $F_T(x) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^T [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)]$  because  $\Psi(-1) = 0$ .

---

270 **Protocol 1**

---

271 1: **Input:** Algorithm  $A \in \mathcal{A}_{\text{zr}}$

272 2: Init  $I = \emptyset$  (all available information) on the server

273 3: Init  $I_i = \emptyset$  (all available information) on worker  $i$  for all  $i \in [n]$

274 4: Run the following **three** loops in parallel. The first **two** on the workers. The **third** on the server.

275 5: **for**  $i = 1, \dots, n$  (in parallel on the workers) **do**

276 6:   **for**  $k = 0, 1, \dots$  **do**

277 7:     Algorithm  $A$  calculates a new point  $x$  based on local information  $I_i$ : (takes 0 seconds)

278 8:     any vector  $x \in \mathbb{R}^d$  such that  $\text{supp}(x) \in \cup_{y \in I_i} \text{supp}(y)$  ( $\text{supp}(v) := \{i \in [d] : v_i \neq 0\}$ )

279 9:     Calculate **one** stochastic gradient <sup>3</sup> $\nabla f(x; \xi)$ ,  $\xi \sim \mathcal{D}_\xi$  ( $\xi$  are i.i.d.) (takes  $h$  seconds)

280 10:     Add  $\nabla f(x; \xi)$  to  $I_i$  (takes 0 seconds)

281 11:   **end for**

282 12: **end for**

283 13: **for**  $i = 1, \dots, n$  (in parallel on the workers) **do**

284 14:   **for**  $k = 0, 1, \dots$  **do**

285 15:     Algorithm  $A$  calculates new points  $\{\bar{s}_i^k\}$  based on local information  $I_i$ : (takes 0 seconds) **any**

286 16:     vector  $\bar{s}_i^k \in \mathbb{R}^d$  such that  $\text{supp}(\bar{s}_i^k) \in \cup_{y \in I_i} \text{supp}(y)$

287 17:     Send  $\bar{C}_i^k(\bar{s}_i^k)$  to the server (takes  $\tau_w \times \bar{P}_i^k$  seconds, where  $\bar{P}_i^k$  is the number of coordinates retained by  $\bar{C}_i^k(\bar{s}_i^k)$ )

288 18:     Add to  $\bar{C}_i^k(\bar{s}_i^k)$  to  $I$

289 19:   **end for**

290 20: **end for**

291 21: **for**  $i = 1, \dots, n$  (in parallel on the server) **do**

292 22:   **for**  $k = 0, 1, \dots$  **do**

293 23:     Algorithm  $A$  calculates a new point  $s_i^k$  based on local information  $I$ : (takes 0 seconds)

294 24:     any vector  $s_i^k \in \mathbb{R}^d$  such that  $\text{supp}(s_i^k) \in \cup_{y \in I} \text{supp}(y)$

295 25:     Algorithm  $A$  compresses the point:  $\bar{C}_i^k(s_i^k) \quad \forall i \in [n]$  (takes 0 seconds)

296 26:     Send  $\bar{C}_i^k(s_i^k)$  to the worker  $i$  (takes  $\tau_s \times P_i^k$  seconds, where  $P_i^k$  is the number of coordinates retained by  $\bar{C}_i^k(s_i^k)$ )

297 27:     Add to  $\bar{C}_i^k(s_i^k)$  to  $I$  (takes 0 seconds)

298 28:   **end for**

299 29: **end for**

300 30: **end for**

301 31:   (a vector may be added to  $I$  or  $I_i$  at the same time as the algorithm calculates a new point; in this case, the protocol adds the vector first (with no delay since the operation takes 0 seconds))

---

304 This is the standard assumption (Carmon et al., 2020) that generalizes the family of methods working  
 305 with the span of vectors (Nesterov, 2018) and holds for the majority of methods, including GD, Adam  
 306 (Kingma & Ba, 2015), DORE (Liu et al., 2020), EF21-P (Grunkowska et al., 2023), MARINA-P, and  
 307 M3 (Grunkowska et al., 2024).

309 2.2 PREVIOUS LOWER BOUND IN THE HETEROGENEOUS SETTING

311 Let us return back to our main question. In order to show that it is impossible to scale with  $n$  in the  
 312 *heterogeneous setting*, Grunkowska et al. (2024) have proposed to use scaled versions of

313

$$314 G_j(x) := n \times \sum_{1 \leq i \leq T \text{ and } (i-1) \bmod n = j-1}^T [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)]$$

315

316 for all  $j \in [n]$ , worker  $i$  has access only a scaled version of  $G_i$  for all  $i \in [n]$ . The idea is that the  
 317 first block from (6) belongs to the first worker, the second block to the second worker,  $\dots$ , and the  
 318  $(n+1)^{\text{th}}$  block to the first worker again, and so on. Notice that  $F_T(x) = \frac{1}{n} \sum_{i=1}^n G_i(x)$ .

319 Notice one important property of this construction: only one worker at a time can discover the next  
 320 coordinate. In other words, if the server sends a new iterate to all workers, only one worker, after  
 321 computing the gradient, can make progress to the next coordinate.

323 <sup>3</sup>i) Multiple queries with the same random variable do not change the lower bound; see Remark E.1 in Section E; ii)  
 In the heterogeneous setup (Section 2.2), worker  $i$  computes  $\nabla f_i(x; \xi)$ , where  $f_i$  is its local function.

The next step in (Gruntkowska et al., 2024), in the proof of the lower bound theorem, was to analyze Protocol 1. They consider<sup>4</sup> RandK with  $K = 1$ . Then, since the compressor sends only one coordinate with probability  $p = 1/d$ , the probability that the server sends the last non-zero coordinate to the worker responsible for the current block of (6) that can progress to the new coordinate is also  $p$ . Thus, the number of consecutive coordinates that the server has to send to the workers is at least  $\sum_{j=1}^T \eta_j$ , where  $\eta_j$  is a geometric-like random variable with  $\mathbb{P}(\eta_j = m | \eta_{j-1}, \dots, \eta_1) \leq p(1-p)^{m-1}$  for all  $m \geq 1$ . Using classical tools from statistical analysis, one can show that  $\sum_{j=1}^T \eta_j \gtrsim T/p \simeq dL\Delta/\varepsilon$  with high probability. Thus, under Assumption 1.4, the communication time complexity cannot be better than  $\Omega(\tau_s dL\Delta/\varepsilon)$ , which does not improve with  $n$ .

### 2.3 FAILURE OF THE PREVIOUS CONSTRUCTION IN THE HOMOGENEOUS SETTING

However, in the *homogeneous* setting, if we want to reuse the idea, arguably the only option we have is to assign (scaled)  $F_T$  to all workers to ensure that they all have the same function. But in this case, the arguments from Section 2.2 no longer apply, because all workers can simultaneously progress to the next coordinate, since they have access to all blocks of (6).

Indeed, if the server sends i.i.d. RandK compressors with  $K = 1$ , then the number of consecutive coordinates that the server has to send before the workers receive the last non-zero coordinate is  $\sum_{j=1}^T \min_{i \in [n]} \eta_{ji}$ , where  $\mathbb{P}(\eta_{ji} = m | \{\eta_{kj}\}_{k < i}) \leq p(1-p)^{m-1}$  for all  $m, j \geq 1, i \in [n]$ . The  $\min_{i \in [n]}$  operation appears because it is sufficient to wait for the first “luckiest” worker. Analyzing this sum, we can only show that

$$\tau_s \sum_{j=1}^T \min_{i \in [n]} \eta_{ji} \gtrsim \tau_s \frac{d}{n} \frac{L\Delta}{\varepsilon}, \quad (8)$$

with high probability, which scales with  $n$  due to min.

There are two options: either  $\Omega(\tau_s dL\Delta/n\varepsilon)$  is tight and it is possible to find a method that matches it, or we need to find another way to improve the lower bound. To prove the latter, we arguably need a different fundamental construction from (6), which we propose in the next section.

## 3 A NEW “WORST-CASE” FUNCTION

In this section, we give a less technical description of our lower bound construction and the main theorem from Section D. Instead of (6), we propose to use another “worst-case” function. For any  $T, K \in \mathbb{N}$ , and  $e \geq a > 1$ , we define the function  $F_{T,K,a} : \mathbb{R}^T \rightarrow \mathbb{R}$  such that

$$F_{T,K,a}(x) = - \sum_{i=1}^T \Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i) + \sum_{i=1}^T \Gamma(x_i), \quad (9)$$

$$\Psi_a(x) = \begin{cases} 0, & x \leq 1/2, \\ \exp \left( \log a \cdot \left( 1 - \frac{1}{(2x-1)^2} \right) \right), & x > 1/2, \end{cases} \quad \Phi(x) = \sqrt{e} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt, \quad (10)$$

$$\Gamma(x) = \begin{cases} -xe^{1/x+1}, & x < 0, \\ 0, & x \geq 0, \end{cases}$$

and  $x_0 = \dots = x_{-K+1} \equiv 1$ . The main modification is that instead of the block  $-\Psi(x_{i-1})\Phi(x_i)$ , we use  $-\Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i)$  (ignore  $a$  for now). In the previous approach, it was sufficient for a worker to have  $x_{i-1} \neq 0$  to discover the next  $i^{\text{th}}$  coordinate. However, in our new construction, the worker needs  $x_{i-1} \neq 0, x_{i-2} \neq 0, \dots, x_{i-K+1} \neq 0$  for that. With this modification, it is not sufficient for the “luckiest” worker to get the non-zero  $i-1^{\text{th}}$  coordinate to discover the next coordinate: to progress to the  $i^{\text{th}}$  coordinate, the worker should also have non-zero  $i-2^{\text{th}}, \dots, i-K+1^{\text{th}}$  coordinates.

<sup>4</sup>In general, they presented a more general setting where the server can zero out coordinates with any probability, capturing not only RandK with  $K = 1$  and  $p = 1/d$ , but also RandK with  $K > 1$  and other compressors.

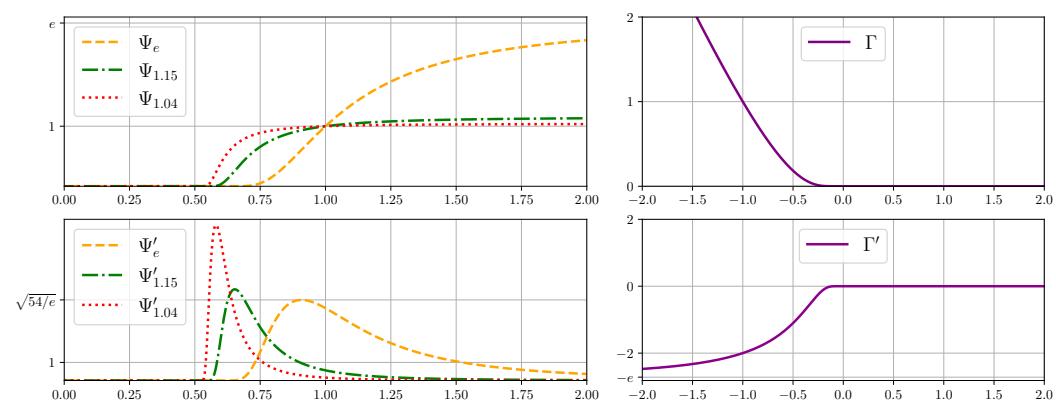


Figure 1: The functions  $\Psi_a(x)$  and  $\Gamma(x)$ , along with their derivatives  $\Psi'_a(x)$  and  $\Gamma'(x)$ . The plots of  $\Phi(x)$  and  $\Phi'(x)$  are shown in (Carmon et al., 2020).

Next, we remove the positive blocks  $\Psi(x_{i-1})\Phi(x_i)$ , which we believe was introduced to prevent the methods from ascending, exploring negative values of  $x_i$ , and finding a nearby stationary point “above.” Instead, we introduce  $\Gamma(x_i)$ , which serves the same purpose: if a method starts exploring negative values, this term prevents it from reaching a stationary point there. Let us define

$$\text{prog}^K(x) := \max\{i \geq 0 \mid x_i \neq 0, x_{i-1} \neq 0, \dots, x_{i-K+1} \neq 0\}.$$

Instead of Lemma 2.1, we prove the following lemma:

**Lemma 3.1** (Lemmas D.4 and D.5). *The function  $F_{T,K,a}$  satisfies:*

1. *For all  $x \in \mathbb{R}^T$ ,  $\text{supp}(\nabla F_{T,K,a}(x)) \in \{1, \dots, \text{prog}^K(x) + 1\} \cup \text{supp}(x)$ , where  $\text{supp}(v) := \{i \in [d] : v_i \neq 0\}$ .*
2. *For all  $x \in \mathbb{R}^T$ , if  $\text{prog}^K(x) < T$ , then  $\|\nabla F_{T,K,a}(x)\| > 1$ .*

The function  $F_{T,K,a}$  remains smooth. However, by multiplying with additional  $\Psi$  terms, we alter its geometry and make it more chaotic: the difference  $F_{T,K,a}(0) - \inf_{x \in \mathbb{R}^T} F_{T,K,a}(x)$ , the smoothness constant, and the maximum  $\ell_\infty$ -norm may increase. To mitigate this, we introduce the parameter  $a$  in (10) that allows us to control these properties. Notice that if  $a = e$ , then  $\Psi_a(x) = \Psi(x)$  for all  $x \in \mathbb{R}^T$ , where  $\Psi$  is defined in (7). Instead of Lemma 2.2, we prove

**Lemma 3.2** (Lemmas D.6, D.7, and D.8). *The function  $F_{T,K,a}$  satisfies:*

1.  $F_{T,K,a}(0) - \inf_{x \in \mathbb{R}^T} F_{T,K,a}(x) \leq \Delta^0(K, a) \cdot T$ , where  $\Delta^0(K, a) := \sqrt{2\pi e} \cdot a^K$ .
2. *The function  $F_{T,K,a}$  is  $\ell_1(K, a)$ -smooth, where  $\ell_1(K, a) := 12\sqrt{2\pi}e^{5/2} \cdot \frac{K^2 a^K}{\log a}$ .*
3. *For all  $x \in \mathbb{R}^T$ ,  $\|\nabla F_{T,K,a}(x)\|_\infty \leq \gamma_\infty(K, a)$ , where  $\gamma_\infty(K, a) := 6\sqrt{2\pi}e^{3/2} \cdot \frac{K a^K}{\sqrt{\log a}}$ .*

Taking  $K = 1$  and  $a = e$ , up to constant factors, Lemmas 3.1 and 3.2 reduce to Lemmas 2.1 and 2.2. The larger the value of  $K$ , the larger the bounds in Lemma 3.2, and this growth can be exponential if  $a = e$ . However, with a proper choice of  $1 < a \ll e$ , we can mitigate the increase caused by  $K$ .

## 4 LOWER BOUND WITH SERVER-TO-WORKER (S2W) COMMUNICATION

We now present informal and formal versions of our main result:

**Theorem 4.1** (Informal Formulation of Theorem 4.2). *Let Assumptions 1.1, 1.2, 1.3, and 1.4 hold. It is infeasible to find an  $\varepsilon$ -stationary point faster than*

$$\tilde{\Omega} \left( \min \left\{ \tau_s d \frac{L\Delta}{\varepsilon}, h \frac{L\Delta}{\varepsilon} + h \frac{\sigma^2 L\Delta}{\varepsilon^2} \right\} \right) \quad (11)$$

432 seconds (up to logarithmic factors), using unbiased compressors (Def. 1.5) based on random sparsification, for all  $L, \Delta, \varepsilon, n, \sigma^2, d, \tau_s, h > 0$  such that  $L\Delta \geq \tilde{\Theta}(\varepsilon)$  and dimension  $d \geq \tilde{\Theta}(L\Delta/\varepsilon)$ .  
 433  
 434

435 **Theorem 4.2.** Let  $L, \Delta, \varepsilon, n, \sigma^2, d, \tau_s, \tau_w, h > 0$  be any numbers such that  $\bar{c}_1 \varepsilon \log^4(n+1) < L\Delta$  and  
 436 dimension  $d \geq \bar{c}_3 \frac{L\Delta}{\log^3(n+1)\varepsilon}$ . Consider Protocol 1. For all  $i \in [n]$  and  $k \geq 0$ , compressor  $C_i^k$  selects  
 437 and transmits  $P_i^k$  uniformly random coordinates without replacement, scaled by any constants<sup>5</sup>, where  
 438  $P_i^k \in \{0, \dots, d\}$  may vary across each compressor<sup>6</sup>. Then, for any algorithm  $A \in \mathcal{A}_{\text{zr}}$  (Def. 2.3),  
 439 there exists a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $f$  is  $L$ -smooth, i.e.,  $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$   
 440 for all  $x, y \in \mathbb{R}^d$ , and  $f(0) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta$ , exists a stochastic gradient oracles that satisfies  
 441 Assumption 1.3, and  $\mathbb{E} \left[ \inf_{y \in S_t} \|\nabla f(y)\|^2 \right] > \varepsilon$  for all  
 442

$$443 \quad t \leq \bar{c}_2 \times \left( \frac{1}{\log^3(n+1)} \cdot \frac{L\Delta}{\varepsilon} \right) \min \left\{ \frac{1}{\log(n+1)} \cdot \tau_s d, \max \left\{ h, \frac{1}{\log^3(n+1)} \cdot \frac{h\sigma^2}{\varepsilon} \right\} \right\}, \quad (12)$$

444 where  $S_t$  is the set of all possible points that can be constructed by  $A$  up to time  $t$  based on  $I$  and  
 445  $\{I_i\}$ . The quantities  $\bar{c}_1$ ,  $\bar{c}_2$ , and  $\bar{c}_3$  are universal constants.  
 446  
 447

448 The formulation of Theorem 4.2 is standard in the literature. However, following Tyurin & Richtárik  
 449 (2023b), we present the lower bound in terms of *time complexities* rather than *iteration complexities*.  
 450 Then, following Huang et al. (2022); He et al. (2023); Tyurin et al. (2024), we consider a subfamily of  
 451 unbiased compressors based from Definition 1.5 on random sparsification to prove the lower bound;  
 452 this is standard practice for taking the “worst-case” compressors from the family (similarly to taking  
 453 the “worst-case” functions (Carmon et al., 2020; Nesterov, 2018)). Moreover, due to the uncertainty  
 454 principle (Safaryan et al., 2022), all unbiased compressors exhibit variance and communication cost  
 455 comparable to those of the RandK sparsifier in the worst case (up to constant factors).  
 456

457 The main observation in Theorems 4.1 and 4.2 is that it is not possible to scale both  $d$  and  $\sigma^2/\varepsilon$  by more  
 458 than  $\log^4(n+1)$  and  $\log^6(n+1)$ , respectively. Asymptotically, this scaling is significantly worse  
 459 than the linear  $n$  and square-root  $\sqrt{n}$  scalings discussed in Section 1.1. For instance, if  $n = 10,000$   
 460 and  $d$  is increased by a factor of 10, we have to increase  $n$  by a factor of  $10^3$  (two factors more) to  
 461 ensure that  $\tau_s d / \log^4(n+1)$  does not change.  
 462

463 In Section A, we present the intuition and the proof sketch of the result.  
 464

## 465 5 LOWER BOUND WITH BOTH W2S AND S2W COMMUNICATION

466 In the previous section, we provide the lower bound without taking into account the communication  
 467 cost  $\tau_w$ . Combining Theorem 4.2 with our new Theorem F.1, which extends the re-  
 468 sult by Tyurin et al. (2024) for our setup, we can obtain the complete lower bound (5) from  
 469 Theorem 1.6 with  $\tau_w > 0$  and  $\tau_s > 0$ . Notice that if  $\tau_s \simeq \tau_w$ , then the lower bound is  
 470  $\tilde{\Omega} \left( \min \left\{ h \left( \frac{\sigma^2}{n\varepsilon} + 1 \right) \frac{L\Delta}{\varepsilon} + \tau_s d \frac{L\Delta}{\varepsilon}, h \frac{L\Delta}{\varepsilon} + h \frac{\sigma^2 L\Delta}{\varepsilon^2} \right\} \right)$ . Up to logarithmic factors, under Assump-  
 471 tions 1.1, 1.2, 1.3, and 1.4, it is infeasible to improve both  $d$  and  $\sigma^2/\varepsilon$  as  $n$  increases.  
 472

### 473 5.1 ALGORITHMS ALMOST MATCHING THE LOWER BOUND

474 Due to the min, there are two regimes in which the lower bound (5) operates. If the second term  
 475 is smaller in (5), then the lower bound is  $\tilde{\Omega} \left( \frac{hL\Delta}{\varepsilon} + \frac{h\sigma^2 L\Delta}{\varepsilon^2} \right)$ , which is matched by the vanilla  
 476 SGD method run locally (without any communication or cooperation). Otherwise, if the first term  
 477 is smaller, then the lower bound is matched by **Batch QSGD**, which has the matching complexity  
 478 (4) (up to logarithmic factors). Moreover, in the latter case, if  $\tau_s \simeq \tau_w$ , the lower bound becomes  
 479  $\tilde{\Omega} \left( \min \left\{ h \left( \frac{\sigma^2}{n\varepsilon} + 1 \right) \frac{L\Delta}{\varepsilon} + \tau_s d \frac{L\Delta}{\varepsilon} \right\} \right)$ , which can be matched by **Batch Synchronized SGD** with the  
 480 complexity (2); thus, if  $\tau_s \simeq \tau_w$ , then unbiased sparsified compression is not needed at all, as it cannot  
 481 help due to the lower bound.  
 482

483 <sup>5</sup>To potentially preserve unbiasedness. For instance, RandK scales by  $d/K$ .  
 484

485 <sup>6</sup>For instance, the compressors can be RandK (see Def. C.1) with any  $K \in [d]$ , PermK (Szlenkak et al.,  
 486 2021), Identity compressor when  $P_i^k = d$ .

486 6 CONCLUSION  
487

488 We prove nearly tight lower bounds for centralized distributed optimization under the computation  
489 and communication Assumption 1.4. We show that *even in the homogeneous scenario*, it is not  
490 possible to scale both  $d$  and  $\sigma^2/\varepsilon$  by more than poly-logarithmic factors in  $n$ . Notice that the family of  
491 **unbiased compressors** contains the family of **biased compressors** (Beznosikov et al., 2020). Therefore,  
492 our lower bounds also apply to methods that use biased compressors, in the sense that there exists a  
493 “worst-case” compressor for which these methods cannot achieve a convergence rate faster than the lower  
494 bound in Theorem 1.6.

495 The lower bounds are tight only up to logarithmic factors. Thus, a possible challenging direction is  
496 to improve the powers of the logarithms, or even eliminate the logarithms entirely. The latter (if at  
497 all possible) may be very challenging and would likely require entirely different constructions and  
498 techniques. Another limitation is that the lower bounds are constructed using random sparsifiers.  
499 Due to the uncertainty principle (Safaryan et al., 2022), we conjecture that the bounds also hold  
500 for the entire family of unbiased compressors, but proving this would require more sophisticated  
501 constructions.

502 In practice, however, biased **and unbiased** compressors, including Top $K$  and Rank $K$  (Alistarh et al.,  
503 2018; Vogels et al., 2019), exhibit significantly better compression properties than those predicted by  
504 worst-case analysis (Beznosikov et al., 2020). When used on the server side in combination with EF  
505 or EF21-P (Gruntkowska et al., 2023; Tyurin et al., 2024), they may help mitigate the pessimistic term  
506  $\tau_s d^{L\Delta/\varepsilon}$ . Moreover, our pessimistic lower bound may potentially be broken under additional assumptions  
507 such as convexity or second-order smoothness.

508  
509 REFERENCES  
510

511 Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-  
512 efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing  
513 Systems (NIPS)*, pp. 1709–1720, 2017.

514 Dan Alistarh, Torsten Hoefer, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric  
515 Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information  
516 Processing Systems*, 31, 2018.

517 Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan.  
518 Second-order information in non-convex stochastic optimization: Power and limitations. In  
519 *Conference on Learning Theory*, pp. 242–299. PMLR, 2020a.

520 Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient  
521 descent with delayed updates. In *Algorithmic Learning Theory*, pp. 111–132. PMLR, 2020b.

522 Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth.  
523 Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50,  
524 2022.

525 Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression  
526 for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

527 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
528 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel  
529 Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler,  
530 Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Ben-  
531 jamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and  
532 Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell,  
533 M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33,  
534 pp. 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.  
535 cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).

536 Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary  
537 points i. *Mathematical Programming*, 184(1):71–120, 2020.

540 Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD.  
 541 *arXiv preprint arXiv:1905.10018*, 2019.  
 542

543 Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex op-  
 544 timization via stochastic path integrated differential estimator. In *NeurIPS Information Processing*  
 545 *Systems*, 2018.

546 Margalit R. Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (Local  
 547 SGD) and continuous perspective. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera  
 548 (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*,  
 549 volume 151 of *Proceedings of Machine Learning Research*, pp. 9050–9090. PMLR, 28–30 Mar  
 550 2022. URL <https://proceedings.mlr.press/v151/glasgow22a.html>.

551 Eduard Gorbunov, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-  
 552 convex distributed learning with compression. In *38th International Conference on Machine*  
 553 *Learning*, 2021.

554 Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and friends: Improved theoreti-  
 555 cal communication complexity for distributed optimization with bidirectional compression. In  
 556 *International Conference on Machine Learning*, pp. 11761–11807. PMLR, 2023.

557 Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional  
 558 communication complexity for nonconvex distributed optimization under function similarity. In  
 559 *Advances in Neural Information Processing Systems*, 2024.

560 Yutong He, Xinmeng Huang, and Kun Yuan. Unbiased compression saves communication in  
 561 distributed optimization: When and how much? *Advances in Neural Information Processing*  
 562 *Systems*, 36:47991–48020, 2023.

563 Samuel Horváth, Chen-Yu Ho, Ľudovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter  
 564 Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific*  
 565 *Machine Learning*, pp. 129–141. PMLR, 2022.

566 Junxian Huang, Feng Qian, Alexandre Gerber, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck.  
 567 A close examination of performance and power characteristics of 4g lte networks. In *Proceedings*  
 568 *of the 10th international conference on Mobile systems, applications, and services*, pp. 225–238,  
 569 2012.

570 Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algo-  
 571 rithms in distributed learning with communication compression. *Advances in Neural Information*  
 572 *Processing Systems*, 2022.

573 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International*  
 574 *Conference on Learning Representations*, 2015.

575 Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and  
 576 Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv*  
 577 *preprint arXiv:1610.05492*, 2016.

578 Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer,  
 579 2020.

580 Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-  
 581 strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34:  
 582 1792–1804, 2021.

583 Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient  
 584 descent in distributed and federated optimization. In *International Conference on Machine*  
 585 *Learning*, 2020.

586 Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for  
 587 efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics*,  
 588 pp. 133–143. PMLR, 2020.

594 Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International*  
 595 *Conference on Machine Learning*, pp. 7111–7123. PMLR, 2021.  
 596

597 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.  
 598 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.  
 599

600 Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning  
 601 with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.  
 602

603 Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear  
 604 programming. Technical report, 1985.  
 605

606 Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan  
 607 Zhang, Denis Rybkin, Zhengxuan Yang, Zhuoqing Morley Mao, et al. A variegated look at 5g in  
 608 the wild: performance, power, and qoe implications. In *Proceedings of the 2021 ACM SIGCOMM*  
 609 *2021 Conference*, pp. 610–625, 2021.  
 610

611 Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method  
 612 efficiency in optimization. 1983.  
 613

614 Yurii Nesterov. A method of solving a convex programming problem with convergence rate  $o(1/k^{**})$ .  
 615 *Doklady Akademii Nauk SSSR*, 269(3):543, 1983.  
 616

617 Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.  
 618

619 Kumar Kshitij Patel, Lingxiao Wang, Blake Woodworth, Brian Bullins, and Nati Srebro. Towards  
 620 optimal communication complexity in distributed non-convex optimization. In *Proceedings*  
 621 *of the 36th International Conference on Neural Information Processing Systems*, 2022. ISBN  
 622 9781713871088.  
 623

624 Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty principle for communication compres-  
 625 sion in distributed and federated learning and the search for an optimal compressor. *Information*  
 626 *and Inference: A Journal of the IMA*, 11(2):557–580, 2022.  
 627

628 Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its  
 629 application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference*  
 630 *of the International Speech Communication Association*, 2014.  
 631

632 Rafal Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster  
 633 distributed nonconvex optimization. In *International Conference on Learning Representations*,  
 634 2021.  
 635

636 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
 637 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
 638 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.  
 639

640 Alexander Tyurin and Peter Richtárik. DASHA: Distributed nonconvex optimization with communi-  
 641 cation compression, optimal oracle complexity, and no client synchronization. *11th International*  
 642 *Conference on Learning Representations (ICLR)*, 2023a.  
 643

644 Alexander Tyurin and Peter Richtárik. Optimal time complexities of parallel stochastic optimization  
 645 methods under a fixed computation model. *Advances in Neural Information Processing Systems*,  
 646 2023b.  
 647

648 Alexander Tyurin and Peter Richtárik. On the optimal time complexities in decentralized stochastic  
 649 asynchronous optimization. *Advances in Neural Information Processing Systems*, 37, 2024.  
 650

651 Alexander Tyurin, Marta Pozzi, Ivan Ilin, and Peter Richtárik. Shadowheart SGD: Distributed  
 652 asynchronous SGD with optimal time complexity under arbitrary computation and communication  
 653 heterogeneity. *Advances in Neural Information Processing Systems*, 37, 2024.  
 654

655 Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient  
 656 compression for distributed optimization. In *Neural Information Processing Systems*, 2019.  
 657

648 Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan McMahan,  
649 Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *International*  
650 *Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020.

651

652 Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives.  
653 *Advances in neural information processing systems*, 29, 2016.

654

655 Blake E Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph  
656 oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in Neural*  
657 *Information Processing Systems*, 31, 2018.

658

659 Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of  
660 distributed stochastic convex optimization with intermittent communication. In *Conference on*  
*Learning Theory*, pp. 4386–4437. PMLR, 2021.

661

662 Hang Xu, Chen-Yu Ho, Ahmed M Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos  
663 Karatsenidis, Marco Canini, and Panos Kalnis. Grace: A compressed communication framework  
664 for distributed machine learning. In *2021 IEEE 41st International Conference on Distributed*  
*Computing Systems (ICDCS)*, pp. 561–572. IEEE, 2021.

665

666 Haoyu Zhao, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. Faster rates for compressed  
667 federated learning with client-variance reduction. *arXiv preprint arXiv:2112.13097*, 2021.

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702	CONTENTS	
703		
704		
705	<b>1 Introduction</b>	<b>1</b>
706	1.1 Related work . . . . .	2
707	1.2 Contributions . . . . .	4
708		
709		
710	<b>2 Preliminaries</b>	<b>4</b>
711	2.1 Family of distributed methods . . . . .	5
712	2.2 Previous Lower Bound in the Heterogeneous Setting . . . . .	6
713	2.3 Failure of the previous construction in the homogeneous setting . . . . .	7
714		
715		
716	<b>3 A New “Worst-Case” Function</b>	<b>7</b>
717		
718	<b>4 Lower Bound with Server-to-Worker (S2W) Communication</b>	<b>8</b>
719		
720	<b>5 Lower Bound with Both W2S and S2W Communication</b>	<b>9</b>
721		
722	5.1 Algorithms almost matching the lower bound . . . . .	9
723		
724	<b>6 Conclusion</b>	<b>10</b>
725		
726	<b>A Proof Sketch</b>	<b>15</b>
727		
728	<b>B Additional Related Work</b>	<b>16</b>
729		
730	<b>C Auxiliary Facts and Notations</b>	<b>16</b>
731		
732	C.1 Notations . . . . .	16
733		
734	<b>D Lower Bound</b>	<b>16</b>
735		
736	D.1 New Construction . . . . .	16
737		
738	D.2 Auxiliary Lemmas . . . . .	17
739	D.3 Proof of lemmas . . . . .	18
740		
741	<b>E Proof of Theorem 4.2</b>	<b>21</b>
742		
743	E.1 Main Concentration Lemma . . . . .	27
744		
745	<b>F Main Theorem with Worker-to-Server Communication</b>	<b>31</b>
746		
747	F.1 Main Concentration Lemma . . . . .	34
748		
749		
750		
751		
752		
753		
754		
755		

756 A PROOF SKETCH  
757

758 We illustrate the main idea behind the proof and how the new “worst-case” function helps to almost  
759 eliminate the scaling with  $n$ . Consider the first  $K$  coordinates of  $F_{T,K,a}$  (which is scaled in the proof  
760 to satisfy Assumptions 1.1 and 1.2). Recall that, due to Lemma 3.1, the only way to discover the  
761  $K + 1^{\text{th}}$  coordinate in any worker is to ensure that all of the first  $K$  coordinates are non-zero.

762 **Reduction to a statistical problem.** There are only two options by which a worker may discover a  
763 new non-zero coordinate: through local stochastic computations or through communication from the  
764 server. In the first option, a worker computes a stochastic gradient, which takes  $h$  seconds. However,  
765 due to the construction of stochastic gradients (Arjevani et al., 2022), even if the computation is  
766 completed, the worker will not make progress or discover a new non-zero coordinate, as it will be  
767 zeroed out with probability  $p_\sigma = \Theta(\varepsilon \cdot \gamma_\infty^2(K,a)/\sigma^2)$ . In the second option, due to the condition of  
768 Theorem 4.2, a worker receives a stream of uniformly sampled coordinates  $\nu_1, \nu_2, \dots$  (workers get  
769 different streams), and the worker can discover a new non-zero coordinate only if random variable  
770  $\nu_i \in [K]$ , which satisfies  $\mathbb{P}(\nu_i \in [K] | \nu_1, \dots, \nu_{i-1}) \leq K/T - i+1 \leq p_K := 2K/T$  for all  $i \leq T/2$ .

771 Next, we define two sets of random variables: (i) let  $\eta_{1,i,k}$  denote the number of stochastic gradient  
772 computations until the first moment when a coordinate is not zeroed out in the stochastic gradient  
773 oracle (see (26)), after the moment when the  $(k-1)^{\text{th}}$  coordinate is no longer zeroed out in worker  
774  $i$ ; (ii) let  $\mu_{1,i,k}$  be the number of received coordinates until the moment when the last received  
775 coordinate belongs to  $[K]$ , after the  $(k-1)^{\text{th}}$  time this has happened. In other words,  $\eta_{1,i,1}$  is the  
776 number of stochastic gradient computations until the moment when the algorithm receives a “lucky”  
777 stochastic gradient where the last coordinate is not zeroed out. The random variable  $\eta_{1,i,2}$  is the  
778 number of computations until it happens for the second time, and so on. Similarly,  $\mu_{1,i,1}$  is the  
779 position of the first coordinate from the stream sent by the server to worker  $i$  that belongs to  $[K]$ . The  
780 random variable  $\mu_{1,i,2}$  refers to the second time this occurs, and so on. By definition, the sequences  
781  $\{\eta_{1,i,k}\}$  and  $\{\mu_{1,i,k}\}$  follow *approximately* geometric-like distributions with parameters  $p_\sigma$  and  $p_K$ ,  
782 respectively.

783 To discover all of the first  $K$  coordinates, either the first or the second process must uncover at  
784 least  $K/2$  coordinates. If worker  $i$  has discovered fewer than  $K/2$  coordinates through stochastic  
785 gradient computations, and fewer than  $K/2$  coordinates through receiving them from the server,  
786 then it will not be able to cover all  $K$  coordinates. Thus, the algorithm should wait at least  
787  $\min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{K/2} \eta_{1,i,k}, \tau_s \sum_{k=1}^{K/2} \mu_{1,i,k} \right\} \right\}$  seconds until the moment when it can potentially  
788 discover the  $K + 1^{\text{th}}$  coordinate, where the outer minimum  $\min_{i \in [n]}$  appears because it is sufficient  
789 for the algorithm to wait for the first “luckiest” worker. Repeating the same arguments  $B := \lfloor T/K \rfloor$   
790 times, the algorithms requires at least

$$791 \quad t_B := \sum_{b=1}^B \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{K/2} \eta_{b,i,k}, \tau_s \sum_{k=1}^{K/2} \mu_{b,i,k} \right\} \right\} \quad (13)$$

794 seconds to discover the  $T^{\text{th}}$  coordinate and potentially find an  $\varepsilon$ -stationary point, where the sequences  
795  $\{\eta_{b,i,k}\}$  and  $\{\mu_{b,i,k}\}$  follow *approximately* geometric-like distributions with  $p_\sigma$  and  $p_K$ , respectively.

796 **Analysis of the concentration.** Hence, we have reduced the lower bound to the analysis of the  
797 sum  $t_B$ . Recall (8), where the lower bound improves with  $n$  due to  $\min_{i \in [n]}$ . In (13), we also  
798 get  $\min_{i \in [n]}$ . However, and this is the main reason for the new construction, there are two sums  
799  $\sum_{k=1}^{K/2}$ , which allows us to mitigate the influence of the  $\min_{i \in [n]}$ . In particular, we can show that  
800  $t_B \gtrsim \frac{BK}{n^{1/K}} \min \{h/p_\sigma, \tau_s/p_K\}$  with high probability. Notice that the first fraction improves with  $n^{\frac{1}{K}}$   
801 instead of  $n$  due to the sums; thus, the larger  $K$ , the smaller the influence of  $n$ .

803 **Putting it all together.** However, we cannot take  $K$  too large due to Lemma 3.2. Substituting the  
804 choice of  $T, p_\sigma$ , and  $p_K$  (defined in the proof of Theorem 4.2 to ensure that Assumptions 1.1, 1.2,  
805 and 1.3 are satisfied and the scaled version of  $F_{T,K,a}$  has the squared norm larger than  $\varepsilon$  while the  
806  $T^{\text{th}}$  is not discovered), we can show that

$$807 \quad t_B \gtrsim \frac{L\Delta}{n^{1/K} \cdot \Delta^0(K,a) \cdot \ell_1(K,a) \cdot \varepsilon} \min \left\{ \max \left\{ h, \frac{h\sigma^2}{\varepsilon \cdot \gamma_\infty^2(K,a)} \right\}, \frac{\tau_s d}{K} \right\},$$

808 with high probability, where  $\Delta^0(K,a)$ ,  $\ell_1(K,a)$ , and  $\gamma_\infty(K,a)$  are defined in Lemma 3.2. The final  
809 step is to choose  $K = \Theta(\log n)$  and  $a = 1 + 1/K$  to obtain the result of Theorem 4.2.

810 **B Additional Related Work**  
811

812 While we focus on lower bounds in the context of stochastic optimization and compressed vectors in  
 813 nonconvex settings, there is much related work in other domains and setups. The seminal works on  
 814 lower bounds were done by Nemirovskij & Yudin (1983); Nesterov (2018), where Nesterov (2018) showed  
 815 that the accelerated gradient descent (Nesterov, 1983) is optimal in the convex setting using a quadratic  
 816 “worst-case” function. In the nonconvex setting, Carmon et al. (2020) provided an alternative function,  
 817 described in the main part of the paper. For convex problems, Woodworth et al. (2018) introduced the  
 818 graph oracle, a generalization of the classical gradient oracle (Nemirovskij & Yudin, 1983; Nesterov, 2018),  
 819 and established lower bounds for a broad class of parallel optimization methods. Arjevani et al. (2020b)  
 820 further analyzed the delayed gradient descent method, which corresponds to Asynchronous SGD with  
 821 constant iteration delays. Tyurin & Richtárik (2023b; 2024); Tyurin et al. (2024) proved lower bounds for  
 822 methods in asynchronous settings. Fang et al. (2018); Patel et al. (2022) studied a different setting from  
 823 Assumption 1.3, where they assumed the mean-squared smoothness property to enable the analysis of  
 824 methods with variance reduction techniques (Fang et al., 2018; Cutkosky & Orabona, 2019). Woodworth  
 825 & Srebro (2016) considered the finite-sum setting in the convex setting. Woodworth et al. (2020; 2021)  
 826 proved that the min-max optimal algorithm for optimizing smooth convex objectives in the intermittent  
 827 communication setting is the best of accelerated local and minibatch SGD, which leads to a similar  
 828 conclusion to ours; however, their results are related to, but not directly comparable with ours, since we  
 829 analyze the limited scalability of improving both stochastic noise and communication complexity through  
 830 compressors. Glasgow et al. (2022) provided sharp lower bounds for local SGD approaches in terms of  
 831 iteration complexity. Huang et al. (2022); He et al. (2023); Gruntkowska et al. (2024) provided lower bounds  
 832 for compression techniques, but in the heterogeneous setting.

833 **C AUXILIARY FACTS AND NOTATIONS**  
834

835 **Definition C.1** (RandK). Assume that  $S$  is a random subset of  $[d]$  such that  $|S| = K$  for some  
 836  $K \in [d]$ . A stochastic mapping  $\mathcal{C} : \mathbb{R}^d \times \mathbb{S}_v \rightarrow \mathbb{R}^d$  is called RandK if

$$837 \quad 838 \quad 839 \quad \mathcal{C}(x; S) = \frac{d}{K} \sum_{j \in S} x_j e_j,$$

840 where  $\{e_i\}_{i=1}^d$  denotes the standard unit basis. The set  $S$  can be produced with a uniform sampling  
 841 of  $[d]$  without replacement.

843 **C.1 NOTATIONS**  
844

845  $\mathbb{N} := \{1, 2, \dots\}$ ;  $\|x\|$  is the output of the standard Euclidean norm for all  $x \in \mathbb{R}^d$ ;  $\langle x, y \rangle =$   
 846  $\sum_{i=1}^d x_i y_i$  is the standard dot product;  $\|A\|$  is the standard spectral/operator norm for all  $A \in$   
 847  $\mathbb{R}^{d \times d}$ ;  $g = \mathcal{O}(f) : \exists C > 0$  such that  $g(z) \leq C \times f(z)$  for all  $z \in \mathcal{Z}$ ;  $g = \Omega(f) : \exists$   
 848  $C > 0$  such that  $g(z) \geq C \times f(z)$  for all  $z \in \mathcal{Z}$ ;  $g = \Theta(f) : g = \mathcal{O}(f)$  and  $g = \Omega(f)$ ;  
 849  $g = \tilde{\mathcal{O}}(f)$ ,  $g = \tilde{\Omega}(f)$ ,  $g = \tilde{\Theta}(f)$  : the same as  $g = \mathcal{O}(f)$ ,  $g = \Omega(f)$ ,  $g = \Theta(f)$ , respectively, but  
 850 up to logarithmic factors;  $g \simeq h : g$  and  $h$  are equal up to universal positive constants;  $g \gtrsim h : g$   
 851 greater or equal to  $h$  up to universal positive constants;  $\mathcal{C}$  is an unbiased compressor (Definition 1.5);  
 852  $\text{supp}(v) = \{i \in [d] : v_i \neq 0\}$ ;  $h$  : maximum time (in seconds) for any worker to compute one stochastic  
 853 gradient;  $\tau_s$  : communication time per coordinate from the server to any worker;  $\tau_w$  : communication time  
 854 per coordinate from any worker to the server;

855 **D LOWER BOUND**  
856857 **D.1 NEW CONSTRUCTION**  
858

860 For any  $T, K \in \mathbb{N}$ , and  $e \geq a > 1$  we define the function  $F_{T, K, a} : \mathbb{R}^T \rightarrow \mathbb{R}$  such that  
 861

$$862 \quad 863 \quad F_{T, K, a}(x) = - \sum_{i=1}^T \Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i) + \sum_{i=1}^T \Gamma(x_i), \quad (14)$$

864 where  $x_i$  is the  $i^{\text{th}}$  coordinate of a vector  $x \in \mathbb{R}^T$  and  
 865

$$866 \quad \Psi_a(x) = \begin{cases} 0, & x \leq 1/2, \\ 867 \quad \exp\left(\log a \cdot \left(1 - \frac{1}{(2x-1)^2}\right)\right), & x > 1/2, \end{cases} \quad \Phi(x) = \sqrt{e} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt,$$

869 and  
 870

$$871 \quad \Gamma(x) = \begin{cases} -xe^{1/x+1}, & x < 0, \\ 872 \quad 0, & x \geq 0. \end{cases}$$

873 We assume that  $x_0 = \dots = x_{-K+1} \equiv 1$ . Importantly, throughout the lower bound analysis, we  
 874 assume that  $e \geq a > 1$ , even if this assumption is not explicitly stated in all theorems.  
 875

876 We additionally define  
 877

$$878 \quad \text{prog}^K(x) := \max\{i \geq 0 \mid x_i \neq 0, x_{i-1} \neq 0, \dots, x_{i-K+1} \neq 0\} \\ 879 \quad (x_0 = \dots = x_{-K+1} \equiv 1),$$

880 which extends  $\text{prog}(x) \equiv \text{prog}^1(x) := \max\{i \geq 0 \mid x_i \neq 0\}$  ( $x_0 \equiv 1$ ).  
 881

## 882 D.2 AUXILIARY LEMMAS

884 In this section, we list useful properties of the functions  $\Phi$ ,  $\Gamma$ ,  $\Psi_a$ , and  $F_{T,K,a}$ . We prove them in  
 885 Section D.3.  
 886

887 **Lemma D.1** (Carmon et al. (2020)). *Function  $\Phi$  is twice differentiable and satisfies*

$$888 \quad 0 \leq \Phi(x) \leq \sqrt{2\pi e}, \quad 0 \leq \Phi'(x) \leq \sqrt{e}, \text{ and } |\Phi''(x)| \leq 27$$

890 for all  $x \in \mathbb{R}$ . Moreover,  $\Phi'(x) > 1$  for all  $-1 < x < 1$ .  
 891

892 **Lemma D.2.** *Function  $\Gamma$  is twice differentiable and satisfies*

$$893 \quad 0 \leq \Gamma(x), \quad -e < \Gamma'(x) \leq 0, \text{ and } 0 \leq \Gamma''(x) \leq 27e^{-2}$$

895 for all  $x \in \mathbb{R}$ . Moreover,  $\Gamma'(x) \leq -2$  for all  $x \leq -1$ .  
 896

897 **Lemma D.3.** *Function  $\Psi_a$  is twice differentiable and satisfies*

$$898 \quad 0 \leq \Psi_a(x) < a, \quad 0 \leq \Psi'_a(x) \leq \frac{2e}{\sqrt{\log a}}, \text{ and } |\Psi''_a(x)| \leq \frac{56e}{\log a}$$

901 for all  $x \in \mathbb{R}$  and  $1 < a \leq e$ . Moreover,  $\Psi_a(x) \geq 1$  for all  $x \geq 1$  and  $1 < a \leq e$ .  
 902

903 **Lemma D.4.** *For all  $x \in \mathbb{R}^T$ ,  $\text{supp}(\nabla F_{T,K,a}(x)) \in \{1, \dots, \text{prog}^K(x) + 1\} \cup \text{supp}(x)$ , where  
 904  $\text{supp}(v) := \{i \in [d] : v_i \neq 0\}$ .*

905 **Lemma D.5.** *For all  $x \in \mathbb{R}^T$ , if  $\text{prog}^K(x) < T$ , then  $\|\nabla F_{T,K,a}(x)\| > 1$ .*  
 906

907 **Lemma D.6.** *Function  $F_{T,K,a}$  satisfies*  
 908

$$909 \quad F_{T,K,a}(0) - \inf_{x \in \mathbb{R}^T} F_{T,K,a}(x) \leq \Delta^0(K, a) \cdot T,$$

911 where  $\Delta^0(K, a) := \sqrt{2\pi e} \cdot a^K$ .  
 912

913 **Lemma D.7.** *For all  $x \in \mathbb{R}^T$ ,  $\|\nabla F_{T,K,a}(x)\|_\infty \leq \gamma_\infty(K, a)$ , where  $\gamma_\infty(K, a) := 6\sqrt{2\pi e}^{3/2} \cdot \frac{Ka^K}{\sqrt{\log a}}$ .*  
 914

915 **Lemma D.8.** *The function  $F_{T,K,a}$  is  $\ell_1(K, a)$ -smooth, i.e.,  $\|\nabla^2 F_{T,K,a}(x)\| \leq \ell_1(K, a)$  for all  
 916  $x \in \mathbb{R}^T$ , where  $\ell_1(K, a) := 12\sqrt{2\pi e}^{5/2} \cdot \frac{K^2 a^K}{\log a}$ .*  
 917

918 D.3 PROOF OF LEMMAS  
919920 **Lemma D.2.** *Function  $\Gamma$  is twice differentiable and satisfies*

921 
$$0 \leq \Gamma(x), \quad -e < \Gamma'(x) \leq 0, \text{ and } 0 \leq \Gamma''(x) \leq 27e^{-2}$$
  
922

923 for all  $x \in \mathbb{R}$ . Moreover,  $\Gamma'(x) \leq -2$  for all  $x \leq -1$ .924  
925 *Proof.* The first fact is due to  $\lim_{\Delta \rightarrow 0} \frac{\Gamma(\Delta)}{\Delta} = 0$ ,  $\Gamma'(0) = 0$ , and  $\lim_{\Delta \rightarrow 0} \frac{\Gamma'(\Delta)}{\Delta} = 0$ .  $\Gamma$  is clearly non-negative. Next, for all  $x \leq 0$ ,

926  
927 
$$\Gamma'(x) = -e^{1/x+1} + \frac{e^{1/x+1}}{x}$$
  
928

929 and  
930

931  
932 
$$\Gamma''(x) = -\frac{e^{1/x+1}}{x^3}.$$
  
933

934 Thus,  $\Gamma'$  is strongly increasing for all  $x \leq 0$ , and  $\lim_{x \rightarrow -\infty} \Gamma'(x) = -e < \Gamma'(x) \leq 0$ . Next,  $\Gamma''(x) \geq 0$   
935 for all  $x \leq 0$ , and  $\max_{x \leq 0} \Gamma''(x) = 27e^{-2}$  for all  $x \leq 0$ .  $\square$   
936937 **Lemma D.3.** *Function  $\Psi_a$  is twice differentiable and satisfies*

938  
939 
$$0 \leq \Psi_a(x) < a, \quad 0 \leq \Psi'_a(x) \leq \frac{2e}{\sqrt{\log a}}, \text{ and } |\Psi''_a(x)| \leq \frac{56e}{\log a}$$
  
940

941 for all  $x \in \mathbb{R}$  and  $1 < a \leq e$ . Moreover,  $\Psi_a(x) \geq 1$  for all  $x \geq 1$  and  $1 < a \leq e$ .942  
943 *Proof.* The differentiability at  $x = \frac{1}{2}$  follows from  $\lim_{\Delta \rightarrow 0} \frac{\Psi_a(\frac{1}{2} + \Delta)}{\Delta} = 0$  for all  $a > 1$ . For all  $x \leq \frac{1}{2}$ ,  
944  $\Psi'_a(x) = 0$ . For all  $x > \frac{1}{2}$ , we get

945  
946 
$$\begin{aligned} 0 \leq \Psi'_a(x) &= \frac{4 \log a}{(2x-1)^3} \exp \left( \log a \left( 1 - \frac{1}{(2x-1)^2} \right) \right) \\ &= \frac{4a}{\sqrt{\log a}} \times \frac{\log^{3/2} a}{(2x-1)^3} \exp \left( -\frac{\log a}{(2x-1)^2} \right). \end{aligned}$$
  
947

948 Taking  $t = \frac{\log^{1/2} a}{(2x-1)} > 0$  and using  $t^3 e^{-t^2} \leq \frac{1}{2}$ , we get

949  
950 
$$\Psi'_a(x) \leq \frac{4a}{\sqrt{\log a}} \times \frac{1}{2} \leq \frac{2e}{\sqrt{\log a}}$$
  
951

952 since  $a \leq e$ .953  
954 Clearly,  $\Psi_a(x) \geq 0$  for all  $x \in \mathbb{R}$ , and  $\Psi_a$  is non-decreasing. Moreover it is strongly monotonic for  
955 all  $x > \frac{1}{2}$ . Thus  $\Psi_a(x) < \lim_{x \rightarrow \infty} \Psi_a(x) = a$  for all  $x \in \mathbb{R}$ .956  
957 The twice differentiability at  $x = \frac{1}{2}$  follows from  $\lim_{\Delta \rightarrow 0} \frac{\Psi'_a(\frac{1}{2} + \Delta)}{\Delta} = 0$  for all  $a > 1$ . For all  $x \leq \frac{1}{2}$ ,  
958  $\Psi''_a(x) = 0$ . For all  $x > \frac{1}{2}$ , taking the second derivative and using simple algebra, we get

959  
960 
$$\begin{aligned} |\Psi''_a(x)| &= \left| -\frac{8 \log a \times (3(2x-1)^2 - 2 \log a)}{(2x-1)^6} \exp \left( \log a \left( 1 - \frac{1}{(2x-1)^2} \right) \right) \right| \\ &= \left| \frac{8a \log a \times (3(2x-1)^2 - 2 \log a)}{(2x-1)^6} \exp \left( -\frac{\log a}{(2x-1)^2} \right) \right| \\ &\leq \left| \frac{24a \log a}{(2x-1)^4} \exp \left( -\frac{\log a}{(2x-1)^2} \right) \right| + \left| \frac{16a \log^2 a}{(2x-1)^6} \exp \left( -\frac{\log a}{(2x-1)^2} \right) \right| \\ &= \frac{24a}{\log a} \times \frac{\log^2 a}{(2x-1)^4} \exp \left( -\frac{\log a}{(2x-1)^2} \right) + \frac{16a}{\log a} \times \frac{\log^3 a}{(2x-1)^6} \exp \left( -\frac{\log a}{(2x-1)^2} \right). \end{aligned}$$
  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

972 Taking  $t = \frac{\log a}{(2x-1)^2} > 0$  and using  $t^2 e^{-t} \leq 1$  and  $t^3 e^{-t} \leq 2$ ,  
 973

$$974 \quad |\Psi_a''(x)| \leq \frac{24a}{\log a} \times 1 + \frac{16a}{\log a} \times 2 \leq \frac{56e}{\log a}$$

975 since  $a \leq e$ . □  
 976

977 **Lemma D.4.** For all  $x \in \mathbb{R}^T$ ,  $\text{supp}(\nabla F_{T,K,a}(x)) \in \{1, \dots, \text{prog}^K(x) + 1\} \cup \text{supp}(x)$ , where  
 978  $\text{supp}(v) := \{i \in [d] : v_i \neq 0\}$ .  
 979

980 *Proof.* Let  $j = \text{prog}^K(x)$  and  $p = \text{prog}^1(x)$ , then  
 981

$$982 \quad F_{T,K,a}(x) = - \sum_{i=1}^{j+1} \Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i) \\ 983 \\ 984 \quad - \sum_{i=j+2}^T \Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i) \\ 985 \\ 986 \quad + \sum_{i=1}^p \Gamma(x_i) + \sum_{i=p+1}^T \Gamma(x_i).$$

987 Since  $j = \text{prog}^K(x)$ , for all  $i \geq j + 2$ , at least one of the values  $x_{i-K}, \dots, x_{i-2}, x_{i-1}$  is zero.  
 988 Noting that  $\Psi_a(0) = \Psi_a'(0) = 0$ , the gradient of the second sum is zero. The first sum depends only  
 989 on the first  $j + 1$  coordinates; thus, the gradient of the first sum is non-zero in at most the  $(j + 1)^{\text{th}}$   
 990 coordinate.  
 991

992 Since  $p = \text{prog}^1(x)$ , the gradient of the last sum is zero because  $\Gamma'(0) = 0$ . Moreover, if  $x_i = 0$ , then  
 993  $\Gamma'(x_i) = 0$ ; thus,  $\nabla(\sum_{i=1}^p \Gamma(x_i)) \in \text{supp}(x)$ . □  
 994

995 **Lemma D.5.** For all  $x \in \mathbb{R}^T$ , if  $\text{prog}^K(x) < T$ , then  $\|\nabla F_{T,K,a}(x)\| > 1$ .  
 996

1000 *Proof.* For all  $j \in [T]$ , the partial derivative of  $F_{T,K,a}$  with respect to  $x_j$  is  
 1001

$$1002 \quad \frac{\partial F_{T,K,a}}{\partial x_j}(x) = \left[ -\Psi_a(x_{j-K}) \dots \Psi_a(x_{j-1}) \Phi'(x_j) \right. \\ 1003 \\ 1004 \quad - \Psi_a(x_{j-K+1}) \dots \Psi_a(x_{j-1}) \Psi_a'(x_j) \Phi(x_{j+1}) \\ 1005 \\ 1006 \quad - \dots \\ 1007 \quad \left. - \Psi_a'(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{\min\{j+K, T\}-1}) \Phi(x_{\min\{j+K, T\}}) \right] + \Gamma'(x_j). \\ 1008$$

1009 We now take the smallest  $j \in [T]$  for which  $x_j < 1$  and  $x_{j-1} \geq 1, \dots, x_{j-K} \geq 1$ .  
 1010

1011 If such  $j$  does not exist, then  $x_1 \geq 1$  due to  $x_0 = \dots = x_{-K+1} \equiv 1$ . Then  $x_2 \geq 1$ , and so  
 1012 on. Meaning that  $x_j \geq 1$  for all  $j \in [T]$ , which contradicts the assumption of the theorem that  
 1013  $\text{prog}^K(x) < T$ .  
 1014

Fixing such  $j$ , consider (15). There are two cases.

1015 *Case 1:*  $x_j > -1$ . Note that  $\Psi, \Phi, \Psi', \Phi' \geq 0$  are non-negative and  $\Gamma' \leq 0$  is non-positive. Thus

$$1016 \quad \frac{\partial F_{T,K,a}}{\partial x_j}(x) \leq -\Psi_a(x_{j-K}) \dots \Psi_a(x_{j-2}) \Psi_a(x_{j-1}) \Phi'(x_j).$$

1019 Since  $x_{j-1} \geq 1, \dots, x_{j-K} \geq 1$  and  $1 > x_j > -1$  (see Lemmas D.1 and D.3), we get  
 1020

$$1021 \quad \frac{\partial F_{T,K,a}}{\partial x_j}(x) < -1.$$

1022 *Case 2:*  $x_j \leq -1$ . Note that  $\Psi, \Phi, \Psi', \Phi' \geq 0$  are non-negative. Thus  
 1023

$$1024 \quad \frac{\partial F_{T,K,a}}{\partial x_j}(x) \leq \Gamma'(x_j).$$

1026 Since  $x_j \leq -1$  (see Lemma D.2), we get  
 1027

$$1028 \frac{\partial F_{T,K,a}}{\partial x_j}(x) < -1. \\ 1029$$

1030 Finally, we can conclude that  
 1031

$$1032 \|\nabla F_{T,K,a}(x)\| \geq \left| \frac{\partial F_{T,K,a}}{\partial x_j}(x) \right| > 1. \\ 1033$$

1034  $\square$

1035 **Lemma D.6.** *Function  $F_{T,K,a}$  satisfies*

$$1037 F_{T,K,a}(0) - \inf_{x \in \mathbb{R}^T} F_{T,K,a}(x) \leq \Delta^0(K, a) \cdot T, \\ 1038$$

1039 where  $\Delta^0(K, a) := \sqrt{2\pi e} \cdot a^K$ .  
 1040

1041 *Proof.* Since  $\Gamma(0) = 0$  and  $\Psi_a, \Phi \geq 0$ , we get  $F_{T,K,a}(0) \leq 0$ . Next, due to  $\Gamma(x) \geq 0$ ,  $0 \leq \Phi(x) \leq \sqrt{2\pi e}$  and  $0 \leq \Psi_a(x) \leq a$  for all  $x \in \mathbb{R}^d$ ,

$$1044 F_{T,K,a}(x) \geq - \sum_{i=1}^T \Psi_a(x_{i-K}) \dots \Psi_a(x_{i-2}) \Psi_a(x_{i-1}) \Phi(x_i) \geq -T \sqrt{2\pi e} \cdot a^K \\ 1045 \\ 1046$$

1047 for all  $x \in \mathbb{R}^T$ .  $\square$

1048 **Lemma D.7.** *For all  $x \in \mathbb{R}^T$ ,  $\|\nabla F_{T,K,a}(x)\|_\infty \leq \gamma_\infty(K, a)$ , where  $\gamma_\infty(K, a) := 6\sqrt{2\pi e}^{3/2} \cdot \frac{Ka^K}{\sqrt{\log a}}$ .*

1051 *Proof.* Using (15),

$$1053 \left| \frac{\partial F_{T,K,a}}{\partial x_j}(x) \right| \leq \left| \Psi_a(x_{j-K}) \dots \Psi_a(x_{j-1}) \Phi'(x_j) \right. \\ 1054 \\ 1055 \quad + \Psi_a(x_{j-K+1}) \dots \Psi_a(x_{j-1}) \Psi_a'(x_j) \Phi(x_{j+1}) \\ 1056 \quad + \dots \\ 1058 \quad \left. + \Psi_a'(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{\min\{j+K, T\}-1}) \Phi(x_{\min\{j+K, T\}}) \right| + |\Gamma'(x_j)|. \\ 1059 \\ 1060 \quad (16)$$

1061 Thus,

$$1062 \left| \frac{\partial F_{T,K,a}}{\partial x_j}(x) \right| \leq a^K \sqrt{e} + K a^{K-1} \sqrt{2\pi e} \frac{2e}{\sqrt{\log a}} + e \leq 6\sqrt{2\pi e}^{3/2} \frac{Ka^K}{\sqrt{\log a}} \\ 1063 \\ 1064$$

due to Lemmas D.1, D.2, and D.3.  $\square$

1065 **Lemma D.8.** *The function  $F_{T,K,a}$  is  $\ell_1(K, a)$ -smooth, i.e.,  $\|\nabla^2 F_{T,K,a}(x)\| \leq \ell_1(K, a)$  for all  
 1066  $x \in \mathbb{R}^T$ , where  $\ell_1(K, a) := 12\sqrt{2\pi e}^{5/2} \cdot \frac{K^2 a^K}{\log a}$ .*

1068 *Proof.* Taking the second partial derivative in (15),

$$1070 \frac{\partial^2 F_{T,K,a}}{\partial x_j^2}(x) = \left[ -\Psi_a(x_{j-K}) \dots \Psi_a(x_{j-1}) \Phi''(x_j) \right. \\ 1071 \quad - \Psi_a(x_{j-K+1}) \dots \Psi_a(x_{j-1}) \Psi_a''(x_j) \Phi(x_{j+1}) \\ 1072 \quad - \dots \\ 1075 \quad \left. - \Psi_a''(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{\min\{j+K, T\}-1}) \Phi(x_{\min\{j+K, T\}}) \right] + \Gamma''(x_j). \\ 1076$$

1077 Due to Lemmas D.1, D.2, and D.3,

$$1078 \left| \frac{\partial^2 F_{T,K,a}}{\partial x_j^2}(x) \right| \leq \left[ 27a^K + K \times \frac{56\sqrt{2\pi e}^{3/2} a^{K-1}}{\log a} \right] + 27e^{-2} \leq 168\sqrt{2\pi e}^{3/2} \cdot \frac{Ka^K}{\log a} \\ 1079$$

1080 Clearly, for all  $\min\{j + K, T\} < i \leq T$ ,

$$\frac{\partial^2 F_{T,K,a}}{\partial x_j \partial x_i}(x) = 0 \quad (20)$$

1081 due to the construction of  $F_{T,K,a}$ . Next, for all  $j < i \leq \min\{j + K, T\}$ ,

$$\begin{aligned} \frac{\partial^2 F_{T,K,a}}{\partial x_j \partial x_i}(x) = & \left[ -\Psi_a(x_{i-K}) \dots \Psi_a(x_{j-1}) \Psi'_a(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{i-1}) \Phi'(x_i) \right. \\ & - \Psi_a(x_{i-K+1}) \dots \Psi_a(x_{j-1}) \Psi'_a(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{i-1}) \Psi'_a(x_i) \Phi(x_{i+1}) \\ & - \dots \\ & \left. - \Psi'_a(x_j) \Psi_a(x_{j+1}) \dots \Psi_a(x_{i-1}) \Psi'_a(x_i) \Psi_a(x_{i+1}) \dots \Psi_a(x_{\min\{j+K, T\}-1}) \Phi(x_{\min\{j+K, T\}}) \right], \end{aligned}$$

1093 and

$$\left| \frac{\partial^2 F_{T,K,a}}{\partial x_j \partial x_i}(x) \right| \leq \frac{2e^{3/2}a^{K-1}}{\sqrt{\log a}} + (K-1) \times \frac{4\sqrt{2\pi}e^{5/2}a^{K-2}}{\log a} \leq 4\sqrt{2\pi}e^{5/2} \cdot \frac{Ka^K}{\log a} \quad (21)$$

1094 for all  $i \neq j \in [T]$  due to Lemmas D.1 and D.3 and  $e \geq a > 1$

1095 Notice that  $\nabla^2 F_{T,K,a}$  is  $(2K+1)$ -diagonal Hessian. Repeating a textbook analysis for completeness  
1096 and denoting temporary  $\mathbf{H} := \nabla^2 F_{T,K,a}$ , we will show that

$$\| \nabla^2 F_{T,K,a}(x) \| \leq (2K+1) \max_{i,j \in [T]} \left| \frac{\partial^2 F_{T,K,a}}{\partial x_j \partial x_i}(x) \right|. \quad (22)$$

1103 for all  $x \in \mathbb{R}^T$ . Indeed, for all  $x \in \mathbb{R}^T$  such that  $\|x\| \leq 1$ ,

$$|x^\top \mathbf{H} x| = \left| \sum_{i=1}^T x_i \sum_{j=1}^T x_j \mathbf{H}_{ij} \right| = \left| \sum_{i=1}^T x_i \sum_{j=\max\{i-K, 1\}}^{\min\{i+K, T\}} x_j \mathbf{H}_{ij} \right| \leq \max_{i,j \in [T]} |\mathbf{H}_{ij}| \left( \sum_{i=1}^T |x_i| \sum_{j=\max\{i-K, 1\}}^{\min\{i+K, T\}} |x_j| \right),$$

1108 where the second equality due to  $\mathbf{H}$  is  $(2K+1)$ -diagonal. Using the Cauchy–Schwarz inequality,

$$|x^\top \mathbf{H} x| \leq \max_{i,j \in [T]} |\mathbf{H}_{ij}| \sqrt{\sum_{i=1}^T x_i^2 \sum_{j=\max\{i-K, 1\}}^{\min\{i+K, T\}} |x_j|^2} \leq \max_{i,j \in [T]} |\mathbf{H}_{ij}| \sqrt{\sum_{i=1}^T \left( \sum_{j=\max\{i-K, 1\}}^{\min\{i+K, T\}} |x_j| \right)^2}$$

1114 since  $\|x\| \leq 1$ . Next, using Jensen’s inequality and  $\|x\| \leq 1$ ,

$$\begin{aligned} |x^\top \mathbf{H} x| & \leq \max_{i,j \in [T]} |\mathbf{H}_{ij}| \sqrt{(2K+1) \sum_{i=1}^T \sum_{j=\max\{i-K, 1\}}^{\min\{i+K, T\}} x_j^2} \leq \max_{i,j \in [T]} |\mathbf{H}_{ij}| \sqrt{(2K+1)^2 \sum_{i=1}^T x_i^2} \\ & \leq (2K+1) \max_{i,j \in [T]} |\mathbf{H}_{ij}|. \end{aligned}$$

1121 We have proved (22). It is left to combine (22), (21), and (19).

1122  $\square$

1123

## 1124 E PROOF OF THEOREM 4.2

1125 **Theorem 4.2.** Let  $L, \Delta, \varepsilon, n, \sigma^2, d, \tau_s, \tau_w, h > 0$  be any numbers such that  $\bar{c}_1 \varepsilon \log^4(n+1) < L\Delta$  and  
1126 dimension  $d \geq \bar{c}_3 \frac{L\Delta}{\log^3(n+1)\varepsilon}$ . Consider Protocol 1. For all  $i \in [n]$  and  $k \geq 0$ , compressor  $\mathcal{C}_i^k$  selects  
1127 and transmits  $P_i^k$  uniformly random coordinates without replacement, scaled by any constants<sup>7</sup>, where  
1128  $P_i^k \in \{0, \dots, d\}$  may vary across each compressor<sup>8</sup>. Then, for any algorithm  $A \in \mathcal{A}_{\text{zr}}$  (Def. 2.3),

1129 <sup>7</sup>To potentially preserve unbiasedness. For instance, RandK scales by  $d/K$ .

1130 <sup>8</sup>For instance, the compressors can be RandK (see Def. C.1) with any  $K \in [d]$ , PermK (Szlenkak et al.,  
1131 2021), Identity compressor when  $P_i^k = d$ .

1134 there exists a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $f$  is  $L$ -smooth, i.e.,  $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$   
 1135 for all  $x, y \in \mathbb{R}^d$ , and  $f(0) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta$ , exists a stochastic gradient oracles that satisfies  
 1136 Assumption 1.3, and  $\mathbb{E} \left[ \inf_{y \in S_t} \|\nabla f(y)\|^2 \right] > \varepsilon$  for all  
 1137

$$1138 \quad t \leq \bar{c}_2 \times \left( \frac{1}{\log^3(n+1)} \cdot \frac{L\Delta}{\varepsilon} \right) \min \left\{ \frac{1}{\log(n+1)} \cdot \tau_s d, \max \left\{ h, \frac{1}{\log^3(n+1)} \cdot \frac{h\sigma^2}{\varepsilon} \right\} \right\}, \quad (12)$$

1140 where  $S_t$  is the set of all possible points that can be constructed by  $A$  up to time  $t$  based on  $I$  and  
 1141  $\{I_i\}$ . The quantities  $\bar{c}_1, \bar{c}_2$ , and  $\bar{c}_3$  are universal constants.  
 1142

1143 *Proof. (Step 1: Construction).* Using the construction from Section D.1, we define a scaled version  
 1144 of it. Let us take any  $\lambda > 0$ ,  $d, T \in \mathbb{N}$ ,  $d \geq T$ , and take the function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  
 1145

$$1146 \quad f(x) := \frac{L\lambda^2}{\ell_1(K, a)} F_{T, K, a} \left( \frac{x_{[T]}}{\lambda} \right),$$

1148 where  $\ell_1(K, a)$  is defined in Lemma D.8 and  $x_{[T]} \in \mathbb{R}^T$  is the vector with the first  $T$  coordinates of  
 1149  $x \in \mathbb{R}^d$ . Notice that the last  $d - T$  coordinates are artificial.  
 1150

1151 First, we have to show that  $f$  is  $L$ -smooth and  $f(0) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta$ , Using Lemma D.8,  
 1152

$$1153 \quad \|\nabla f(x) - \nabla f(y)\| = \frac{L\lambda}{\ell_1(K, a)} \left\| \nabla F_{T, K, a} \left( \frac{x_{[T]}}{\lambda} \right) - \nabla F_{T, K, a} \left( \frac{y_{[T]}}{\lambda} \right) \right\| \leq L\lambda \left\| \frac{x_{[T]}}{\lambda} - \frac{y_{[T]}}{\lambda} \right\| \\ 1154 \quad = L \|x_{[T]} - y_{[T]}\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

1155 Let us take

$$1157 \quad T = \left\lfloor \frac{\Delta \cdot \ell_1(K, a)}{L\lambda^2 \cdot \Delta^0(K, a)} \right\rfloor.$$

1158 Due to Lemma D.6,

$$1159 \quad f(0) - \inf_{x \in \mathbb{R}^d} f(x) = \frac{L\lambda^2}{\ell_1(K, a)} (F_{T, K, a}(0) - \inf_{x \in \mathbb{R}^T} F_{T, K, a}(x)) \leq \frac{L\lambda^2 \Delta^0(K, a) T}{\ell_1(K, a)} \leq \Delta,$$

1160 where  $\Delta^0(K, a)$  is defined in Lemma D.6. We also choose  
 1161

$$1164 \quad \lambda = \frac{\sqrt{2\varepsilon} \ell_1(K, a)}{L} \quad (23)$$

1166 to ensure that

$$1168 \quad \|\nabla f(x)\|^2 = \frac{L^2 \lambda^2}{\ell_1^2(K, a)} \left\| \nabla F_{T, K, a} \left( \frac{x_{[T]}}{\lambda} \right) \right\|^2 = 2\varepsilon \left\| \nabla F_{T, K, a} \left( \frac{x_{[T]}}{\lambda} \right) \right\|^2 > 2\varepsilon \cdot \mathbb{1} [\text{prog}^K(x_{[T]}) < T],$$

1170 where the last inequality due to Lemma D.5. Note that

$$1173 \quad T = \left\lfloor \frac{L\Delta}{2\Delta^0(K, a) \cdot \ell_1(K, a) \cdot \varepsilon} \right\rfloor. \quad (25)$$

1175 **(Step 2: Stochastic Oracle).**

1176 We take the stochastic oracle construction form (Arjevani et al., 2022). For all  $j \in [d]$ ,

$$1178 \quad [\nabla f(x; \xi)]_j := \nabla_j f(x) \left( 1 + \mathbb{1} [j > \text{prog}^K(x)] \left( \frac{\xi}{p_\sigma} - 1 \right) \right) \quad \forall x \in \mathbb{R}^d, \quad (26)$$

1180 and  $\mathcal{D}_\xi = \text{Bernoulli}(p_\sigma)$  for all  $j \in [n]$ , where  $p_\sigma \in (0, 1]$ . We denote  $[x]_j$  as the  $j^{\text{th}}$  index of a  
 1181 vector  $x \in \mathbb{R}^d$ . It is left to show this mapping is unbiased and  $\sigma^2$ -variance-bounded. Indeed,

$$1183 \quad \mathbb{E} [[\nabla f(x, \xi)]_i] = \nabla_i f(x) \left( 1 + \mathbb{1} [i > \text{prog}^K(x)] \left( \frac{\mathbb{E}[\xi]}{p_\sigma} - 1 \right) \right) = \nabla_i f(x)$$

1185 for all  $i \in [d]$ , and

$$1187 \quad \mathbb{E} \left[ \|\nabla f(x; \xi) - \nabla f(x)\|^2 \right] \leq \max_{j \in [d]} |\nabla_j f(x)|^2 \mathbb{E} \left[ \left( \frac{\xi}{p_\sigma} - 1 \right)^2 \right]$$

1188 because the difference is non-zero only in one coordinate. Thus  
 1189

$$\begin{aligned} \mathbb{E} \left[ \|\nabla f(x, \xi) - \nabla f(x)\|^2 \right] &\leq \frac{\|\nabla f(x)\|_\infty^2 (1 - p_\sigma)}{p_\sigma} = \frac{L^2 \lambda^2 \|F_{T,K,a} \left( \frac{x_{[T]}}{\lambda} \right)\|_\infty^2 (1 - p_\sigma)}{\ell_1^2(K, a) p_\sigma} \\ &\leq \frac{L^2 \lambda^2 \gamma_\infty^2(K, a) (1 - p_\sigma)}{\ell_1^2(K, a) p_\sigma}, \end{aligned}$$

1195 where we use Lemma D.7 in the last inequality. Taking

$$p_\sigma = \min \left\{ \frac{L^2 \lambda^2 \gamma_\infty^2(K, a)}{\sigma^2 \ell_1^2(K, a)}, 1 \right\} \stackrel{(23)}{=} \min \left\{ \frac{2\varepsilon \gamma_\infty^2(K, a)}{\sigma^2}, 1 \right\}, \quad (27)$$

1199 we ensure that  $\mathbb{E} \left[ \|\nabla f(x, \xi) - \nabla f(x)\|^2 \right] \leq \sigma^2$ .  
 1200

1201 **(Step 3: Reduction to the Analysis of Concentration).** At the beginning, due to Definition 2.3,  
 1202  $s_i^0 = 0$  for all  $i \in [n]$ . Thus, if  $k = 0$ , all workers can receive zero vectors from the server. Thus, at  
 1203 the beginning, all workers can only calculate stochastic gradients at the point zero.

1204 Let  $t^{1,i}$  denote the earliest time at which all of the first  $K$  coordinates become non-zero in the local  
 1205 information available to worker  $i$ . In other words,  $t^{1,i}$  is the first time when worker  $i$  has *discovered*  
 1206 all of the first  $K$  coordinates. Consequently, prior to time  
 1207

$$t_1 := \min_{i \in [n]} t_{1,i} \quad (28)$$

1210 neither the server nor any worker is able to discover (filled with non-zero values) the  $(K + 1)^{\text{th}}$  and  
 1211 subsequent coordinates due to Lemma D.4.

1212 **There are two options by which a worker may discover a new non-zero coordinate: through  
 1213 local stochastic computations or through communication from the server.**

1215 **Option 1:** In the first option, a worker computes a stochastic gradient, which takes  $h$  sec-  
 1216 onds. However, due to the construction of stochastic gradients, even if the computation is completed,  
 1217 the worker will not make progress or discover a new non-zero coordinate, as it will be zeroed out  
 1218 with probability  $p_\sigma$ . Due to Lemma D.4, each worker can discover at most one coordinate at position  
 1219  $\text{prog}^K(x) + 1$  before time  $t_1$  in the first  $K$  coordinates, where  $x$  is a query point.

1220 *Remark E.1.* For this reason, making multiple queries with the same random variable instead of a single  
 1221 query does not help the algorithm progress: if the coordinates are zeroed out, then they are zeroed out in  
 1222 all vectors.

1223 Let  $\eta_{1,i,1}$  be the number of stochastic gradients computations<sup>9</sup> until the first moment when a co-  
 1224 ordinate is not zeroed out in (26) in worker  $i$ . Assume that  $\xi_1, \xi_2, \dots$  is a stream of i.i.d. random  
 1225 Bernoulli variables from (26) in worker  $i$  (all workers have different streams), then

$$\mathbb{P}(\eta_{1,i,1} \leq t) \leq \sum_{k=1}^{\lfloor t \rfloor} \mathbb{P}(\xi_k = 1, \xi_{k-1} = 0, \dots, \xi_1 = 0) = \sum_{k=1}^{\lfloor t \rfloor} p_\sigma (1 - p_\sigma)^{j-1} \leq t p_\sigma.$$

1226 for all  $t \geq 0$ . Similarly, let  $\eta_{1,i,k}$  denote the number of stochastic gradient computations until the first  
 1227 moment when a coordinate is not zeroed out in (26), after the moment when the  $(k-1)^{\text{th}}$  coordinate is  
 1228 no longer zeroed out in worker  $i$ . In other words, worker  $i$  should calculate  $\eta_{1,i,1}$  stochastic gradients  
 1229 to discover the first coordinate, calculate  $\eta_{1,i,2}$  stochastic gradients to discover the second coordinate,  
 1230 and so on. Since the draws of  $\xi$  in (26) are i.i.d., we can conclude that

$$\mathbb{P}(\eta_{1,i,k} \leq t | \eta_{1,i,k-1}, \dots, \eta_{1,i,1}) \leq t p_\sigma$$

1235 for all  $k \geq 1$  and  $t \geq 0$ .  
 1236

1238 **Option 2:** In the second option, worker  $i$  receives  $P \in \{0, \dots, d\}$  random coordinates  
 1239 with the set of indices  $\{\nu_{1,1}, \dots, \nu_{1,P}\}$  without replacement, where it takes  $\tau_s$  seconds to receive one  
 1240

1241 <sup>9</sup>It is possible that  $\mathbb{P}(\eta_{1,i,1} = \infty) > 0$  if, for instance, the algorithm decides to stop calculating stochastic  
 1242 gradients. And even  $\mathbb{P}(\eta_{1,i,1} = \infty) = 1$  if it does not calculate at all.

1242 coordinate. Then, the worker receives  $\bar{P} \in \{0, \dots, d\}$  random coordinates with the set of indices  
 1243  $\{\nu_{2,1}, \dots, \nu_{2,\bar{P}}\}$  *without replacement*, and so on (all workers get different sets; we drop the indices  
 1244 of the workers in the notations).

1245 Consequently, the worker receives a stream of coordinate indices  $(\nu_1, \nu_2, \dots)$  where we concatenated  
 1246 the sets of indices, preserving the exact order in which the server sampled them. Note that all workers  
 1247 have different streams, and we now focus on one worker.

1248 Notice that

$$1250 \quad \mathbb{P}(\nu_1 \in [K]) = \frac{K}{d}$$

1252 since  $\nu_1$  is uniformly random coordinate from the set  $[d]$ . Next,

$$1254 \quad \mathbb{P}(\nu_2 \in [K] | \nu_1) \leq \frac{K}{d-1},$$

1256 because either  $\nu_1 \in [K]$ , in which case the probability is  $\frac{K-1}{d-1}$ , or  $\nu_1 \notin [K]$ , in which case the  
 1257 probability is  $\frac{K}{d-1}$ . Using the same reasoning,

$$1259 \quad \mathbb{P}(\nu_i \in [K] | \nu_1, \dots, \nu_{i-1}) \leq \frac{K}{d-i+1}.$$

1262 for all  $i \leq d$ .

1263 Hence, the worker receives a stream of coordinates  $\nu_1, \nu_2, \dots$  such that  $\mathbb{P}(\nu_i \in [K] | \nu_1, \dots, \nu_{i-1}) \leq$   
 1264  $\frac{K}{d-i+1}$  for all  $i \leq d$ . Let  $\mu_{1,i,1}$  be the number of received coordinates until the moment when the  
 1265 last received coordinate belongs to  $[K]$  in worker  $i$ . Similarly, let  $\mu_{1,i,k}$  be the number of received  
 1266 coordinates until the moment when the last received coordinate belongs to  $[K]$ , after the  $(k-1)^{\text{th}}$   
 1267 time this has happened in worker  $i$ . In other words, worker  $i$  should receive  $\mu_{1,i,1}$  coordinates to  
 1268 obtain a coordinate that belongs to  $[K]$ . To get the next coordinate that belongs to  $[K]$ , the worker  
 1269 should receive  $\mu_{1,i,2}$  coordinates, and so on. Then,

$$1270 \quad \mathbb{P}(\mu_{1,i,1} = j) = \mathbb{P}(\nu_j \in [K], \nu_{j-1} \notin [K], \dots, \nu_1 \notin [K]) \\ 1271 \quad = \mathbb{P}(\nu_j \in [K] | \nu_{j-1} \notin [K], \dots, \nu_1 \notin [K]) \mathbb{P}(\nu_{j-1} \notin [K], \dots, \nu_1 \notin [K]) \leq \frac{K}{d-j+1}$$

1274 for all  $1 \leq j \leq d$ , and

$$1276 \quad \mathbb{P}(\mu_{1,i,1} \leq t) = \sum_{j=1}^{\lfloor t \rfloor} \mathbb{P}(\mu_{1,i,1} = j) \leq \frac{Kt}{d-t+1}. \quad (29)$$

1279 for all  $0 \leq t \leq d$ . Similarly,

$$1281 \quad \mathbb{P}(\mu_{1,i,k} = j | \mu_{1,i,k-1}, \dots, \mu_{1,i,1}) \\ 1282 \quad = \mathbb{P}(\nu_{u+j} \in [K], \nu_{u+j-1} \notin [K], \dots, \nu_{u+1} \notin [K] | \nu_u \in [K], \dots, \nu_1 \notin [K]) \\ 1283 \quad \leq \frac{K}{\max\{d-u, 0\} - j + 1},$$

1286 where  $u = \sum_{j=1}^{k-1} \mu_{1,i,j}$ , for all  $j \leq \max\{d-u, 0\}$ . Thus,

$$1288 \quad \mathbb{P}(\mu_{1,i,k} \leq t | \mu_{1,i,k-1}, \dots, \mu_{1,i,1}) \leq \frac{Kt}{\max\{d - \sum_{j=1}^{k-1} \mu_{1,i,j}, 0\} - t + 1},$$

1290 for all  $0 \leq t \leq \max\{d - \sum_{j=1}^{k-1} \mu_{1,i,j}, 0\}$ .

1292 Recall that the workers can discover new non-zero coordinates only through the stochastic processes  
 1293 discussed above. To discover all of the first  $K$  coordinates, either the first or the second process  
 1294 must uncover at least  $\frac{K}{2}$  coordinates<sup>10</sup>. If worker  $i$  has discovered fewer than  $\frac{K}{2}$  coordinates through

1295 <sup>10</sup>At the end of the proof, we take  $K \bmod 2 = 0$ .

1296 stochastic gradient computations and fewer than  $\frac{K}{2}$  coordinates through receiving coordinates from  
 1297 the server, then it will not be able to cover all  $K$  coordinates. Hence,  
 1298

$$1301 \quad t_1 \geq \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{1,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{1,i,k} \right\} \right\}, \quad (30)$$

1304 where  $t_1$  is defined in (28). This is because  $h \sum_{k=1}^{\frac{K}{2}} \eta_{1,i,k}$  is the time required to obtain  $\frac{K}{2}$  “lucky”  
 1305 stochastic gradients, those for which the coordinates are not zeroed out, and  $\tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{1,i,k}$  is the  
 1306 time required to receive  $\frac{K}{2}$  “lucky” coordinates that belong to  $[K]$ .  
 1307

1308 *Remark E.2.* The previous derivations hold for all  $\tau_w > 0$ . If we start taking the communication time  $\tau_w$   
 1309 into account, then the bound on  $t_1$  in (30) may only increase. For all  $\tau_w > 0$ , worker  $i$  still has to discover  
 1310 new non-zero coordinates either through stochastic gradient computations or by receiving coordinates from  
 1311 the server and it will take at least

$$1313 \quad \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{1,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{1,i,k} \right\} \right\}$$

1316 seconds to discover all of the first  $K$  coordinates.

1317 Once the workers have discovered the first  $K$  coordinates, the discovery process repeats for the set  
 1318  $\{K+1, \dots, 2K\}$ , which similarly requires at least

$$1320 \quad \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{2,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{2,i,k} \right\} \right\}$$

1324 seconds, where  $\{\eta_{b,i,k}\}$  and  $\{\mu_{b,i,k}\}$  are random variables such that

$$1325 \quad \mathbb{P}(\eta_{b,i,k} \leq t | \eta_{b,i,k-1}, \dots, \eta_{b,i,1}, \mathcal{G}_{b-1}) \leq tp_\sigma \quad (31)$$

1327 for all  $b \geq 1, k \geq 1, i \in [n], t \geq 0$ , and

$$1328 \quad \mathbb{P}(\mu_{b,i,k} \leq t | \mu_{b,i,k-1}, \dots, \mu_{b,i,1}, \mathcal{G}_{b-1}) \leq \frac{Kt}{\max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\} - t + 1} \quad (32)$$

1331 for all  $b \geq 1, k \geq 1, i \in [n]$ , and  $t \leq \max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\}$ , where  $\mathcal{G}_{b-1}$  is the sigma-algebra  
 1332 generated by  $\{\eta_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$  and  $\{\mu_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$  and  $u = \sum_{j=1}^{k-1} \mu_{b,i,j}$ .  
 1333

1334 More formally,  $\eta_{2,i,k}$  can be defined as the number of stochastic gradient computations until the first  
 1335 moment when a coordinate is not zeroed out in (26), after the moment when the  $(k-1)^{\text{th}}$  coordinate  
 1336 is no longer zeroed out, when  $\text{prog}^K$  of the input points to the stochastic gradients is  $\geq K$ , and  $\mu_{2,i,k}$   
 1337 be the number of received coordinates until the moment when the last received coordinate belongs to  
 1338  $\{K+1, \dots, 2K\}$ , after the  $(k-1)^{\text{th}}$  time this has happened, when  $\text{prog}^1$  of the input points to the  
 1339 compressor is  $\geq K+1$ , and so on.

1340 We define

$$1342 \quad p_K := \frac{2K}{d}. \quad (33)$$

1344 Finally, to discover the  $T^{\text{th}}$  coordinates it takes at least

$$1346 \quad \sum_{b=1}^B \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{b,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{b,i,k} \right\} \right\}$$

1349 seconds, where  $B = \lfloor \frac{T}{K} \rfloor$ . It is left to use the following lemma.

1350 **Lemma E.3.** Let  $\{\eta_{b,i,j}\}_{i,j,b \geq 0}$  and  $\{\mu_{b,i,j}\}_{i,j,b \geq 0}$  be random variables such that

$$1352 \quad \mathbb{P}(\eta_{b,i,k} \leq t | \eta_{b,i,k-1}, \dots, \eta_{b,i,1}, \mathcal{G}_{b-1}) \leq tp_\sigma \quad (34)$$

1353 for all  $b \geq 1, k \geq 1, i \in [n], t \geq 0$ , and

$$1355 \quad \mathbb{P}(\mu_{b,i,k} \leq t | \mu_{b,i,k-1}, \dots, \mu_{b,i,1}, \mathcal{G}_{b-1}) \leq \frac{Kt}{\max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\} - t + 1}, \quad (35)$$

1357 for all  $b \geq 1, k \geq 1, i \in [n], 0 \leq t \leq \max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\}$ , and  $1 \leq K \leq d$ , where  $\mathcal{G}_{b-1}$  is the  
1358 sigma-algebra generated by  $\{\eta_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$  and  $\{\mu_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$ . Then

$$1361 \quad \mathbb{P}\left(\sum_{b=1}^B \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{b,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{b,i,k} \right\} \right\} \leq \bar{t}\right) \leq \delta$$

1364 with

$$1365 \quad \bar{t} := \frac{BK + \log \delta}{e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n))} \min \left\{ \frac{h}{p_\sigma}, \frac{\tau_s}{p_K} \right\}, \quad (36)$$

1368 where

$$1369 \quad p_K := \frac{2K}{d}.$$

1372 **(Step 4: Endgame).** Thus, with probability at least  $1 - \delta$ , any zero-respecting algorithm requires at  
1373 least  $\bar{t}$  seconds to discover the  $T^{\text{th}}$  coordinate. Since  $\text{prog}^K(x) \leq \text{prog}^1(x)$  for all  $x \in \mathbb{R}^T$ , and due  
1374 to (24),

$$1375 \quad \inf_{y \in S_t} \|\nabla f(y)\|^2 > 2\varepsilon \inf_{y \in S_t} \mathbb{1}[\text{prog}^1(y_{[T]}) < T],$$

1377 where  $S_t$  is the set of all possible candidate points to be an  $\varepsilon$ -stationary point up to time  $t$ , which can  
1378 be computed by  $A$ . Taking  $\delta = \frac{1}{2}$ ,

$$1380 \quad \mathbb{E} \left[ \inf_{y \in S_t} \|\nabla f(y)\|^2 \right] > 2\varepsilon \mathbb{E} \left[ \inf_{y \in S_t} \mathbb{1}[\text{prog}^1(y_{[T]}) < T] \right] \geq \varepsilon,$$

1382 for  $t = \frac{1}{2}\bar{t}$  because  $\text{prog}^1(y_{[T]}) < T$  for all  $y \in S_t$  with probability at least  $\frac{1}{2}$ .

1384 It is left to choose  $K$  and  $a$ , and substitute all quantities to  $\bar{t}$ . Using  $B = \lfloor \frac{T}{K} \rfloor$ ,

$$1386 \quad \bar{t} = \frac{\lfloor \frac{T}{K} \rfloor K - \log 8}{e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n))} \min \left\{ \frac{h}{p_\sigma}, \frac{\tau_s}{p_K} \right\}$$

$$1388 \quad \geq \frac{T - K - \log 8}{e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n))} \min \left\{ \frac{h}{p_\sigma}, \frac{\tau_s}{p_K} \right\}.$$

1391 Due to (25), (27), and (33),

$$1393 \quad \bar{t} \geq \frac{\left\lfloor \frac{L\Delta}{2\Delta^0(K,a) \cdot \ell_1(K,a) \cdot \varepsilon} \right\rfloor - K - \log 8}{e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n))} \min \left\{ \max \left\{ h, \frac{h\sigma^2}{2\varepsilon\gamma_\infty^2(K,a)} \right\}, \frac{\tau_s d}{2K} \right\}.$$

1395 Using the definitions of  $\Delta^0(K,a)$ ,  $\gamma_\infty(K,a)$ , and  $\ell_1(K,a)$ ,

$$1397 \quad \bar{t} \geq \left( e^4(2n)^{2/K} \left( 4 + \frac{2}{K} \log(2n) \right) \right)^{-1} \left( \left\lfloor \frac{L\Delta \log a}{48\pi e^3 K^2 a^{2K} \varepsilon} \right\rfloor - K - \log 8 \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2 \log a}{144\pi e^3 K^2 a^{2K} \varepsilon} \right\}, \frac{\tau_s d}{2K} \right\}.$$

1400 We can take any  $a$  from the interval  $(1, e]$ . We choose  $a = 1 + \frac{1}{K}$ , then  $\log a = \log(1 + \frac{1}{K}) \geq \frac{1}{2K}$   
1401 for all  $K \geq 1$ ,  $a^{2K} \leq e^2$  for all  $K \geq 1$ , and

$$1403 \quad \bar{t} \geq \left( e^4(2n)^{2/K} \left( 4 + \frac{2}{K} \log(2n) \right) \right)^{-1} \left( \left\lfloor \frac{L\Delta}{96\pi e^5 K^3 \varepsilon} \right\rfloor - K - \log 8 \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2}{288\pi e^5 K^3 \varepsilon} \right\}, \frac{\tau_s d}{2K} \right\}.$$

1404 Taking  $K = 2 \lceil 2 \log(2n) \rceil$ ,  $(2n)^{2/K} \leq e$ ,  $K \leq 16 \log(n+1)$  and  
 1405  
 1406  $\bar{t} \geq \frac{1}{5e^5} \left( \left\lfloor \frac{L\Delta}{96 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} \right\rfloor - 32 \log(n+1) \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2}{288 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} \right\}, \frac{\tau_s d}{32 \log(n+1)} \right\}$   
 1407  
 1408  $\geq \frac{1}{5e^5} \left( \frac{L\Delta}{96 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} - 36 \log(n+1) \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2}{288 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} \right\}, \frac{\tau_s d}{32 \log(n+1)} \right\}.$   
 1409  
 1410

1411 We assume  $\frac{L\Delta}{\varepsilon} \geq \bar{c}_1 \log^4(n+1)$  for a universal constant  $\bar{c}_1$ . Taking  $\bar{c}_1$  large enough, one can see  
 1412 that

1413  $\bar{t} \geq \frac{1}{5e^5} \left( \frac{L\Delta}{2 \cdot 96 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2}{288 \cdot 8^4 \pi e^5 \log^3(n+1) \varepsilon} \right\}, \frac{\tau_s d}{32 \log(n+1)} \right\}.$   
 1414  
 1415 For a small enough universal  $\bar{c}_2$ , we get the inequality

1416  
 1417  $\bar{t} \geq \bar{c}_2 \times \left( \frac{L\Delta}{\log^3(n+1) \varepsilon} \right) \min \left\{ \max \left\{ h, \frac{h\sigma^2}{\log^3(n+1) \varepsilon} \right\}, \frac{\tau_s d}{\log(n+1)} \right\},$   
 1418

1419 which finishes the proof. Notice that we can take

1420  
 1421  $d \geq T = \left\lfloor \frac{L\Delta \log a}{48\pi e^3 K^2 a^{2K} \varepsilon} \right\rfloor = \Theta \left( \frac{L\Delta}{\log^3(n+1) \varepsilon} \right).$   
 1422

1423  $\square$   
 1424

## 1425 E.1 MAIN CONCENTRATION LEMMA

1426  
 1427 **Lemma E.3.** Let  $\{\eta_{b,i,j}\}_{i,j,b \geq 0}$  and  $\{\mu_{b,i,j}\}_{i,j,b \geq 0}$  be random variables such that

1428  $\mathbb{P}(\eta_{b,i,k} \leq t | \eta_{b,i,k-1}, \dots, \eta_{b,i,1}, \mathcal{G}_{b-1}) \leq tp_\sigma \quad (34)$   
 1429

1430 for all  $b \geq 1, k \geq 1, i \in [n], t \geq 0$ , and

1431  
 1432  $\mathbb{P}(\mu_{b,i,k} \leq t | \mu_{b,i,k-1}, \dots, \mu_{b,i,1}, \mathcal{G}_{b-1}) \leq \frac{Kt}{\max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\} - t + 1}, \quad (35)$   
 1433

1434 for all  $b \geq 1, k \geq 1, i \in [n], 0 \leq t \leq \max\{d - \sum_{j=1}^{k-1} \mu_{b,i,j}, 0\}$ , and  $1 \leq K \leq d$ , where  $\mathcal{G}_{b-1}$  is the  
 1435 sigma-algebra generated by  $\{\eta_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$  and  $\{\mu_{b',i,k}\}_{i \in [n], k \in [\frac{K}{2}], b' < b}$ . Then

1436  
 1437  $\mathbb{P} \left( \sum_{b=1}^B \min_{i \in [n]} \left\{ \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{b,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{b,i,k} \right\} \right\} \leq \bar{t} \right) \leq \delta$   
 1438

1439 with

1440  
 1441  $\bar{t} := \frac{BK + \log \delta}{e^4 (2n)^{2/K} (4 + \frac{2}{K} \log(2n))} \min \left\{ \frac{h}{p_\sigma}, \frac{\tau_s}{p_K} \right\}, \quad (36)$   
 1442

1443 where

1444  
 1445  $p_K := \frac{2K}{d}.$   
 1446

1447  
 1448 *Proof.* Let us temporarily define  $\beta_{b,i} := \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{b,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{b,i,k} \right\}$ . Using Chernoff's  
 1449 method, we get

1450  
 1451  $\mathbb{P} \left( \sum_{b=1}^B \min_{i \in [n]} \beta_{b,i} \leq \bar{t} \right) = \mathbb{P} \left( \exp \left( - \sum_{b=1}^B \lambda \min_{i \in [n]} \beta_{b,i} \right) \geq \exp(-\lambda \bar{t}) \right)$   
 1452  
 1453  $\leq \exp(\lambda \bar{t}) \mathbb{E} \left[ \exp \left( - \sum_{b=1}^B \lambda \min_{i \in [n]} \beta_{b,i} \right) \right] \quad (37)$   
 1454  
 1455  
 1456  $= \exp(\lambda \bar{t}) \mathbb{E} \left[ \mathbb{E} \left[ \exp \left( - \lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \exp \left( - \sum_{b=1}^{B-1} \lambda \min_{i \in [n]} \beta_{b,i} \right) \right]$   
 1457

1458 for all  $\lambda > 0$  since  $\beta_{1,i}, \dots, \beta_{B-1,i}$  are  $\mathcal{G}_{B-1}$ -measurable. Consider the inner expectation separately:  
1459

$$1460 \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] = \mathbb{E} \left[ \max_{i \in [n]} \exp (-\lambda \beta_{B,i}) \middle| \mathcal{G}_{B-1} \right] \leq \sum_{i=1}^n \mathbb{E} [\exp (-\lambda \beta_{B,i}) \mid \mathcal{G}_{B-1}],$$

1463 where we bound max by  $\sum$ . Using the temporal definitions of  $\{\beta_{B,i}\}$ ,

$$\begin{aligned} 1464 \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \\ 1465 &\leq \sum_{i=1}^n \mathbb{E} \left[ \exp \left( -\lambda \min \left\{ h \sum_{k=1}^{\frac{K}{2}} \eta_{B,i,k}, \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{B,i,k} \right\} \right) \middle| \mathcal{G}_{B-1} \right] \\ 1466 \\ 1467 &= \sum_{i=1}^n \mathbb{E} \left[ \max \left\{ \exp \left( -\lambda h \sum_{k=1}^{\frac{K}{2}} \eta_{B,i,k} \right), \exp \left( -\lambda \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{B,i,k} \right) \right\} \middle| \mathcal{G}_{B-1} \right] \\ 1468 \\ 1469 &\leq \underbrace{\sum_{i=1}^n \mathbb{E} \left[ \exp \left( -\lambda h \sum_{k=1}^{\frac{K}{2}} \eta_{B,i,k} \right) \middle| \mathcal{G}_{B-1} \right]}_{I_1 :=} + \underbrace{\sum_{i=1}^n \mathbb{E} \left[ \exp \left( -\lambda \tau_s \sum_{k=1}^{\frac{K}{2}} \mu_{B,i,k} \right) \middle| \mathcal{G}_{B-1} \right]}_{I_2 :=}. \end{aligned} \quad (38)$$

1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479 Using the tower property,

$$1480 \\ 1481 I_1 = \sum_{i=1}^n \mathbb{E} \left[ \underbrace{\mathbb{E} \left[ \exp \left( -\lambda h \eta_{B,i,\frac{K}{2}} \right) \middle| \eta_{B,i,\frac{K}{2}-1}, \dots, \eta_{B,i,1}, \mathcal{G}_{B-1} \right]}_{J_1 :=} \exp \left( -\lambda h \sum_{k=1}^{\frac{K}{2}-1} \eta_{B,i,k} \right) \middle| \mathcal{G}_{B-1} \right]. \quad (39)$$

1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213  
2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267  
2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321  
2322  
2323  
2324  
2325  
2326  
2327  
2328  
2329  
2330  
2331  
2332  
2333  
2334  
2335  
2336  
2337  
2338  
2339  
2340  
2341  
2342  
2343  
2344  
2345  
2346  
2347  
2348  
2349  
2350  
2351  
2352  
2353  
2354  
2355  
2356  
2357  
2358  
2359  
2359  
2360  
2361  
2362  
2363  
2364  
2365  
2366  
2367  
2368  
2369  
2369  
2370  
2371  
2372  
2373  
2374  
2375  
2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429  
2429  
2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483  
2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2496  
2497  
2498  
2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
2509  
2510  
2511  
2512  
2513  
2514  
2515  
2516  
2517  
2518  
2519  
2519  
2520  
2521  
2522  
2523  
2524  
2525  
2526  
2527  
2528  
2529  
2529  
2530  
2531  
2532  
2533  
2534  
2535  
2536  
2537  
2538  
2539  
2539  
2540  
2541  
2542  
2543  
2544  
2545  
2546  
2547  
2548  
2549  
2549  
2550  
2551  
2552  
2553  
2554  
2555  
2556  
2557  
2558  
2559  
2559  
2560  
2561  
2562  
2563  
2564  
2565  
2566  
2567  
2568  
2569  
2569  
2570  
2571  
2572  
2573  
2574  
2575  
2576  
2577  
2578  
2579  
2579  
2580  
2581  
2582  
2583  
2584  
2585  
2586  
2587  
2588  
2589  
2589  
2590  
2591  
2592  
2593  
2594  
2595  
2596  
2597  
2598  
2599  
2599  
2600  
2601  
2602  
2603  
2604  
2605  
2606  
2607  
2608  
2609  
2609  
2610  
2611  
2612  
2613  
2614  
2615  
2616  
2617  
2618  
2619  
2619  
2620  
2621  
2622  
2623  
2624  
2625  
2626  
2627  
2628  
2629  
2629  
2630  
2631  
2632  
2633  
2634  
2635  
2636  
2637  
2638  
2639  
2639  
2640  
2641  
2642  
2643  
2644  
2645  
2646  
2647  
2648  
2649  
2649  
2650  
2651  
2652  
2653  
2654  
2655  
2656  
2657  
2658  
2659  
2659  
2660  
2661  
2662  
2663  
2664  
2665  
2666  
2667  
2668  
2669  
2669  
2670  
2671  
2672  
2673  
2674  
2675  
2676  
2677  
2678  
2679  
2679  
2680  
2681  
2682  
2683  
2684  
2685  
2686  
2687  
2688  
2689  
2689  
2690  
2691  
2692  
2693  
2694  
2695  
2696  
2697  
2698  
2699  
2699  
2700  
2701  
2702  
2703  
2704  
2705  
2706  
2707  
2708  
2709  
2709  
2710  
2711  
2712  
2713  
2714  
2715  
2716  
2717  
2718  
2719  
2719  
2720  
2721  
2722  
2723  
2724  
2725  
2726  
2727  
2728  
2729  
2729  
2730  
2731  
2732  
2733  
2734  
2735  
2736  
2737  
2738  
2739  
2739  
2740  
2741  
2742  
2743  
2744  
2745  
2746  
2747  
2748  
2749  
2749  
2750  
2751  
2752  
2753  
2754  
2755  
2756  
2757  
2758  
2759  
2759  
2760  
2761  
2762  
2763  
2764  
2765  
2766  
2767  
2768  
2769  
2769  
2770  
2771  
2772  
2773  
2774  
2775  
2776  
2777  
2778  
2779  
2779  
2780  
2781  
2782  
2783  
2784  
2785  
2786  
2787  
2788  
2789  
2789  
2790  
2791  
2792  
2793  
2794  
2795  
2796  
2797  
2798  
2799  
2799  
2800  
2801  
2802  
2803  
2804  
2805  
2806  
2807  
2808  
2809  
2809  
2810  
2811  
2812  
2813  
2814  
2815  
2816  
2817  
2818  
2819  
2819  
2820  
2821  
2822  
2823  
2824  
2825  
2826  
2827  
2828  
2829  
2829  
2830  
2831  
2832  
2833  
2834  
2835  
2836  
2837  
2838  
2839  
2839  
2840  
2841  
2842  
2843  
2844  
2845  
2846  
2847  
2848  
2849  
2849  
2850  
2851  
2852  
2853  
2854  
2855  
2856  
2857  
2858  
2859  
2859  
2860  
2861  
2862  
2863  
2864  
2865  
2866  
2867  
2868  
2869  
2869  
2870  
2871  
2872  
2873  
2874  
2875  
2876  
2877  
2878  
2879  
2879  
2880  
2881  
2882  
2883  
2884  
2885  
2886  
2887  
2888  
2889  
2889  
2890  
2891  
2892  
2893  
2894  
2895  
2896  
2897  
2898  
2899  
2899  
2900  
2901  
2902  
2903  
2904  
2905  
2906  
2907  
2908  
2909  
2909  
2910  
2911  
2912  
2913  
2914  
2915  
2916  
2917  
2918  
2919  
2919  
2920  
2921  
2922  
2923  
2924  
2925  
2926  
2927  
2928  
2929  
2929  
2930  
2931  
2932  
2933  
2934  
2935  
2936  
2937  
2938  
2939  
2939  
2940  
2941  
2942  
2943  
2944  
2945  
2946  
2947  
2948  
2949  
2949  
2950  
2951  
2952  
2953  
2954  
2955  
2956  
2957  
2958  
2959  
2959  
2960  
2961  
2962  
2963  
2964  
2965  
2966  
2967  
2968  
2969  
2969  
2970  
2971  
2972  
2973  
2974  
2975  
2976  
2977  
2978  
2979  
2979  
2980  
2981  
2982  
2983  
2984  
2985  
2986  
2987  
2988  
2989  
2989  
2990  
2991  
2992  
2993  
2994  
2995  
2996  
2997  
2998  
2998  
2999  
2999  
3000  
3001  
3002  
3003  
3004  
3005  
3006  
3007  
3008  
3009  
3009  
3010  
3011  
3012  
3013  
3014  
3015  
3016  
3017  
3018  
3019  
3019  
3020  
3021  
3022  
3023  
3024  
3025  
3026  
3027  
3028  
3029  
3029  
3030  
3031  
3032  
3033  
3034  
3035  
3036  
3037  
3038  
3039  
3039  
3040  
3041  
3042  
3043  
3044  
3045  
3046  
3047  
3048  
3049  
3049  
3050  
3051  
3052  
3053  
3054  
3055  
3056  
3057  
3058  
3059  
3059  
3060  
3061  
3062  
3063  
3064  
3065  
3066  
3067  
3068  
3069  
3069  
3070  
3071  
3072  
3073  
3074  
3075  
3076  
3077  
3078  
3079  
3079  
3080  
3081  
3082  
3083  
3084  
3085  
3086  
3087  
3088  
3089  
3089  
3090  
3091  
3092  
3093  
3094  
3095  
3096  
3097  
309

1512 for all  $t \geq 0$ . The analysis of  $I_2^{K/2}$  a little bit more evolved. For all  $1 \leq j \leq \frac{K}{2}$ ,

$$\begin{aligned} 1514 \quad I_2^j &:= \sum_{i=1}^n \mathbb{E} \left[ \exp \left( -\lambda \tau_s \sum_{k=1}^j \mu_{B,i,k} \right) \middle| \mathcal{G}_{B-1} \right] \\ 1515 \quad &= \sum_{i=1}^n \mathbb{E} \left[ \underbrace{\mathbb{E} [\exp (-\lambda \tau_s \mu_{B,i,j}) | \mu_{B,i,j-1}, \dots, \mu_{B,i,1}, \mathcal{G}_{B-1}] \exp \left( -\lambda \tau_s \sum_{k=1}^{j-1} \mu_{B,i,k} \right)}_{K_2 :=} \middle| \mathcal{G}_{B-1} \right] \\ 1516 \quad & \\ 1517 \quad & \\ 1518 \quad & \\ 1519 \quad & \\ 1520 \quad & \\ 1521 \quad & \end{aligned}$$

1522 and  $I_2^0 = n$ . Let us define  $u = \sum_{k=1}^{j-1} \mu_{B,i,k}$ . If  $u \geq \frac{d}{2}$ , then

$$1524 \quad K_2 = \mathbb{E} [\exp (-\lambda \tau_s \mu_{B,i,j}) | \mu_{B,i,j-1}, \dots, \mu_{B,i,1}, \mathcal{G}_{B-1}] \exp (-\lambda \tau_s u) \leq \exp \left( -\frac{\lambda \tau_s d}{2} \right) \\ 1525 \quad \\ 1526 \quad \\ 1527 \quad \text{Otherwise, if } u < \frac{d}{2}, \text{ then, for all } t \geq 0,$$

$$\begin{aligned} 1528 \quad & \mathbb{E} [\exp (-\lambda \tau_s \mu_{B,i,j}) | \mu_{B,i,j-1}, \dots, \mu_{B,i,1}, \mathcal{G}_{B-1}] \\ 1529 \quad & \leq \exp (-\lambda t) + \mathbb{P} \left( \mu_{B,i,j} \leq \frac{t}{\tau_s} \middle| \mu_{B,i,j-1}, \dots, \mu_{B,i,1}, \mathcal{G}_{B-1} \right) \\ 1530 \quad & \\ 1531 \quad & \leq \exp (-\lambda t) + \frac{K \frac{t}{\tau_s}}{d - u - \frac{t}{\tau_s} + 1} \leq \exp (-\lambda t) + \frac{2Kt}{\tau_s d - 2t} \\ 1532 \quad & \\ 1533 \quad & \\ 1534 \quad & \end{aligned}$$

1535 due to (35) and  $u < \frac{d}{2}$ . Combining both cases,

$$1536 \quad K_2 \leq \max \left\{ \left( \exp (-\lambda t) + \frac{2Kt}{\tau_s d - 2t} \right) \exp \left( -\lambda \tau_s \sum_{k=1}^{j-1} \mu_{B,i,k} \right), \exp \left( -\frac{\lambda \tau_s d}{2} \right) \right\} \\ 1537 \quad \\ 1538 \quad \\ 1539 \quad \\ 1540 \quad \text{for all } t < \frac{\tau_s d}{2} \text{ and } u \geq 0, \text{ and}$$

$$\begin{aligned} 1541 \quad I_2^j &\leq \left( \exp (-\lambda t) + \frac{2Kt}{\tau_s d - 2t} \right) \sum_{i=1}^n \mathbb{E} \left[ \exp \left( -\lambda \tau_s \sum_{k=1}^{j-1} \mu_{B,i,k} \right) \middle| \mathcal{G}_{B-1} \right] + n \exp \left( -\frac{\lambda \tau_s d}{2} \right) \\ 1542 \quad & \\ 1543 \quad & \\ 1544 \quad & \\ 1545 \quad & \\ 1546 \quad & \end{aligned} \tag{41}$$

1547 where we use the inequality  $\max\{a, b\} \leq a + b$  for all  $a, b \geq 0$ . Substituting (40) to (38),

$$1548 \quad \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \leq n \left( \exp (-\lambda t) + \frac{tp_\sigma}{h} \right)^{\frac{K}{2}} + I_2^{\frac{K}{2}}, \\ 1549 \quad \\ 1550 \quad \\ 1551 \quad \text{where } t, \lambda \geq 0 \text{ are free parameters. Taking } t = \frac{4 + \frac{2}{K} \log(2n)}{\lambda}, \\ 1552 \quad \\ 1553 \quad \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \leq n \left( \frac{e^{-4}}{(2n)^{2/K}} + \frac{(4 + \frac{2}{K} \log(2n))p_\sigma}{\lambda h} \right)^{\frac{K}{2}} + I_2^{\frac{K}{2}}. \\ 1554 \quad \\ 1555 \quad \\ 1556 \quad \text{Choosing } \lambda = e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n)) \max \left\{ \frac{p_\sigma}{h}, \frac{p_K}{\tau_s} \right\}, \\ 1557 \quad \\ 1558 \quad \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \leq n \left( \frac{2e^{-4}}{(2n)^{2/K}} \right)^{\frac{K}{2}} + I_2^{\frac{K}{2}} = \frac{1}{2} (2e^{-4})^{\frac{K}{2}} + I_2^{\frac{K}{2}}. \\ 1559 \quad \\ 1560 \quad \\ 1561 \quad \text{With this choice of } \lambda \text{ and } t \text{ in (41), we get}$$

$$\begin{aligned} 1562 \quad I_2^j &\leq \left( \frac{3e^{-4}}{(2n)^{2/K}} \right) I_2^{j-1} + n \exp \left( -e^4(2n)^{2/K}(4K + 2 \log(2n)) \right) \\ 1563 \quad & \\ 1564 \quad & \\ 1565 \quad & \end{aligned}$$

1566 for all  $j \geq 1$  because  $t \leq \frac{\tau_s d}{2e^4(2n)^{2/K}K}$ ,  $\lambda = \frac{4 + \frac{2}{K} \log(2n)}{t} \geq 2e^4(2n)^{2/K}(4 + \frac{2}{K} \log(2n))\frac{K}{\tau_s d}$ . In the  
 1567 third inequality, we unrolled the recursion with  $I_2^0 = n$  and use  $\sum_{j=0}^{\infty} \left(\frac{3e^{-4}}{(2n)^{2/K}}\right)^j \leq 2$ .  
 1568

1569  
 1570 Finally,  $I_2^{K/2} \leq \frac{1}{2} (3e^{-4})^{\frac{K}{2}} + 2e^{-2e^4 K} \leq \frac{1}{2} e^{-K}$  and  
 1571

$$1572 \mathbb{E} \left[ \exp \left( -\lambda \min_{i \in [n]} \beta_{B,i} \right) \middle| \mathcal{G}_{B-1} \right] \leq \frac{1}{2} (2e^{-4})^{\frac{K}{2}} + I_2^{K/2} \leq \frac{1}{2} e^{-K} + \frac{1}{2} e^{-K} \leq e^{-K}. \\ 1573$$

1574 Substituting the last inequality to (37) and repeating the steps  $B - 1$  more times, we get  
 1575

$$1576 \mathbb{P} \left( \sum_{b=1}^B \min_{i \in [n]} \beta_{b,i} \leq \bar{t} \right) \leq \exp(\lambda \bar{t} - BK). \\ 1577 \\ 1578$$

1579 It is left to take  $\bar{t} = \frac{BK + \log \delta}{\lambda}$ . □  
 1580

1581  
 1582  
 1583  
 1584  
 1585  
 1586  
 1587  
 1588  
 1589  
 1590  
 1591  
 1592  
 1593  
 1594  
 1595  
 1596  
 1597  
 1598  
 1599  
 1600  
 1601  
 1602  
 1603  
 1604  
 1605  
 1606  
 1607  
 1608  
 1609  
 1610  
 1611  
 1612  
 1613  
 1614  
 1615  
 1616  
 1617  
 1618  
 1619

1620 **F MAIN THEOREM WITH WORKER-TO-SERVER COMMUNICATION**  
 1621

1622 In this section, we extend the result of Theorem 4.2 by taking into account the communication time  
 1623  $\tau_w$ . However, in this section, we ignore the communication times from the server to the workers **in the**  
 1624 **analysis**, which will be sufficient to obtain an almost tight lower bound if combined with Theorem 4.2.

1625 Tyurin et al. (2024) consider a similar setup with  $\tau_s = 0$ . However, their protocol does not allow the  
 1626 workers to modify the iterate computed by the server and operates in the primal space. For instance,  
 1627 the workers are not allowed to run local steps. Moreover,  $\bar{P}_i^k$  are fixed in their version of Protocol 1.  
 1628 We improve upon this in the following theorem:

1629 **Theorem F.1.** *Let  $L, \Delta, \varepsilon, \sigma^2, n, d, \tau_w, \tau_s, h > 0$  be any numbers such that  $\varepsilon < c_1 L \Delta$  and  $d \geq \frac{\Delta L}{c_2 \varepsilon}$ .*

1630 *Consider Protocol 1. For all  $i \in [n]$  and  $k \geq 0$ , compressor  $\bar{C}_i^k$  selects and transmits  $\bar{P}_i^k$  uniformly*  
 1631 *random coordinates without replacement, scaled by any constants, where  $\bar{P}_i^k \in \{0, \dots, d\}$  may*  
 1632 *vary across each compressor. For any algorithm  $A \in \mathcal{A}_{\text{sr}}$ , there exists a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$*   
 1633 *such that  $f$  is  $L$ -smooth,  $f(0) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta$ , exists a stochastic gradient oracle that satisfies*  
 1634 *Assumption 1.3, and  $\mathbb{E} \left[ \inf_{y \in S_t} \|\nabla f(y)\|^2 \right] > \varepsilon$  for all*

$$1636 \quad t \leq c_3 \times \frac{L\Delta}{\varepsilon \log(n+1)} \cdot \min \left\{ \max \left\{ \frac{h\sigma^2}{n\varepsilon}, \frac{\tau_w d}{n}, \sqrt{\frac{h\sigma^2\tau_w d}{n\varepsilon}}, h, \tau_w \right\}, \max \left\{ \frac{h\sigma^2}{\varepsilon}, h \right\} \right\},$$

1637 where  $S_t$  is the set of all possible points that can be constructed by  $A$  up to time  $t$  based on  $I$  and  
 1638  $\{I_i\}$ . The quantities  $c_1$ ,  $c_2$ , and  $c_3$  are universal constants.

1639 *Proof.* The proof closely follows the analysis from (Tyurin et al., 2024; Tyurin & Richtárik, 2024)  
 1640 and the proof of Theorem 4.2, but with some important modifications. In this proof, it is sufficient to  
 1641 work with (6) and Lemmas 2.1 and 2.2.

1642 Let us fix  $\lambda > 0$  and define the function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that

$$1643 \quad f(x) := \frac{L\lambda^2}{\ell_1} F_T \left( \frac{x_{[T]}}{\lambda} \right),$$

1644 where the function  $F_T$  is given in (6) and  $x_{[T]} \in \mathbb{R}^T$  is the vector with the first  $T$  coordinates of  
 1645  $x \in \mathbb{R}^d$ . Notice that the last  $d - T$  coordinates are artificial.

1646 First, we have to show that  $f$  is  $L$ -smooth and  $f(0) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta$ . Using Lemma 2.2,

$$1647 \quad \begin{aligned} \|\nabla f(x) - \nabla f(y)\| &= \frac{L\lambda}{\ell_1} \left\| \nabla F_T \left( \frac{x_{[T]}}{\lambda} \right) - \nabla F_T \left( \frac{y_{[T]}}{\lambda} \right) \right\| \leq L\lambda \left\| \frac{x_{[T]}}{\lambda} - \frac{y_{[T]}}{\lambda} \right\| \\ 1648 \quad &= L \|x_{[T]} - y_{[T]}\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^d. \end{aligned}$$

1649 Taking

$$1650 \quad T = \left\lceil \frac{\Delta \ell_1}{L \lambda^2 \Delta^0} \right\rceil,$$

$$1651 \quad f(0) - \inf_{x \in \mathbb{R}^d} f(x) = \frac{L\lambda^2}{\ell_1} (F_T(0) - \inf_{x \in \mathbb{R}^T} F_T(x)) \leq \frac{L\lambda^2 \Delta^0 T}{\ell_1} \leq \Delta.$$

1652 due to Lemma 2.2.

1653 Next, we construct a stochastic gradient mapping. For our lower bound, we define

$$1654 \quad [\nabla f(x; \xi)]_j := \nabla_j f(x) \left( 1 + \mathbb{1}[j > \text{prog}(x)] \left( \frac{\xi}{p_\sigma} - 1 \right) \right) \quad \forall x \in \mathbb{R}^d, \quad (42)$$

1655 and let  $\mathcal{D}_\xi = \text{Bernoulli}(p_\sigma)$  for all  $j \in [n]$ , where  $p_\sigma \in (0, 1]$ . We denote  $[x]_j$  as the  $j^{\text{th}}$  coordinate  
 1656 of a vector  $x \in \mathbb{R}^d$ . We choose

$$1657 \quad p_\sigma := \min \left\{ \frac{L^2 \lambda^2 \gamma_\infty^2}{\sigma^2 \ell_1^2}, 1 \right\}.$$

1674 Then this mapping is unbiased and  $\sigma^2$ -variance-bounded. Indeed,  
 1675

$$1676 \mathbb{E}[[\nabla f(x, \xi)]_i] = \nabla_i f(x) \left( 1 + \mathbb{1}[i > \text{prog}(x)] \left( \frac{\mathbb{E}[\xi]}{p_\sigma} - 1 \right) \right) = \nabla_i f(x)$$

1677 for all  $i \in [d]$ , and  
 1678

$$1680 \mathbb{E} \left[ \|\nabla f(x; \xi) - \nabla f(x)\|^2 \right] \leq \max_{j \in [d]} |\nabla_j f(x)|^2 \mathbb{E} \left[ \left( \frac{\xi}{p_\sigma} - 1 \right)^2 \right]$$

1682 because the difference is non-zero only in one coordinate. Thus  
 1683

$$1684 \mathbb{E} \left[ \|\nabla f(x, \xi) - \nabla f(x)\|^2 \right] \leq \frac{\|\nabla f(x)\|_\infty^2 (1 - p_\sigma)}{p_\sigma} = \frac{L^2 \lambda^2 \|F_T\left(\frac{x_{[T]}}{\lambda}\right)\|_\infty^2 (1 - p_\sigma)}{\ell_1^2 p_\sigma}$$

$$1685 \leq \frac{L^2 \lambda^2 \gamma_\infty^2 (1 - p_\sigma)}{\ell_1^2 p_\sigma} \leq \sigma^2,$$

1687 where we use Lemma 2.2.  
 1688

1689 Taking  
 1690

$$1691 \lambda = \frac{\sqrt{2\varepsilon} \ell_1}{L},$$

1692 we ensure that  
 1693

$$1694 \|\nabla f(x)\|^2 = \frac{L^2 \lambda^2}{\ell_1^2} \left\| \nabla F_T \left( \frac{x_{[T]}}{\lambda} \right) \right\|^2 > 2\varepsilon \mathbb{1}[\text{prog}(x_{[T]}) < T] \quad (43)$$

1695 for all  $x \in \mathbb{R}^d$ , where we use Lemma 2.1. Thus  
 1696

$$1697 T = \left\lfloor \frac{\Delta L}{2\varepsilon \ell_1 \Delta^0} \right\rfloor \quad (44)$$

1698 and  
 1699

$$1700 p_\sigma = \min \left\{ \frac{2\varepsilon \gamma_\infty^2}{\sigma^2}, 1 \right\}.$$

1701 Using the same reasoning as in Tyurin et al. (2024) and our Theorem 4.2, we define two sets of random  
 1702 variables. Let  $\eta_{1,i}$  be the first computed stochastic gradient when the oracle draws a “successful”  
 1703 Bernoulli trial in (42) at worker  $i$ . Then,  
 1704

$$1705 \mathbb{P}(\eta_{1,i} \leq t) \leq \sum_{i=1}^{\lfloor t \rfloor} (1 - p_\sigma)^{i-1} p_\sigma \leq p_\sigma \lfloor t \rfloor$$

1706 for  $t \geq 0$ , and  
 1707

$$1708 \mathbb{P}(\eta_{1,i} \leq t) \leq \min\{p_\sigma \lfloor t \rfloor, 1\}$$

1709 For all  $i \in [n]$ , the server receives a stream of coordinates from worker  $i$ . Let  $\mu_{1,i}$  be the number of  
 1710 received coordinates by the server from worker  $i$  until the moment when the index of the last received  
 1711 coordinate is 1. Let us define  
 1712

$$1713 p_d := \frac{2}{d}.$$

1714 Similarly to the proof of Theorem 4.2 with  $K = 1$  (see (29)),  
 1715

$$1716 \mathbb{P}(\mu_{1,i} \leq t | \eta_{1,i}) = \sum_{j=1}^{\lfloor t \rfloor} \mathbb{P}(\mu_{1,i,1} = j) \leq \frac{\lfloor t \rfloor}{d - t + 1}.$$

1717 for all  $t \leq d$ . Thus,  
 1718

$$1719 \mathbb{P}(\mu_{1,i} \leq t | \eta_{1,i}) \leq \begin{cases} \frac{\lfloor t \rfloor}{d - t + 1}, & t \leq \frac{d}{2} \\ 1, & t > \frac{d}{2} \end{cases} \leq \begin{cases} \lfloor t \rfloor p_d, & t \leq \frac{d}{2} \\ 1, & t > \frac{d}{2} \end{cases} \leq \min\{2 \lfloor t \rfloor p_d, 1\}$$

1728 for all  $t \geq 0$ .  
 1729

1730 There are two ways in which worker  $i$  can discover the first coordinate. Either the worker is “lucky”  
 1731 and draws a successful Bernoulli random variable locally, or it gets to discover the first coordinate  
 1732 through the server. Thus, worker  $i$  requires at least

$$1733 y_{1,i} := \min \left\{ h\eta_{1,i}, \min_{j \in [n], j \neq i} \{h\eta_{1,j} + \tau_w \mu_{1,j}\} \right\} = \min_{j \in [n]} \{h\eta_{1,j} + \mathbb{1}[i \neq j] \tau_w \mu_{1,j}\}$$

1735 seconds because  $h\eta_{1,i}$  is the minimal time to discover the first coordinate locally, and  
 1736  $\min_{j \in [n], j \neq i} \{h\eta_{1,j} + \tau_w \mu_{1,j}\}$  is the minimal time to discover the first coordinate from other workers  
 1737 via the server, which can transmit it to worker  $i$ .

1738 *Remark F.2.* The previous derivations hold for all  $\tau_s > 0$ . If we start taking the communication time  
 1739  $\tau_s$  into account, then  $y_{1,i}$  may only increase. For all  $\tau_s > 0$ , worker  $i$  still requires at least  $y_{1,i}$  sec-  
 1740 onds for the same reason that  $h\eta_{1,i}$  is the minimal time to discover the first coordinate locally, and  
 1741  $\min_{j \in [n], j \neq i} \{h\eta_{1,j} + \tau_w \mu_{1,j}\}$  is the minimal time to discover the first coordinate from other workers. If  
 1742 we start taking into account the communication time from  $\tau_s$ , the lower bound

$$1743 \min \left\{ h\eta_{1,i}, \min_{j \in [n], j \neq i} \{h\eta_{1,j} + \tau_w \mu_{1,j}\} \right\}$$

1744 still holds.

1745 Using the same reasoning, worker  $i$  requires at least

$$1746 y_{k,i} := \min_{j \in [n]} \{h\eta_{k,j} + \mathbb{1}[i \neq j] \tau_w \mu_{k,j} + y_{k-1,j}\}$$

1747 seconds to discover the  $k^{\text{th}}$  coordinate for all  $k \geq 2$ , where

$$1748 \mathbb{P}(\eta_{k,i} \leq t | \mathcal{G}_{k-1}) \leq \min\{\lfloor t \rfloor p_\sigma, 1\} \quad (45)$$

1749 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , and

$$1750 \mathbb{P}(\mu_{k,i} \leq t | \eta_{k,i}, \mathcal{G}_{k-1}) \leq \min\{2 \lfloor t \rfloor p_d, 1\} \quad (46)$$

1751 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , where  $\mathcal{G}_{k-1}$  is the sigma-algebra generated by  $\{\eta_{k',i}\}_{i \in [n], k' < k}$  and  
 1752  $\{\mu_{k',i}\}_{i \in [n], k' < k}$ . Thus, the first possible time when the workers and the server can discover the  $T^{\text{th}}$   
 1753 coordinate is

$$1754 y_T := \min_{i \in [n]} y_{T,i}.$$

1755 For this random variable, we prove the lemma below (see Section F.1).

1756 **Lemma F.3.** Let  $\{\eta_{k,i}\}_{i,k \geq 0}$  and  $\{\mu_{k,i}\}_{i,k \geq 0}$  be random variables such that

$$1757 \mathbb{P}(\eta_{k,i} \leq t | \mathcal{G}_{k-1}) \leq \min\{\lfloor t \rfloor p_\sigma, 1\} \quad (47)$$

1758 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , and

$$1759 \mathbb{P}(\mu_{k,i} \leq t | \eta_{k,i}, \mathcal{G}_{k-1}) \leq \min\{2 \lfloor t \rfloor p_d, 1\}, \quad (48)$$

1760 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , where  $\mathcal{G}_{k-1}$  is the sigma-algebra generated by  $\{\eta_{k',i}\}_{i \in [n], k' < k}$  and  
 1761  $\{\mu_{k',i}\}_{i \in [n], k' < k}$ . Then

$$1762 \mathbb{P}(y_T \leq \bar{t}) \leq \delta$$

1763 with

$$1764 \bar{t} := \frac{T - \log n + \log \delta}{32 \log(8n)} \cdot \min \left\{ \max \left\{ \frac{h}{p_\sigma n}, \frac{\tau_w}{p_d n}, \frac{\sqrt{h\tau_w}}{\sqrt{p_\sigma p_d n}}, h, \tau_w \right\}, \frac{h}{p_\sigma} \right\}, \quad (49)$$

1765 where

$$1766 y_T := \min_{i \in [n]} y_{T,i},$$

$$1767 y_{k,i} := \min_{j \in [n]} \{h\eta_{k,j} + \mathbb{1}[i \neq j] \tau_w \mu_{k,j} + y_{k-1,j}\}$$

1768 for all  $k \geq 1$ ,  $i \in [n]$  and  $y_{0,i} = 0$  for all  $i \in [n]$ .

1782 Thus, with probability at least  $1 - \delta$ , any zero-respecting algorithm requires at least  $\bar{t}$  seconds to  
 1783 discover the last coordinate. Due to (43),  
 1784

$$1785 \inf_{y \in S_t} \|\nabla f(y)\|^2 > 2\varepsilon \inf_{y \in S_t} \mathbb{1} [\text{prog}(y_{[T]}) < T],$$

1787 where  $S_t$  is the set of all possible candidate points to be an  $\varepsilon$ -stationary point up to time  $t$ , which can  
 1788 be computed by  $A$ . Taking  $\delta = \frac{1}{2}$ ,  
 1789

$$1790 \mathbb{E} \left[ \inf_{y \in S_t} \|\nabla f(y)\|^2 \right] > 2\varepsilon \mathbb{E} \left[ \inf_{y \in S_t} \mathbb{1} [\text{prog}(y_{[T]}) < T] \right] \geq \varepsilon,$$

1792 for  $t = \frac{1}{2}\bar{t}$  because  $\text{prog}(y_{[T]}) < T$  for all  $y \in S_t$  with probability at least  $\frac{1}{2}$ . It is left to substitute all  
 1793 quantities:  
 1794

$$1795 t = \frac{\left\lfloor \frac{\Delta L}{2\varepsilon\ell_1\Delta^0} \right\rfloor - \log n + \log \frac{1}{2}}{64\log(8n)} \cdot \min \left\{ \max \left\{ \frac{h}{n} \max \left\{ \frac{\sigma^2}{2\varepsilon\gamma_\infty^2}, 1 \right\}, \frac{\tau_w d}{2n}, \frac{\sqrt{hd\tau_w}}{\sqrt{2n}} \sqrt{\frac{\sigma^2}{2\varepsilon\gamma_\infty^2}}, h, \tau_w \right\}, h \max \left\{ \frac{\sigma^2}{2\varepsilon\gamma_\infty^2}, 1 \right\} \right\}.$$

1798 Since  $\ell_1, \Delta^0, \gamma_\infty$  are universal constants, assuming  $\varepsilon < c_1 L \Delta$  for some small universal  $c_1 > 0$ , we  
 1799 get  
 1800

$$1801 t \geq c_3 \times \frac{L\Delta}{\varepsilon \log(n+1)} \cdot \min \left\{ \max \left\{ \frac{h\sigma^2}{n\varepsilon}, \frac{\tau_w d}{n}, \sqrt{\frac{h\sigma^2\tau_w d}{n\varepsilon}}, h, \tau_w \right\}, \max \left\{ \frac{h\sigma^2}{\varepsilon}, h \right\} \right\}$$

1803 for some small universal  $c_3 > 0$ . Notice that we can take any dimension  $d$  such that  
 1804

$$1805 d \geq T = \Theta \left( \frac{L\Delta}{\varepsilon} \right).$$

□

## 1809 F.1 MAIN CONCENTRATION LEMMA

1811 **Lemma F.3.** *Let  $\{\eta_{k,i}\}_{i,k \geq 0}$  and  $\{\mu_{k,i}\}_{i,k \geq 0}$  be random variables such that*

$$1813 \mathbb{P}(\eta_{k,i} \leq t | \mathcal{G}_{k-1}) \leq \min\{\lfloor t \rfloor p_\sigma, 1\} \quad (47)$$

1814 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , and  
 1815

$$1816 \mathbb{P}(\mu_{k,i} \leq t | \eta_{k,i}, \mathcal{G}_{k-1}) \leq \min\{2 \lfloor t \rfloor p_d, 1\}, \quad (48)$$

1817 for all  $k \geq 1$ ,  $i \in [n]$ , and  $t \geq 0$ , where  $\mathcal{G}_{k-1}$  is the sigma-algebra generated by  $\{\eta_{k',i}\}_{i \in [n], k' < k}$   
 1818 and  $\{\mu_{k',i}\}_{i \in [n], k' < k}$ . Then  
 1819

$$1820 \mathbb{P}(y_T \leq \bar{t}) \leq \delta$$

1821 with  
 1822

$$1823 \bar{t} := \frac{T - \log n + \log \delta}{32\log(8n)} \cdot \min \left\{ \max \left\{ \frac{h}{p_\sigma n}, \frac{\tau_w}{p_d n}, \frac{\sqrt{h\tau_w}}{\sqrt{p_\sigma p_d n}}, h, \tau_w \right\}, \frac{h}{p_\sigma} \right\}, \quad (49)$$

1825 where  
 1826

$$1827 y_T := \min_{i \in [n]} y_{T,i},$$

$$1829 y_{k,i} := \min_{j \in [n]} \{h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j} + y_{k-1,j}\}$$

1830 for all  $k \geq 1$ ,  $i \in [n]$  and  $y_{0,i} = 0$  for all  $i \in [n]$ .  
 1831

1832 *Proof.* Using the Chernoff method for any  $s > 0$  and  $k \geq 1$ , we get  
 1833

$$1835 \mathbb{P}(y_k \leq \bar{t}) = \mathbb{P}(-sy_k \geq -s\bar{t}) = \mathbb{P}\left(e^{-sy_k} \geq e^{-s\bar{t}}\right) \leq e^{s\bar{t}} \mathbb{E}[e^{-sy_k}]$$

$$1836 = e^{s\bar{t}} \mathbb{E} \left[ \exp \left( -s \min_{j \in [n]} y_{k,j} \right) \right] = e^{s\bar{t}} \mathbb{E} \left[ \max_{j \in [n]} \exp(-s y_{k,j}) \right].$$

1838 Bounding the maximum by the sum,

$$1840 \mathbb{P}(y_k \leq t) \leq e^{st} \sum_{j=1}^n \mathbb{E} [\exp(-s y_{k,j})] \leq n e^{st} \max_{j \in [n]} \mathbb{E} [\exp(-s y_{k,j})]. \quad (50)$$

1842 We focus on the last exponent separately. For all  $i \in [n]$ ,

$$\begin{aligned} 1845 \mathbb{E} [\exp(-s y_{k,i})] &= \mathbb{E} \left[ \exp \left( -s \min_{j \in [n]} \{h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j} + y_{k-1,j}\} \right) \right] \\ 1846 &= \mathbb{E} \left[ \max_{j \in [n]} \exp(-s(h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j} + y_{k-1,j})) \right] \\ 1847 &\leq \sum_{j=1}^n \mathbb{E} [\exp(-s(h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j} + y_{k-1,j}))] \\ 1849 &= \sum_{j=1}^n \mathbb{E} \left[ \underbrace{\mathbb{E} [\exp(-s(h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j})) | \mathcal{G}_{k-1}]}_{I_1 :=} \exp(-s y_{k-1,j}) \right] \\ 1852 & \end{aligned} \quad (51)$$

1856 Considering the inner expectation separately:

$$\begin{aligned} 1857 I_1 &= \mathbb{E} [\exp(-s(h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j})) | \mathcal{G}_{k-1}] \\ 1858 &\leq \exp(-st) + \mathbb{P}(h\eta_{k,j} + \mathbb{1}[i \neq j]\tau_w \mu_{k,j} \leq t | \mathcal{G}_{k-1}) \end{aligned}$$

1860 for all  $t \geq 0$ . Using the properties of condition expectations,

$$\begin{aligned} 1861 I_1 &\leq \exp(-st) + \mathbb{P}(h\eta_{k,j} \leq t, \mathbb{1}[i \neq j]\tau_w \mu_{k,j} \leq t | \mathcal{G}_{k-1}) \\ 1862 &= \exp(-st) + \mathbb{E} [\mathbb{1}[h\eta_{k,j} \leq t] \mathbb{1}[\mathbb{1}[i \neq j]\tau_w \mu_{k,j} \leq t] | \mathcal{G}_{k-1}] \\ 1863 &= \exp(-st) + \mathbb{E} [\mathbb{E} [\mathbb{1}[\mathbb{1}[i \neq j]\tau_w \mu_{k,j} \leq t] | \eta_{k,j}, \mathcal{G}_{k-1}] \mathbb{1}[h\eta_{k,j} \leq t] | \mathcal{G}_{k-1}] \\ 1864 &= \exp(-st) + \mathbb{E} [\mathbb{P}(\mathbb{1}[i \neq j]\tau_w \mu_{k,j} \leq t | \eta_{k,j}, \mathcal{G}_{k-1}) \mathbb{1}[h\eta_{k,j} \leq t] | \mathcal{G}_{k-1}]. \end{aligned}$$

1866 If  $i = j$ , then we bound the probability by 1 and get

$$\begin{aligned} 1868 I_1 &\leq \exp(-st) + \mathbb{E} [\mathbb{1}[h\eta_{k,j} \leq t] | \mathcal{G}_{k-1}] \\ 1869 &= \exp(-st) + \mathbb{P}(h\eta_{k,j} \leq t | \mathcal{G}_{k-1}) \\ 1870 &\leq \exp(-st) + \left\lfloor \frac{t}{h} \right\rfloor p_\sigma, \end{aligned}$$

1872 for all  $t \geq 0$ , where we use (47). Otherwise, if  $i \neq j$ , using (48) and (47),

$$\begin{aligned} 1874 I_1 &\leq \exp(-st) + \mathbb{E} \left[ \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \mathbb{1}[h\eta_{k,j} \leq t] \middle| \mathcal{G}_{k-1} \right] \\ 1875 &= \exp(-st) + \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \mathbb{P}(h\eta_{k,j} \leq t | \mathcal{G}_{k-1}) \\ 1876 &\leq \exp(-st) + \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \end{aligned}$$

1881 for all  $t \geq 0$ . Substituting the inequalities to (51),

$$\begin{aligned} 1883 \mathbb{E} [\exp(-s y_{k,i})] &= \sum_{j \neq i} \left( \exp(-st) + \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right) \mathbb{E} [\exp(-s y_{k-1,j})] \\ 1884 &+ \left( \exp(-st) + \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right) \mathbb{E} [\exp(-s y_{k-1,i})]. \end{aligned}$$

1888 for all  $i \in [n]$  and  $t \geq 0$ . Thus,

$$1889 \max_{i \in [n]} \mathbb{E} [\exp(-s y_{k,i})]$$

$$\begin{aligned}
&\leq \left[ (n-1) \left( \exp(-st) + \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right) + \left( \exp(-st) + \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right) \right] \\
&\quad \times \max_{i \in [n]} \mathbb{E} [\exp(-sy_{k-1,i})] \\
&= \left[ n \exp(-st) + (n-1) \left( \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right) + \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right] \max_{i \in [n]} \mathbb{E} [\exp(-sy_{k-1,i})].
\end{aligned}$$

Taking  $s = \frac{\log(8n)}{t}$ , we get

$$\max_{i \in [n]} \mathbb{E} [\exp(-sy_{k,i})] \leq \left[ \frac{1}{8} + \underbrace{(n-1) \left( \min \left\{ 1, 2 \left\lfloor \frac{t}{\tau_w} \right\rfloor p_d \right\} \min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\} \right)}_{I_2 :=} + \underbrace{\min \left\{ 1, \left\lfloor \frac{t}{h} \right\rfloor p_\sigma \right\}}_{I_3 :=} \right] \max_{i \in [n]} \mathbb{E} [\exp(-sy_{k-1,i})]. \quad (52)$$

Next, we take  $t = \min\{t_1, t_2\}$ , where

$$t_1 := \max \left\{ \frac{h}{32p_\sigma n}, \frac{\tau_w}{32p_d n}, \frac{\sqrt{h\tau_w}}{32\sqrt{p_\sigma p_d n}}, \frac{h}{32}, \frac{\tau_w}{32} \right\},$$

and

$$t_2 := \frac{h}{32p_\sigma}$$

to ensure that

$$I_3 \leq \min \left\{ 1, \frac{t_2 p_\sigma}{h} \right\} \leq \frac{1}{16}.$$

There are five possible values of  $t_1$ .

If  $t_1 = \frac{h}{32p_\sigma n}$ , then

$$I_2 \leq (n-1) \min \left\{ 1, \frac{t_1 p_\sigma}{h} \right\} \leq \frac{1}{16},$$

If  $t_1 = \frac{\tau_w}{32p_d n}$ , then

$$I_2 \leq (n-1) \min \left\{ 1, \frac{2t_1 p_d}{\tau_w} \right\} \leq \frac{1}{16}.$$

If  $t_1 = \frac{\sqrt{h\tau_w}}{32\sqrt{p_\sigma p_d n}}$ , then

$$I_2 \leq (n-1) \frac{2t_1^2 p_d p_\sigma}{\tau_w h} \leq \frac{1}{16}.$$

If  $t_1 = \frac{h}{32}$ , then

$$I_2 \leq (n-1) \min \left\{ 1, \left\lfloor \frac{t_1}{h} \right\rfloor p_\sigma \right\} = 0.$$

Finally, if  $t_1 = \frac{\tau_w}{32}$ , then

$$I_2 \leq (n-1) \min \left\{ 1, 2 \left\lfloor \frac{t_1}{\tau_w} \right\rfloor p_d \right\} = 0.$$

Thus, using (52), we obtain

$$\max_{i \in [n]} \mathbb{E} [\exp(-sy_{k,i})] \leq \left[ \frac{1}{8} + \frac{1}{16} + \frac{1}{16} \right] \max_{i \in [n]} \mathbb{E} [\exp(-sy_{k-1,i})] \leq e^{-1} \max_{i \in [n]} \mathbb{E} [\exp(-sy_{k-1,i})]$$

for our choice of  $t$ . Unrolling the recursion and using  $y_{0,i} = 0$  for all  $i \in [n]$ ,

$$\max_{i \in [n]} \mathbb{E} [\exp(-sy_{k,i})] \leq e^{-k}.$$

We substitute it to (50), to get

$$\mathbb{P} (y_k \leq \bar{t}) \leq e^{s\bar{t} + \log n - k}.$$

It is left to choose  $\bar{t} = \frac{k - \log n + \log \delta}{s}$ .  $\square$