Distributed Optimization for Overparameterized Problems: Achieving Optimal Dimension Independent Communication Complexity

Bingqing Song
Department of ECE
University of Minnesota
email:song0409@umn.edu

Ioannis Tsaknakis
Department of ECE
University of Minnesota
email:tsakn001@umn.edu

Chung-Yiu Yau
Department of SEEM
Chinese University of Hong Kong
email:cyyau@se.cuhk.edu.hk

Hoi-To Wai
Department of SEEM
Chinese University of Hong Kong
email:htwai@cuhk.edu.hk

Mingyi Hong
Department of ECE
University of Minnesota
email:mhong@umn.edu

Abstract

Decentralized optimization are playing an important role in applications such as training large machine learning models, among others. Despite its superior practical performance, there has been some lack of fundamental understanding about its theoretical properties. In this work, we address the following open research question: To train an overparameterized model over a set of distributed nodes, what is the *minimum* communication overhead (in terms of the bits got exchanged) that the system needs to sustain, while still achieving (near) zero training loss? We show that for a class of overparameterized models where the number of parameters D is much larger than the total data samples N, the best possible communication complexity is $\Omega(N)$, which is independent of the problem dimension D. Further, for a few specific overparameterized models (i.e., the linear regression, and certain multi-layer neural network with one wide layer), we develop a set of algorithms which uses certain linear compression followed by adaptive quantization, and show that they achieve dimension independent, near-optimal communication complexity. To our knowledge, this is the first time that dimension independent communication complexity has been shown for distributed optimization.

1 Introduction

The research of decentralized/distributed optimization has recently gained tremendous momentum, partly due to the fact that a well-designed decentralized/distributed algorithm is capble of training large-scale machine learning models, by utilizing massively distributed computation and data resources. In a typical decentralized optimization setting, multiple local agents hold small to moderately sized datasets, and they collaborate by iteratively solving their local problems while sharing some information with other agents. The majority of the state-of-the-art distributed/decentralized optimization algorithms are deeply rooted in classical consensus-based approaches [Tsitsiklis, 1984, Tsitsiklis and Luo, 1987], where the agents repetitively share the local optimization variables, or the gradients of local functions, so to reach an optimal *consensual* solution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

It is important to note that in those classical settings, the agents typically solve *small-scale* problems where the problem dimension (i.e., the dimension of the optimization variable, and the gradients) is relatively small (e.g., in the order of hundreds or thousands). Therefore, it is reasonable that the agents should iteratively share their local parameters/gradients to improve the quality of the models. However, many modern learning models are in the *overparameterized* regime with very high-dimensional parameters [He et al., 2016, Vaswani et al., 2017], such as deep and/or wide neural networks. This practice poses some significant challenges to the classical consensus-based approaches, because by directly sharing the model parameters and/or gradients, the communication burden increases as problem size increases.

A line of recent works have been developed, which adopt strategies such as gradient/model compression, lazy communication, etc, and they are able to significantly reduce the overall communication complexities [Xu et al., 2020, Sun et al., 2020, Li et al., 2019, Karimireddy et al., 2019]. A few works focus on designing compressors [Safaryan et al., 2020, Horváth et al., 2019, Beznosikov et al., 2020, Konečnỳ et al., 2016, Safaryan et al., 2021]. Although the majority of these works are able to reduce *per round* communication burden, it is by no means clear if, to achieve certain solution quality, the *total* amount of communication can be saved. In fact, it has been shown in [Arjevani and Shamir, 2015], that the total communication achievable is dependent on the problem dimension, which could be huge. More importantly, there is a general lack of fundamental understanding about the performance limits for decentralized and distributed optimization, when the models are 'large'. Without such characterization, it is by no means clear if the existing algorithms are *optimal* in reducing the communication complexities.

In this work, we focus on distributed optimization for 'large' machine learning models (i.e., overparameterized problems, to be defined shortly), and we ask the following question:

(Q) What is the *best possible* communication complexity (in terms of number of bits needed to obtain certain solution accuracy), achievable by *any* distributed algorithms for a class of overparameterized problems? Which class of algorithms achieves such complexity bounds?

Related Works. A number of recent works are devoted to reducing the communication cost in distributed/decentralized optimization by combining the optimization algorithms with *compression*, e.g., sparsification, quantization, etc. For the distributed setting with the server-worker architecture, relevant works include [Alistarh et al., 2017, Wen et al., 2017, Wang et al., 2018, Bernstein et al., 2018, Haddadpour et al., 2021, Mishchenko et al., 2019, Gorbunov et al., 2021]; see [Richtárik et al., 2021] for a recent survey. For the decentralized setting without a central server, recent works such as [Koloskova et al., 2019, Reisizadeh et al., 2019, Tang et al., 2018, Vogels et al., 2020, Magnússon et al., 2020] proposed algorithms that utilize error compensation, low-rank compression, etc. While they have demonstrated appealing performance in practice, the theoretical guarantees do not scale for overparameterized problems. For instance, [Koloskova et al., 2019] showed an $\mathcal{O}(1/T + 1/(\omega T)^2)$ iteration complexity upper bound for their algorithm. The quantity $1/\omega$ is the compressor parameter which is at least $\Omega(\sqrt{D})$, where D is the problem dimension.

Lower communication complexity bounds for distributed optimization problems have been studied since the 1980s. There are two major categories of lower bound results in literature depending on how we count the "communication" required to reach a solution of given accuracy, i.e., 1) in terms of the number of communication rounds, or 2) in terms of the number of bits got communicated. The first category is not directly related to the current work; we refer to readers to Appendix A for discussions. In second category, the first lower bound is given in [Tsitsiklis and Luo, 1987] where the authors consider a distributed setting with two nodes and provide a $\Omega\left(D \log\left(D/\epsilon\right)\right)$ bound for the class of quadratic objectives, where ϵ is the solution accuracy, and D is the variable dimension. In [Zhang et al., 2013] lower communication bounds are derived for certain statistical estimation tasks, such as linear and probit regression. Moreover, in [Vempala et al., 2020] the communication complexity of several distributed optimization tasks is studied, including ℓ_p regression, linear, semidefinite and convex optimization. In [Korhonen and Alistarh, 2021], an $\Omega\left(KD \log\left(D/\epsilon\right)\right)$ bound is derived for quadratic problems on a multi-node setting, for both deterministic and randomized protocols, where K is the number of clients.

Recently, overparameterized problem has drawn great research interests. It has been showed that remarkable performance can be achieved with overparameterized networks [Jacot et al., 2018, Du et al., 2019, Allen-Zhu et al., 2019]. Some recent works study the overparameterization of decentralized problems, which guarantee the global convergence [Huang et al., 2021, Deng et al., 2022]. However,

the high communication cost caused by the large model is a key problem in decentralized algorithm. For example, the ReLU fully connected networks designed in the aforementioned two works are of widths $\Omega(N^4)$ and $\Omega(N^{18})$ respectively, where N is the number of data samples available.

Contributions. This work addresses both questions in **(Q)**. The contributions are highlighted below:

(i) Regarding to the *best possible* communication complexity (i.e., the communication complexity lower bound), we first show that for a class of Polyak-Lojasiewicz (PL) problems, the communication lower bound is $\Omega(KD\log(D/\epsilon))$, which is similar to the known results for strongly convex problems. Further, we show that for a class of overparameterized models where the loss function is a composition of a quadratic problem and some nonlinear functions, and that the number of parameters D is much larger than the total data samples N, the communication complexity lower bound becomes $\Omega(KN\log(K/\epsilon))$, which is *independent* of the problem dimension D.

(ii) We develop a distributed algorithm which can be used to optimize both classes of problems, in the setting where multiple clients are coordinated by a server. We show that the lower bound for the PL problem can be matched; while by using certain linear random compression/decompression scheme, followed by adaptive quantization, the lower bound for the overparameterized problem can be matched (up to a $\log(D)$ factor) for two of its important special cases, i.e., linear regression, and certain multi-layer neural network with one wide layer. Further, we show that the proposed algorithm can be extended to a fully decentralized setting without the coordination of a server.

To our knowledge, these are the first set of results that attempt to comprehensively address the issue of optimal communication complexities for overparameterized problems. We expect that results provided in this work will open doors for understanding more complex communication-efficient distributed algorithms, as well as larger classes of overparameterized problems.

2 The System Model

Consider a system having K distributed agents, and a total of N samples $\{(X_i, y_i)\}_{i=1}^N$. Suppose each agent has N_k data samples, with $\sum_{k=1}^K N_k = N$; let set $\mathcal{X}_k$ (resp. $\mathcal{Y}_k$) collect agent k's data samples (resp. labels). They cooperatively solve an optimization problem of the following form:

$$\min_{\theta\in\Theta}\ f(\theta;\mathcal{X},\mathcal{Y}) := \sum_{k=1}^K f_k(\theta;\mathcal{X}_k,\mathcal{Y}_k) \tag{1}$$

where $f_k:\Theta\to\mathbb{R}$ is the local loss function for agent k; θ is the optimization variable, whose domain is $\Theta\subseteq\mathbb{R}^D$. For notational simplicity, we will ignore the reference for the data samples $\mathcal{X},\mathcal{Y}$ and $\mathcal{X}_k,\mathcal{Y}_k$ whenever possible, i.e., we will write $f(\theta)$ and $f_k(\theta)$ as the global and local loss functions.

Throughout the paper, we will consider a few special cases of (1), where the local function f_k's are *overparameterized*. The first of such problem class is referred to as the *distributed PL problems*. To introduce this class, let us provide a few assumptions.

Assumption 1. *The local objective functions satisfy:*

$$2\mu_k\cdot(f_k(\theta)-f_k(\theta^*_{(k)}))\le\|\nabla f_k(\theta)\|^2,\quad\forall\,\theta,\quad\forall\,k, \tag{2}$$

where $\theta^*_{(k)}$ is a global minimum of $f_k(\cdot)$; μ_k's some positive constants.

Assumption 2. *There exists positive constants L_k's and $L\ge 1$ such that:*

$$\|\nabla f_k(\theta)-\nabla f_k(\theta')\|\le L_k\|\theta-\theta'\|,\quad\|\nabla f(\theta)-\nabla f(\theta')\|\le L\|\theta-\theta'\|,\quad\forall\,\theta,\theta',\ \forall\,k.$$

Note that (2) is the so-called PL assumption, which was introduced in [Polyak, 1963]. It is weaker than strong convexity, while it can still be used to ensure the linear convergence of gradient descent. It is also less restrictive than several other related conditions under which linear global convergence can be derived, such as the restricted secant inequality [Karimi et al., 2016]. The second inequality in (2) has also been used in the literature; see, e.g., [Oymak and Soltanolkotabi, 2019], in which such a bound for overparameterized network has been showed. Assumption 2 is the standard Lipschitz gradient condition.

Definition 1 (Distributed PL problems)**.** *Consider problems of the form* (1)*. Define:*

$$\mathcal{C}_{\rm pl} := \left\{ f(\theta)=\sum_{k=1}^K f_k(\theta) : f_k(\theta)\ \textit{satisfies Assumptions 1 – 2},\ \forall\,k\in[K]\right\}.$$

The second class of problems is referred to as the *distributed overparameterized* problem, where the loss function is a composition of least square loss and a non-linear function. This form of loss function often appears in analyzing neural networks, see, e.g., [Oymak and Soltanolkotabi, 2019].

Definition 2 (Distributed overparameterized problems)**.** *Consider problems of the form* (1)*. Define functions $G_k(\cdot) : \Theta \to \mathbb{R}^{N_k}$ and constants $b_k \in \mathbb{R}^{N_k}$, for all k. We call $G_k(\cdot)$ overparameterized if for every $b_k \in \mathbb{R}^{N_k}$, there exists $\theta \in \Theta$ such that $G_k(\theta) = b_k$. Define the following class of problems:*

$$\mathcal{C}_{\text{op}} := \left\{ f(\theta) = \sum_{k=1}^{K} f_k(\theta),\ f_k(\theta) = \frac{1}{2}\|G_k(\theta) - b_k\|_2^2 : G_k(\cdot) : \Theta \to \mathbb{R}^{N_k} \text{ overparameterized.} \right\}$$

3 Communication Complexity Lower Bounds for Deterministic Algorithms

In this section we derive the communication complexity lower bounds for classes $\mathcal{C}_{\text{op}}$ and $\mathcal{C}_{\text{pl}}$. Throughout this section and the paper, our focus is given to *deterministic* algorithms where no randomness is involved. All the proofs for results presented in this section are given in Appendix B.

We start with a description of the *coordinator model* (also see [Braverman et al., 2013, Korhonen and Alistarh, 2021]), which will be used later as the basis of our derivation. Consider a set of K distinct nodes/agents, and a single coordinator node. Each node and the coordinator have a two-way communication channel, over which binary messages are exchanged. Consider a *class of problems* $\mathcal{C}$ of the form (1), that is $f(\theta) = \sum_{k=1}^{K} f_k(\theta)$, where each node k has access only to objective $f_k(\theta)$. Moreover, let Π be a set of protocols that are executed over the coordinator model. Then, given a problem class $\mathcal{C}$, a *protocol* $\pi \in \Pi(\epsilon)$ takes as input a problem $c \in \mathcal{C}$, performs a number of message exchanges (over the coordinator model), and outputs an ϵ-approximate solution $\tilde{\theta} \in \Theta$ of problem (1). That is, the output $\tilde{\theta}$ satisfies (where θ^* is a point where the exact minimum is attained):

$$f(\theta^*) \leq f(\tilde{\theta}) \leq f(\theta^*) + \epsilon, \tag{3}$$

Next, let us look at π in more detail. We assume that for $\pi \in \Pi(\epsilon)$, it performs $T(\epsilon)$ rounds of communication and computation, where during each round the following steps are carried out:

- At round $t = 0$, each agent/node k is given as input the local objective f_k (e.g., the data samples $\mathcal{X}_k, \mathcal{Y}_k$ assigned to each node);
- At round $1 \leq t < T(\epsilon)$,
  - The coordinator sends a message $m_{c\to k}^{(t)} \in \{0, 1\}$ to some node k.
  - Node k sends message $m_{k\to c}^{(t)} \in \{0, 1\}$ to the coordinator.
- At the final round $t = T(\epsilon)$ the coordinator outputs an ϵ-approximate solution.

Next, we define the communication complexity of the distributed protocol. Let us denote the total number of bits exchanged under protocol π, for problem $c \in \mathcal{C}$, as $M(c, \pi, \epsilon)$; Use the following definition, which has been used extensively in related works such as [Tsitsiklis and Luo, 1987].

Definition 3 (Communication complexity)**.** *The communication complexity of protocol π over problem class $\mathcal{C}$ is defined as the maximum number of bits exchanged across the problems in $\mathcal{C}$, i.e., $M(\mathcal{C}, \pi, \epsilon) := \max_{c\in\mathcal{C}} M(c, \pi, \epsilon)$. Also, the communication complexity of problem class $\mathcal{C}$ is the minimum communication complexity across all protocols that solve $\mathcal{C}$, i.e.,*

$$M(\mathcal{C}, \epsilon) := \min_{\pi\in\Pi(\epsilon)} M(\mathcal{C}, \pi, \epsilon) = \min_{\pi\in\Pi(\epsilon)} \max_{c\in\mathcal{C}} M(c, \pi, \epsilon).$$

Next, we derive communication complexity lower bounds for certain problem classes $\mathcal{C}$. Our construction strategies follow those from [Korhonen and Alistarh, 2021], that is, we will reduce a problem with known communication complexity, to hard problem instances within classes of our interest. The main difference from the aforementioned work (which only deals with quadratic problems) is that, we need to construct hard PL and overparameterized problems instances.

Below we derive communication complexity lower bounds for the class of deterministic distributed problems where the local loss functions belong to the class of PL functions defined in Def. 1.

Theorem 3.1. *Consider the problem class $\mathcal{C}_{\text{pl}}$ with given parameters K, D, ϵ, for which it holds that $\frac{D}{\epsilon} = \Omega(1)$. Then we have $M(\mathcal{C}_{\text{pl}}, \epsilon) = \Omega(KD \log\left(\frac{D}{\epsilon}\right))$.*

Note that the condition $D/\epsilon = \Omega(1)$ is reasonable since we will mainly focus on the regime where D is large. Next we consider $\mathcal{C}_{\rm op}$. We have the following communication complexity lower bound.

Theorem 3.2. *Consider the problem class $\mathcal{C}_{\rm op}$, with given parameters K, N, D, ϵ, for which it holds that $\frac{N}{\epsilon} = \Omega(1)$. Then we have that $M(\mathcal{C}_{\rm op}, \epsilon) = \Omega(KN \log\left(\frac{N}{\epsilon}\right))$.*

Clearly, the lower bound for class $\mathcal{C}_{\rm op}$ does not explicitly depend on the problem dimension D, while that for the $\mathcal{C}_{\rm pl}$ does. This appears to be reasonable, partly because in the class of $\mathcal{C}_{\rm pl}$ the number of data samples N has not been explicitly modeled. Nevertheless, at this point it is not clear if such a *dimension-independent* lower bound is tight or not, because we are yet to find an algorithm to match such a lower bound. To our knowledge, these are the first set of results that characterize the communication complexity for the classes $\mathcal{C}_{\rm pl}$ and $\mathcal{C}_{\rm op}$, and they will serve to guide our subsequent algorithm development and analysis.

4 Algorithms Achieving the Optimal Communication Complexities

In this section, we design algorithms and analyze the *achievable* communication complexities, for the PL and the overparameterized problems in Def. 1 and 2. We will begin with the simple setting where all the agents are connected to a coordinator, and then extend to the case where the agents are decentralized. Throughout this section, for simplicity of exposition, we will assume that $\Theta \equiv \mathbb{R}^D$.

Let us begin with a generic algorithm to deal with the server-agent setting. The algorithm is an adaptation of the classical quantize-then-communicate strategy [Tsitsiklis and Luo, 1987, Magnússon et al., 2020]. At each iteration, each agent first computes its gradient, followed by certain *adaptive* quantization scheme, before getting transmitted to the collaborating agents.

Specifically, at t-th iteration, each agent uses a function $F : \mathbb{R}^D \to \mathbb{R}^H$ to compress the local gradient; Then the output is quantized as $q_k^t \in \mathbb{R}^H$ (use the method in Def. 4). Each agent then shares q_k^t to the server, and the aggregated $\mathbf{q}^t$ is broadcast to all agents. Each agent then uses a function $\tilde{F} : \mathbb{R}^H \to \mathbb{R}^D$ to process the received vector, and updates the local parameters. One key difference compared with [Tsitsiklis and Luo, 1987, Magnússon et al., 2020] is that, the algorithm uses a pair of *compression and decompression functions* $F(\cdot)$ and $\tilde{F}(\cdot)$ to process the gradients transmitted/received. We will see shortly that such a pair plays an important role in matching the lower bound. For the detailed steps, see Algorithm 1 below.

Algorithm 1 Limited Communication Distributed Optimization Algorithm

Initialize: Fix $F(\cdot)$ and $\tilde{F}(\cdot)$, $\eta > 0$, $\gamma^t \in \mathbb{R}_+$, $b \in \mathbb{Z}_+$; set $\theta_k^0 = \theta^0$, choose q_k^0 which can be represented by b_0 bits for each entry, such that $\|q_k^0 - F(\nabla f_k(\theta^0))\|_\infty \leqslant \frac{\sqrt{f_k(\theta^0)}}{C \cdot D^3 \sqrt{K}}$, where C is some constant independent of D.
for $t = 0, 1, \ldots, T-1$ **do**
  `Update`: Each agent updates $\theta_k^{t+1} = \theta^t - \eta \tilde{F}(\sum_{k=1}^{K} q_k^t)$; set $\theta^{t+1} = \theta_k^{t+1}$.
  `Compute`: The function of gradient of each agent $g_k^{t+1} = F(\nabla f_k(\theta^{t+1}))$;
  `Quantize`: Each agent quantizes $q_k^{t+1} = \text{quant}(g_k^{t+1}, q_k^t, \gamma^t, b)$, defined in Def. 4;
  `Communicate`: Each agent sends q_k^{t+1} to the server;
  `Aggregation`: The server collects $\mathbf{q}^{t+1} = (q_1^{t+1}, q_2^{t+1}, \cdots, q_K^{t+1})$;
  `Broadcast`: The server broadcasts $\mathbf{q}^{t+1} = (q_1^{t+1}, q_2^{t+1}, \cdots, q_K^{t+1})$ to each agent;
end for
Return: Parameters θ^T.

In Algorithm 1, we note that as long as the initial solution θ^0 is the same among all the agents, then θ_k^{t+1} computed in the '`Update`' step will always be the same across the agents. This is the reason that we can directly set θ_k^{t+1} as θ^{t+1}. Subsequently we will extend this algorithm to a *fully decentralized* setting, where precise consensus is no longer required. We choose the quantization method for the first iteration defined in Appendix C. In the subsequent iterations, we use the following adaptive quantization scheme; also see [Magnússon et al., 2020].

Definition 4. *Fix $r \in \mathbb{R}_+$ and let $b \in \mathbb{N}$ be an positive integer. Let* quant*: $\mathbb{R}^H \times \mathbb{R}^H \times \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}^H$ be the quantization function defined component-wise as follows, where $\delta(r,b) := r/\left(2^b - 1\right)$,*

$$[\text{quant}(c,p,r,b)]_j = \begin{cases} p_j - r & \text{if } p_j \leq c_j - r + \delta(r,b) \\ p_j + r & \text{if } p_j \geq c_j + r - \delta(r,b) \\ p_j - r + 2\delta(r,b) \left\lfloor \frac{c_j - p_j + r + \delta(r,b)}{2\delta(r,b)} \right\rfloor & \text{otherwise.} \end{cases}$$

Essentially, the quantization scheme searches for a closest point on the grid (which has a width $2r$ and is centered at $\mathbf{p} = (p_1, \ldots, p_H)$. The distance between the neighbor points on the grid is $\delta(r,b)$.

Lemma 4.1. *Fix $p \in \mathbb{R}^H$, then for all $c \in \mathbb{R}^H, Q \in \mathbb{R}^{D \times H}$, such that $\|c - p\|_\infty \leq r$ we have*

$$\|\text{quant}\,(c,p,r,b) - c\|_\infty \leq \frac{r}{2^b - 1}, \quad \|Q \cdot \text{quant}\,(c,p,r,b) - Q \cdot c\|_\infty \leq \frac{rH\|Q\|_\infty}{2^b - 1}.$$

4.1 Communication Complexity Upper Bounds for $\mathcal{C}_{\text{pl}}$

We will begin with analyzing the class of problem $\mathcal{C}_{\text{pl}}$, as defined in Def. 1. Before we proceed, let us make an additional assumption.

Assumption 3. *Assume that each $f_k(\cdot)$ is non-negative in its domain; Further, assume that the optimal objective value for problem (1) is zero, that is, there exists $\theta^* \in \mathbb{R}^D$ such that $f(\theta^*) = 0$.*

The above 'zero loss' assumption has been used by a number of recent works to characterize problems with a large number of parameters, e.g., in overparameterized neural network [Liu et al., 2022, Nguyen and Mondelli, 2020]; in optimization literature, this assumption is also used to improve the analysis of gradient-based algorithms; see, e.g. [Razaviyayn et al., 2019].

Lemma 4.2. *Suppose Assumption 2 – 3 are satisfied, then we have the following inequality:*

$$\|\nabla f(\theta)\|^2 \leqslant 2L \cdot f(\theta), \quad \forall\, \theta.$$

The above lemma shows that together with the PL condition, Assumptions 2, 3 imply:

$$2\mu \cdot f(\theta) \leq \|\nabla f(\theta)\|^2 \leq 2L \cdot f(\theta), \quad \forall\, \theta.$$

For convenience, throughout our analysis we set $\mu < 1, L > 1$.

We will refer to the class of problems satisfying Assumption 1 to 3 as $\widetilde{\mathcal{C}}_{\text{pl}}$. We note that $\widetilde{\mathcal{C}}_{\text{pl}}$ is a strict subset of the class $\mathcal{C}_{\text{pl}}$. Next, let us present our result for $\mathcal{C}_{\text{pl}}$. The proof can be found in Appendix C.

Theorem 4.1. *Suppose Assumptions 2 – 3 are satisfied. Consider using Alg. 1 to solve $\mathcal{C}_{\text{pl}}$; set*

$$\eta = (2L^{\frac{3}{2}})^{-1}, \tau = (\sqrt{C}D)^{-1}, b = \max\left(\log_2(\sqrt{C}D + 1), b_0\right), \tilde{F}(x) = F(x) = x,$$

and $C := \max\left(\sqrt{16L/\mu^2 + L^{\frac{3}{2}}/\mu}, 100\right)$. Then the following holds true:

$$f\left(\theta^{t+1}\right) \leq \left(1 - \mu/(8L^{\frac{3}{2}})\right) f\left(\theta^t\right), \quad \forall\, t = 0, 1, \cdots.$$

Further, to compute an ϵ-optimal solution satisfying (3), each agent is required to transmit:

$$8b \cdot D \cdot \mu^{-1} L^{\frac{3}{2}} \log(f(\theta^0)/\epsilon) \quad \textit{bits/agent}.$$

The quantization strategy at the first iteration (in Appendix C) requires b_0 to be $\mathcal{O}\big(\log(D)\big)$, so we can conclude $b = \mathcal{O}\big(\log(D)\big)$. Thus Theorem 4.1 gives the total communication complexity for each agent as $\mathcal{O}\big(D\log(D) \cdot \log(\frac{1}{\epsilon})\big)$, which matches the lower bound in Theorem 3.1, up to a gap $\log(D)$.

4.2 Communication Complexity Upper Bounds for $\mathcal{C}_{\text{op}}$

Next, we develop algorithms for the distributed overparameterized problem defined in Def. 2. Note that functions in the class $\mathcal{C}_{\text{op}}$ can be non-convex, therefore one may not be able to find an algorithm that can solve the entire class of problems in polynomial time. Our strategy is to identify special cases of $\mathcal{C}_{\text{op}}$ for which the analysis is possible. Again, although these problems are restrictions of $\mathcal{C}_{\text{op}}$, we do not expect the lower bound in Theorem 3.2 to be improved in either the dependency on N or on ϵ.

The Case for Quadratic Problems. Towards this end, let us consider an overparameterized quadratic problem. To proceed, assume that each agent k has data tuple (A_k, b_k) with N_k samples, where $A_k \in \mathbb{R}^{N_k \times D}$ is the feature matrix, and $b_k \in \mathbb{R}^{N_k \times 1}$ is the label. Denote $A := [A_1; \cdots; A_K] \in \mathbb{R}^{N \times D}$, where $N = \sum_{k=1}^{K} N_k$. Suppose $N < D$. Then the generic problem (1) reduces to:

$$f(x) = \frac{1}{2}\|A\theta - b\|^2 = \frac{1}{2}\sum_{k=1}^{K}\|A_k\theta - b_k\|^2. \tag{4}$$

Assuming that each A_k has full row rank, then the above problem is a special case of $\mathcal{C}_{\text{op}}$.

Next, we specialize Alg. 1. Recall that the lower bound derived in Theorem 3.2 is *independent* of D, so we cannot directly apply the algorithm we used for the class $\mathcal{C}_{\text{pl}}$, in which each agent transmits a dimension D vector at every iteration. To proceed, let us consider the following *linear compression* scheme, in which we use a matrix B of size $H \times D$ to compress (resp. decompress) the gradients that get transmitted (resp. received), where H is a fixed positive constant satisfying $N \leqslant H \ll D$. In particular, let us specialize the two processing operators in Alg. 1 as: $F(x) = Bx,\ \tilde{F}(y) = B^\top y$. Further, define $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ (resp. $s_{\max}(\cdot)$ and $s_{\min}(\cdot)$) as the largest and smallest eigenvalues (resp. singular values) for a matrix. Then we have the following result; see Appendix D for proof.

Theorem 4.2. *Consider using Alg. 1 to solve the quadratic problem (4), with A being full row rank. Choose B such that* $\text{rank}(BA^\top) = N$*; set $\eta = \frac{\sigma_{\min} Z)}{\sigma^2_{\max}(Z)}$; set b and C according to (30), and choose $\tilde{F}(\cdot)$ and $F(\cdot)$ as $F(x) = Bx,\ \tilde{F}(y) = B^\top y$. Then the following holds true:*

$$f(\theta^{t+1}) \leq (1 - 1/2\kappa^4) f(\theta^t),$$

where κ is the condition number of the matrix $Z := AB^\top BA^\top$ given by $\kappa(Z) := \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)}$. Further, to compute an ϵ-optimal solution (3), each agent is required to transmit:

$$2b \cdot \kappa^4 \cdot H \log(f(\theta^0)/\epsilon) \quad \textit{bits/agent}. \tag{5}$$

In Theorem 4.2, it is required that matrix $BA^\top$ is full rank, and this can be satisfied with probability 1 if we choose B to be a random matrix. Further, the convergence rate depends on the condition number $\kappa(Z)$. Although we have showed the per iteration communication cost can be reduced to H by linear compression, it is not clear yet whether the dimension D can affect the value of $\kappa(Z)$. In the following, we show $\kappa(Z)$ is independent of D; The proof can be found in Appendix E.

Proposition 1. *Consider the matrices A and B in Theorem 4.2, and define $Z := AB^\top BA^\top$. Suppose each entry in B satisfies $B_{hj} \overset{i.i.d}{\sim} \mathcal{N}(0,1), \forall\, h \in [H], j \in [D]$. Then the following holds:*
(1) The condition number $\kappa(Z)$ is independent of D.
(2) Let $A^\top = V\Sigma W$ be the compact form of SVD, where $V \in \mathbb{R}^{D\times N}$ satisfies $V^T V = \mathbf{I}_N$; $\Sigma \in \mathbb{R}^{N\times N}$ is the diagonal matrix whose diagonal entries are singular values of A; and $W \in \mathbb{R}^{N\times N}$ is a unitary matrix. Then for any fixed A, we have

$$\mathbb{P}\left(\kappa(Z) \leqslant \left(\frac{s_{\max}(\Sigma)}{s_{\min}(\Sigma)}\right)^2 \cdot \left(\frac{\sqrt{H}+\sqrt{N}+t}{\sqrt{H}-\sqrt{N}-t}\right)^2\right) \geqslant 1 - e^{-t^2/2}.$$

Proposition 1 implies that, as long as the compression matrix is standard normal, then $\kappa(Z)$ is independent of D. This means that the total bits required for each agent, as provided in (8), will only be dependent on N, H and $\log(f(\theta^0)/\epsilon)$, but not on D. To further understand the dependency of $\kappa(Z)$ with N and H, let us treat the ratio between the largest and smallest singular values as being fixed [1]. Then set $2e^{-t^2/2} = \delta > 0$, we have $t = \sqrt{2\log(\frac{2}{\delta})}$, so the upper bound for κ becomes $\kappa(AA^\top) \cdot \frac{\sqrt{H}+\sqrt{N}+\sqrt{2\log(\frac{2}{\delta})}}{\sqrt{H}-\sqrt{N}-\sqrt{2\log(\frac{2}{\delta})}}$. Assuming that $N \geqslant 4t^2, H = 4N$, then we can conclude that $\mathbb{P}(\kappa(Z) \leq 49) \geq 1 - \delta$. Applying this result to (8), we can conclude that with probability $1-\delta$, the total number of bits transmitted by each agent is $8b \cdot \kappa^4(A^T A) \cdot N \log(f(\theta^0)/\epsilon)$. Similarly as in Theorem 4.1, b is in the order of $\mathcal{O}(\log(D))$ (see Appendix C.). Therefore the total communication complexity of each agent can be reduced to $\mathcal{O}\big(H \log(D) \log(\frac{1}{\epsilon})\big)$. When $H = \Theta(N)$, then the complexity bound matches the lower bound in Theorem 3.2, up to a gap $\mathcal{O}(\log(D))$. Additionally,

[1] This is reasonable, since even when analyzing strongly convex distributed problems, the dependency on the condition number is unavoidable; see [Magnússon et al., 2020, Sec. III-C].

we note that by fixing the seed of random number generator, the random matrix B can be generated by each agent locally, without needing to transmit any information.

Decentralized Algorithm. We propose a *fully decentralized algorithm* for solving (4) with a communication complexity that approaches the lower bound in Theorem 3.2. Our idea is similar to Alg. 1 that adopts a dimension reduction step for compression, yet we shall encapsulate the `Quantize`, `Communicate`, `Aggregation`, `Broadcast` steps through a decentralized average consensus protocol with support for quantized communication.

In particular, we consider a setting where K agents are represented as nodes on a connected graph (V, E) with a doubly stochastic mixing matrix $W \in \mathbb{R}_+^{K\times K}$. Observe the following fact:

Fact 1. *Consider a set of vectors $v_1, \ldots, v_K \in \mathbb{R}^H$ held by K agents on (V, E). For any $\epsilon > 0$, the* `CHOCO-GOSSIP` *protocol [Koloskova et al., 2019] is described as the map:*

$$\bar{v}_1, \ldots, \bar{v}_K = \texttt{CHOCO-GOSSIP}(v_1, \ldots, v_K; \epsilon)$$

such that $\bar{v}_k$ is the vector available at agent k satisfying $\|\bar{v}_k - K^{-1}\sum_{j=1}^K v_j\|^2 \leq \epsilon$. Let $\xi = \sum_{k=1}^K \|v_k - K^{-1}\sum_{j=1}^K v_j\|^2$ be the initial consensus error, the protocol requires at most

$$(82/\delta^2\omega)\log(\xi/\epsilon) \text{ rounds of quantized communication},$$

where $\delta = 1 - |\lambda_2(W)|$ and $\omega \in (0, 1]$ is controlled by the number of bits in the quantizer. Details about the `CHOCO-GOSSIP` *protocol can be found in Appendix F.1.*

We replace line 5–line 8 in Alg. 1 by `CHOCO-GOSSIP` to yield the decentralized algorithm:

Algorithm 2 Decentralized Gradient Descent with Compressed Comm. via Linear Compression

Input: step size η; initial iterate $\theta^0 \in \mathbb{R}^D$.
`Initialize`: $\theta_i^0 = \bar{\theta}^0$, $\forall i \in [K]$
for t **in** $0, \ldots, T-1$ **do**
  `Compute`: Each agent compute the compressed local gradient $g_k^t = F(\nabla f_k(\theta_k^t))$.
  `Consensus`: Agents exchange gradients $(\bar{g}_1^t, \ldots, \bar{g}_K^t) = \texttt{CHOCO-GOSSIP}(g_1^t, \ldots, g_K^t; \frac{\bar{\epsilon}}{t+1})$.
  `Update`: Each agent updates $\theta_k^{t+1} = \theta_k^t - \eta\tilde{F}(\bar{g}_k^t)$.
end for
Output: the last iterate θ_i^T

Theorem 4.3. *Consider using Alg. 2 to solve the quadratic problem (4), with A being full row rank. Choose B such that $\mathrm{rank}(BA^\top) = N$; set $\eta = \frac{\sigma_{\min}(Z)}{4\sigma_{\max}^2(Z)}$, and choose $\tilde{F}(\cdot)$ and $F(\cdot)$ as $F(x) = Bx$, $\tilde{F}(y) = B^\top y$. For any $T \geq 1$, the following holds true:*

$$f(\theta_k^T) \leq \left(1 - 1/4\kappa^2\right)^T f(\bar{\theta}^0) + \mathcal{O}(\bar{\epsilon}^2),\ k = 1, \ldots, K.$$

where κ was defined in Theorem 4.2. Moreover, suppose that A, B are constructed as in Proposition 1, to compute an ϵ-optimal solution (3), each agent is required to transmit:

$$\mathcal{O}\left(\frac{\kappa^2 H \log H}{\delta^2}\left[\log\frac{1}{\epsilon} + \log\left(1 + \frac{\sqrt{D} + \sqrt{H} + \sqrt{\log 1/\zeta}}{n\sigma_{\max}^2(Z)}\right)\right]\log\frac{1}{\epsilon}\right) \text{ bits},$$

with probability at least $1 - \zeta$.

Proposition 1 shows that the condition number κ is $\mathcal{O}(1)$. To satisfy $\mathrm{rank}(BA^\top) = N$, we need $H \geq N$. Altogether, this yields the communication complexity $\mathcal{O}(N\log(N)\log(D)\log(1/\epsilon)^2)$ to match Theorem 3.2 up to a $\log(D)$ factor.

The Case of Overparameterized Neural Network. We now explore the possibility of extending the analysis of quadratic problems to a specific overparameterized neural network, still belonging to class $\mathcal{C}_{\mathrm{op}}$. We will customize Alg. 1 to partially compress the gradients of the loss function. Our idea is to design $F(\cdot)$ (resp. $\tilde{F}(\cdot)$) so that we compress (resp. decompress) the gradient with respect to the parameters in the widest layer to $\mathcal{O}(N)$ bits, while directly quantize the remaining gradients. Note that for a few popular network architectures such as Neural Tangent Kernel (NTK) [Huang

et al., 2021], the widest layer contains most of the parameters $\Omega(N^4)$. Therefore, compressing these parameters into $\mathcal{O}(N)$ bits already represents significant communication . We will leave the design and analysis of more aggressive schemes which compress the parameters of all layers for future work.

To proceed, let us first state the structure of the network. Let $X \in \mathbb{R}^{N \times n_0}$ denote the data sample; let $a(\cdot)$ be an activation function; let $O_l \in \mathbb{R}^{N \times n_l}$ as the output of each layer, which can be expressed as

$$O_l = \begin{cases} X & l = 0 \\ a\left(O_{l-1} W_l\right) & l \in [L-1] \\ O_{L-1} W_L & l = L \end{cases},$$

where $W_l \in \mathbb{R}^{n_{l-1} \times n_l}$ is the weight matrix in l-th layer. Further, denote $O_{L,k}$ as the output of the network for k-th agent. Let $\theta = (W_l)_{l=1}^L$ collect all the weights. Then objective function becomes

$$f(\theta) = \sum_{k=1}^{K} f_k(\theta) = \frac{1}{2} \sum_{k=1}^{K} \|O_{L,k} - y_k\|^2. \tag{6}$$

Further, we define the vectorized gradient of each layer as $u_l = \mathrm{vec}(\nabla_{W_l} f(\theta))$ and $u_{l,k}$ for each agent. Now we describe the assumptions on the network.

Assumption 4. *(Pyramidal network topology) Let $n_1 \geq N$ and $n_2 \geq n_3 \geq \ldots \geq n_L = 1$.*

Assumption 5. *(Activation function) Fix $\nu \in (0,1)$ and $\rho > 0$. Let a satisfy that: $(i) a'(x) \in [\nu, 1]$, (ii) $|a(x)| \leqslant |x|$ for every $x \in \mathbb{R}$, and (iii) a' is ρ-Lipschitz.*

As showed in [Nguyen and Mondelli, 2020], with Assumption 4 and 5, objective (6) belongs to $\mathcal{C}_{\mathrm{op}}$.

With the above Pyramidal structure, layer $l = 2$ is the widest layer and contains the largest number of parameters, denoted as $u_2 \in \mathbb{R}^{n_1 \times n_2}$. As mentioned above, we will design F and $\tilde{F}$ to compress and decompress u_2. To proceed, Let us choose $B = \mathbf{I}_{n_2} \otimes \tilde{B} \in \mathbb{R}^{H n_2 \times n_1 n_2}$ where $\tilde{B} \in \mathbb{R}^{H \times n_1}$, such that $N \leqslant H \ll n_1$. Then F and $\tilde{F}$ are defined as:

$$F(u_l) = \begin{cases} 0 & l = 1, \\ B u_l & l = 2, \\ u_l & l \geqslant 3, \end{cases} \quad \tilde{F}(F(u_l)) = \begin{cases} 0 & l = 1, \\ B^\top B u_l & l = 2, \\ u_l & l \geqslant 3. \end{cases} \tag{7}$$

This means that we compress the parameters in the second layer, and perform usual gradient descent for the layers $l \geq 3$; further, the parameters in the first layer is left frozen at its initialization. Note that we freeze the first layer for some technical reasons, but layer freezing itself is a useful technique to accelerate the training in practice, which can provide strong generalization ability [Yosinski et al., 2014, Advani et al., 2020].

Let us denote O_1^0 as the output of the first hidden layer at initialization. We have the following result. Its proof as well as the expressions for various constants, can be found in Appendix G.

Theorem 4.4. (Informal) *Consider using Alg. 1 to solve the quadratic problem (6), with X being full row rank. Suppose θ^0 is initialized properly (see Appendix G for details). Choose $\tilde{B}$ such as $\mathrm{rank}(\tilde{B} O_1^{0\top}) = N$; choose $\tilde{F}(\cdot)$ and $F(\cdot)$ as in (7). Then there exists stepsize η, constant ϕ, C, b such that to compute an ϵ-optimal solution (6), each agent is required to transmit:*

$$\frac{4b}{\eta\phi} \cdot \left(H n_2 + \sum_{l=2}^{L-1} n_l n_{l+1}\right) \log(f(\theta^0)/\epsilon) \quad \textit{bits/agent}, \tag{8}$$

where η, ϕ are related to X, ν, ρ and singular values of the weight matrices at initialization.

Theorem 4.4 shows that if we choose $H = \Theta(N)$, and when n_1 is very large, then the total communication complexity can be reduced by a factor of n_1/H.

Preliminary Numerical Experiments. We conclude by presenting a numerical experiment for the UCI Tom's Hardware[2] dataset using Alg. 2 where we applied $\lfloor \log(t+100) \rfloor$ rounds of communication at the t-th iteration for the `CHOCO-GOSSIP` subroutine; see Appendix F.1. We consider a ring network with $K = 5$ agents, each one has 500 or 1000 samples (thus making $N = 2500$, or $N = 5000$).

[2] Available: https://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+

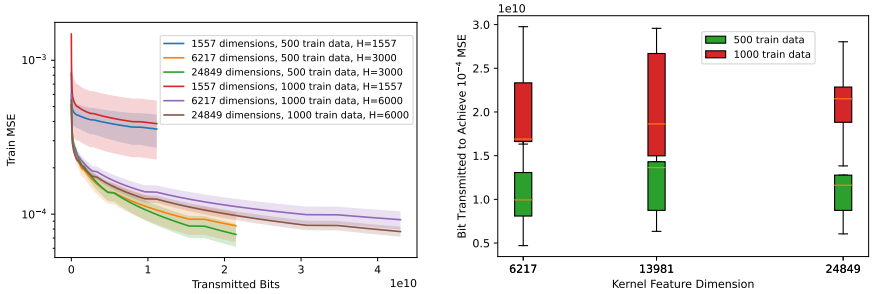

Figure 1: **Overparameterized Kernel Regression with Alg. 2.** Both figures show statistics of 5 random seeded runs. (Left) Training MSE against the number of bit transmitted on the whole network. (Right) Communication cost to achieve 10^{-4} MSE on training dataset, compared over different dimensions of kernel features.

We construct D-dimensional features from the dataset as NTK features [Bietti and Mairal, 2019]. In Fig. 1, we train a least square regression model in the overparameterized regime. We observe that the communication cost needed is proportional to the number of samples, instead of the feature dimension. Particularly, our results indicate that the communication cost required to reach a target train MSE (of 10^{-4}) is proportional to the number of training samples.

Acknowledgments and Disclosure of Funding We sincerely thank the anonymous reviewers for the suggestions. M. Hong and B. Song are supported by NSF grants CIF-1910385, CMMI-1727757 and CNS-2003033.

References

M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30: 1709–1720, 2017.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Y. Arjevani and O. Shamir. Communication complexity of distributed convex learning and optimization. In *Proceedings of NIPS*, pages 1756–1764, 2015.

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c4ef9c39b300931b69a36fb3dbb8d60e-Paper.pdf.

M. Braverman, F. Ellen, R. Oshman, T. Pitassi, and V. Vaikuntanathan. A tight bound for set disjointness in the message-passing model. *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 668–677, 2013.

Y. Deng, M. M. Kamani, and M. Mahdavi. Local sgd optimizes overparameterized neural networks in polynomial time. In *International Conference on Artificial Intelligence and Statistics*, pages 6840–6861. PMLR, 2022.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik. Marina: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

B. Huang, X. Li, Z. Song, and X. Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pages 4423–4434. PMLR, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019.

A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

J. H. Korhonen and D. Alistarh. Towards tight communication lower bounds for distributed optimisation. *Advances in Neural Information Processing Systems*, 34, 2021.

D. Kovalev, E. Gasanov, A. Gasnikov, and P. Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34, 2021.

D. Kovalev, A. Beznosikov, A. Sadiev, M. Persiianov, P. Richtárik, and A. V. Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *CoRR*, abs/2202.02771, 2022. URL https://arxiv.org/abs/2202.02771.

J. Li, Y. Liu, and W. Wang. Distributed learning with random features. *arXiv preprint arXiv:1906.03155*, 2019.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

Y. Lu and C. De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pages 7111–7123. PMLR, 2021.

S. Magnússon, H. Shokri-Ghadikolaei, and N. Li. On maintaining linear convergence of distributed learning and optimization under limited communication. *IEEE Transactions on Signal Processing*, 68:6101–6116, 2020.

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

Q. N. Nguyen and M. Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

K. K. Patel and A. Dieuleveut. Communication trade-offs for synchronized distributed sgd with large step size. *arXiv preprint arXiv:1904.11325*, 2019.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

M. Razaviyayn, M. Hong, N. Reyhanian, and Z.-Q. Luo. A linearly convergent doubly stochastic gauss–seidel algorithm for solving linear equations and a certain class of over-parameterized optimization problems. *Mathematical Programming*, 176(1):465–496, 2019.

A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing*, 67(19):4934–4947, 2019.

P. Richtárik, I. Sokolov, and I. Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

M. Safaryan, E. Shulgin, and P. Richtárik. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *arXiv preprint arXiv:2002.08958*, 2020.

M. Safaryan, F. Hanzely, and P. Richtárik. Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization. *Advances in Neural Information Processing Systems*, 34:25688–25702, 2021.

K. Scaman, F. Bach, S. Bubeck, Y. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. *arXiv preprint arXiv:1702.08704*, 2017.

K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2740–2749, 2018.

S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

H. Sun and M. Hong. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. *IEEE Transactions on Signal processing*, 67(22):5912–5928, 2019.

J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang. Lazily aggregated quantized gradient innovation for communication-efficient federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu. Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, 31, 2018.

J. N. Tsitsiklis. Problems in decentralized decision making and computation. Technical report, Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

J. N. Tsitsiklis and Z.-Q. Luo. Communication complexity of convex optimization. *Journal of Complexity*, 3(3):231–243, 1987.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

S. S. Vempala, R. Wang, and D. P. Woodruff. The communication complexity of optimization. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1733–1752. SIAM, 2020.

T. Vogels, S. P. Karimireddy, and M. Jaggi. Practical low-rank communication compression in decentralized deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 14171–14181, 2020.

H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in Neural Information Processing Systems*, 31, 2018.

J. Wang and G. Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 30, 2017.

H. Xu, C.-Y. Ho, A. M. Abdelmoniem, A. Dutta, E. H. Bergou, K. Karatsenidis, M. Canini, and P. Kalnis. Compressed communication for distributed deep learning: Survey and quantitative evaluation. Technical report, 2020.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Related Literature and Comparison

In the first category, where the commmunication complexities are measured by *rounds* of communications (where in each round real-valued vectors get exchanged), recent works [Kovalev et al., 2021, 2022] provides lower bounds for a sum of smooth and strongly convex functions over time-varying networks, and for strongly monotone variational inequality problems in a stochastic (finite-sum) setting. Other related works include [Scaman et al., 2017, 2018, Arjevani and Shamir, 2015]. In addition, lower bounds for non-convex problems are considered in [Sun and Hong, 2019, Lu and De Sa, 2021]. Further, there are many works that derive rounds of communication upper bounds for decentralized and federated learning algorithms, see, e.g.,[Stich and Karimireddy, 2019, Stich et al., 2018, Patel and Dieuleveut, 2019, Yu et al., 2019, Wang and Joshi, 2018, Gorbunov et al., 2021].

A recent work [Gorbunov et al., 2021] has analyzed the communication efficient algorithm to solve functions that satisfy PL conditions. However, it has been focused on analyzing the number of communication rounds needed to achieve certain ϵ optimal solution, while the current paper is focused on finding the minimum bits to be communicated. Therefore, although the two works are both about developing communication efficient algorithms for PL functions, the bounds obtained in these works represent different physical quantities, thus cannot be directly compared.

B Communication Complexity Lower Bounds

Let us introduce the so-called equality problem Vempala et al. [2020], denoted as EQUAL.

Definition 5 (Equality Problem)**.** *Consider a set of K agents, each is given an input $c_k \in \{0,1\}^m$; m is the length of the binary input. Then, the* EQUAL *problem is defined as follows:*

$$\mathsf{EQUAL}_m(c_1, \ldots, c_K) = \begin{cases} 1, & \text{if } c_1 = \ldots = c_K \\ 0, & \text{otherwise} \end{cases}$$

For any deterministic algorithms, the communication complexity lower bound of $\mathsf{EQUAL}_m(c_1, \ldots, c_K)$ is $\Omega(Km)$ [Vempala et al., 2020, Thm 3.5].

Further, the intermediate steps required to derive the lower bounds involve packing arguments. Therefore, we provide below a lower bound for the maximum number of points that we can pack into a compact set $[0,1]^n$, such that the distance between each pair of points is at least δ.

Definition 6 (Packing Problem)**.** *We define the following:*

- *For a given $\delta > 0$ we define the set $S(\delta) \subseteq [0,1]^n$ such that $\|x - y\| > \delta, \forall x, y \in S(\delta)$.*
- *Assuming that $|S(\delta)| \geq 2^m$, we define a function $h : \{0,1\}^m \to S(\delta)$. For $u, v \in \{0,1\}^m$ it holds that $u \neq v \Leftrightarrow h(u) \neq h(v)$.*

Lemma B.1. *([Korhonen and Alistarh, 2021, Lemma 2]) For a set $S(\delta) \subseteq [0,1]^n$ defined in Def. 6, it holds that $|S(\delta)| \geq \left(\frac{\sqrt{2n}}{\sqrt{\pi e}\delta}\right)^n$.*

Next, we repeat here for completeness the Assumptions that the local functions in the Distributed PL problem class satisfy (Def. 1).

Assumption 6. *The local objective functions satisfy:*

$$2\mu_k \cdot (f_k(\theta) - f_k(\theta^*_{(k)})) \leq \|\nabla f_k(\theta)\|^2, \quad \forall\, \theta, \quad \forall\, k,$$

where $\theta^*_{(k)}$ is a global minimum of $f_k(\cdot)$; μ_k's some positive constants.

Assumption 7. *There exists positive constants L_k's and L such that:*

$$\|\nabla f_k(\theta) - \nabla f_k(\theta')\| \leq L_k \|\theta - \theta'\|, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\|, \quad \forall\, \theta, \theta',\ \forall\, k.$$

In the Lemma below we show that the function instance we are going to use in Theorem 3.1 satisfies the PL condition and both Assumptions 6 and 7.

Lemma B.2. *The function $f_k(\theta) = \frac{1}{2}\|\theta - c\|_2^2 + \sin^2(\theta_k - c_k), c \in \mathbb{R}^D$ (where with c_k, θ_k we denote the kth component of vectors c, θ, respectively) satisfies the PL condition with $\mu = \frac{1}{8}$. That is, it holds that $f_k(\theta) - f_k(\theta^*) \leq \frac{1}{2\mu}\|\nabla f_k(\theta)\|^2, \forall \theta \in \mathbb{R}^D$, where θ^* is the global minimum of $f_k(\cdot)$.*

Remark 1. *The constructed function $f_k(\theta)$ and the respective sum across nodes $\sum_{k=1}^{K} f_k(\theta)$ satisfy Assumptions 6, 7. Specifically, Assumption 6 follows trivially from Lemma B.2. In addition, it can be shown that $f_k(\theta)$ and $\sum_{k=1}^{K} f_k(\theta)$ have bounded Hessians, which implies the Lipschitz gradient property of Assumption 7.*

Proof of Lemma B.2. To begin with we are going to show that the (one dimensional) function

$$\widetilde{f}_k(\theta_k) = \frac{1}{2}\left(\theta_k - c_k\right)^2 + \sin^2\left(\theta_k - c_k\right),$$

where $\theta_k, c_k \in \mathbb{R}$, satisfies the PL condition with $\mu = \frac{1}{8}$; notice that $\widetilde{f}_k(\theta_k)$ has a unique minimum with value 0 attained at $\theta_k = c_k$, and we have that $\widetilde{f}_k'(\theta_k) = \theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)$. Then, we are going to use this result to prove the PL condition for the function $f_k(\theta)$.

So, let us define the function

$$g(\theta_k) = 4\left[\theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)\right]^2 - \frac{1}{2}(\theta_k - c_k)^2 - \sin^2\left(\theta_k - c_k\right).$$

The gradient of the above function is given by

$$\begin{aligned} g'(\theta_k) &= 8\left[\theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)\right]\left[1 + 2\cos\left(2(\theta_k - c_k)\right)\right] - (\theta_k - c_k) - \sin\left(2(\theta_k - c_k)\right) \\ &= \left[\theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)\right]\left[7 + 16\cos\left(2(\theta_k - c_k)\right)\right]. \end{aligned}$$

In order to show that $\widetilde{f}_k(\theta_k)$ satisfies the PL property it suffices to prove that $g(\theta_k) \geq 0, \forall \theta_k \in \mathbb{R}$. Then, it will hold that

$$4\left[\theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)\right]^2 \geq \frac{1}{2}(\theta_k - c_k)^2 - \sin^2\left(\theta_k - c_k\right), \forall \theta_k \in \mathbb{R},$$

that is the PL condition of $\widetilde{f}_k(\theta_k)$ will be satisfied.

First, notice that

$$\begin{aligned} g(\theta_k) &= 4\left(\theta_k - c_k\right)^2 + 8\left(\theta_k - c_k\right)\sin\left(2(\theta_k - c_k)\right) + 4\sin^2\left(2(\theta_k - c_k)\right) - \frac{1}{2}(\theta_k - c_k)^2 - \sin^2\left(\theta_k - c_k\right) \\ &\geq \frac{7}{2}\left(\theta_k - c_k\right)^2 + 8\left(\theta_k - c_k\right)\sin\left(2(\theta_k - c_k)\right) - 1. \end{aligned}$$

It is clear from the above expression that the term $\frac{7}{2}\left(\theta_k - c_k\right)^2$ dominates the value of the objective for large values of $|\theta_k - c_k|$. Therefore, the objective does not become unbounded below.

Secondly, consider the stationary points of $g(\theta_k)$. Since the objective does not become unbounded below the only possible global minima of $g(\theta_k)$ are its stationary points. We are going to show that the values of the objective at those points is non-negative, effectively proving that $g(\theta_k) \geq 0, \forall \theta_k \in \mathbb{R}$. The stationary points of g are defined by the following expressions:

- $\sin\left(2(\theta_k - c_k)\right) = -\left(\theta_k - c_k\right)$

  Notice that in the interval $\theta_k \in \left(c_k, \frac{\pi}{2} + c_k\right]$ it holds that $\sin\left[2(\theta_k - c_k)\right] \geq 0$ but $-\left(\theta_k - c_k\right) < 0$. Similarly, for $\theta_k \in \left[-\frac{\pi}{2} + c_k, c_k\right)$ it holds that $\sin\left[2(\theta_k - c_k)\right] \leq 0$ but $-\left(\theta_k - c_k\right) > 0$. Also, for $\theta_k \notin \left[-\frac{\pi}{2} + c_k, c_k\right) \cup \left(c_k, \frac{\pi}{2} + c_k\right] \cup \{c_k\}$ it holds that $|\theta_k - c_k| \geq \frac{\pi}{2} > 1$. Therefore, in all the above cases the equation of the stationary points does not have a solution (and thus there are no stationary points). Finally, note that the equation has a trivial solution at $\theta_k = c_k$, which corresponds to the only stationary point we can get from this equation. For that point it holds that $g(c_k) = 0$.

- $\cos\left(2(\theta_k - c_k)\right) = -\frac{7}{16}$

  From the above equations, and by using the proper trigonometric identities it follows that $\sin^2\left(2(\theta_k - c_k)\right) = \frac{207}{256}$, $\sin^2\left(\theta_k - c_k\right) = \frac{23}{32}$ and $\theta_k - c_k = \kappa\pi \pm \frac{1}{2}\arccos\left(-\frac{7}{16}\right) \approx \kappa\pi \pm \frac{2.02}{2}$. Then if we plug the above values into $g(\theta_k)$ we get

$$g(\theta_k) \approx \frac{7}{2}\left(\kappa\pi \pm 1.01\right)^2 + 8\left(\kappa\pi \pm 1.01\right)\sin\left[2\kappa\pi \pm 2.02\right] + 4\frac{207}{256} - \frac{23}{32}$$

$$= \frac{7}{2}\left(\kappa\pi \pm 1.01\right)^2 + 8\left(\kappa\pi \pm 1.01\right)\sin\left[\pm 2.02\right] + \frac{644}{256}$$
$$= \left(\kappa\pi \pm 1.01\right)\left[\frac{7}{2}\left(\kappa\pi \pm 1.01\right) + 8\cdot\left(\pm 0.9\right)\right] + \frac{161}{64}.$$

It can be easily verified that $\left(\kappa\pi \pm 1.01\right)\left[\frac{7}{2}\left(\kappa\pi \pm 1.01\right) + 8\cdot\left(\pm 0.9\right)\right] + \frac{161}{64} > 0$ for all $\kappa \in \mathbb{Z}$.

In conclusion for all the stationary points $\hat{\theta}_k$ it holds that $g(\hat{\theta}_k) \geq 0$. As a result, we can claim that $g(\theta_k) \geq 0, \forall \theta_k \in \mathbb{R}$, and the function $\widetilde{f}(\theta_k)$ satisfies the PL property.

Next, consider the PL property (with $\mu = \frac{1}{8}$) for the function $f_k(\theta)$, that is

$$\frac{1}{2}\|\theta - c\|_2^2 + \sin^2\left(\theta_k - c_k\right) \leq 4\left\|\theta - c + \sin\left(2(\theta_k - c_k)\right)e_k\right\|^2, \tag{9}$$

where e_k is a vector of all zeros except at index k.

We have shown above that

$$\frac{1}{2}(\theta_k - c_k)^2 + \sin^2\left(\theta_k - c_k\right) \leq 4\left[\theta_k - c_k + \sin\left(2(\theta_k - c_k)\right)\right]^2. \tag{10}$$

Also, it trivially holds that

$$\frac{1}{2}\sum_{i=1, i\neq k}^{D}\left(\theta_i - c_i\right)^2 \leq 4\sum_{i=1, i\neq k}^{D}\left(\theta_i - c_i\right)^2. \tag{11}$$

Adding inequalities (10) and (11) we obtain condition (9), which ensures the PL property for the objective $f_k(\theta)$, with $\mu = \frac{1}{8}$. This completes the proof.

□

Proof of Theorem 3.1. To begin with, let $\pi \in \Pi(\epsilon)$ be an arbitrary protocol that solves the problems in $\mathcal{C}_{\text{pl}}$. Then, for any input (i.e., for any specific problem within the class $\mathcal{C}_{\text{pl}}$) the protocol returns an ϵ-approximate minimum $\tilde{\theta}$ as given in Def. 3.

Moreover, consider the set $S(\delta) \subseteq [0,1]^D$ introduced in Def. 6 (where in place of n we have D) with $\delta = 2\sqrt{2\epsilon}$. Then, Lemma B.1 implies that

$$|S(\delta)| \geq \left(\frac{\sqrt{2D}}{2\sqrt{2\pi e\epsilon}}\right)^D = \left(\frac{1}{2\sqrt{\pi e}}\right)^D\left(\frac{D}{\epsilon}\right)^{D/2}.$$

Then, we set $m = \mathbf{\Theta}\left(\log_2|S(\delta)|\right) = \mathbf{\Theta}\left(D\log\left(\frac{D}{\epsilon}\right)\right)$, where the exact value of m is selected such that $|S(\delta)| > 2^m$ holds. As a result, under the assumption that $\frac{D}{\epsilon} = \Omega(1)$, it holds that $|S(\delta)| \geq 2$, and $m \geq 1$.

Next, we will show that protocol π also solves the EQUAL problem. That is, we will reduce every instance of $\mathsf{EQUAL}_m(u_1, \ldots, u_K)$ to a problem in $\mathcal{C}_{\text{pl}}$. Towards this end, for an arbitrary input $(u_1, \ldots, u_K) \in \{0,1\}^{Km}$ we select the following function from $\mathcal{C}_{\text{pl}}$,

$$f(\theta) = \sum_{k=1}^{K} f_k(\theta) \text{ with } f_k(\theta) = \frac{1}{2}\|\theta - h(u_k)\|_2^2 + \sin^2\left(\theta_k - \left(h(u_k)\right)_k\right), \tag{12}$$

where $h(\cdot)$ is introduced in Def. 6, and $h(u_k) \in S(\delta) \subseteq [0,1]^D$; $\left(h(u_k)\right)_k, \theta_k$ denote the kth component of $h(u_k)$ and θ, respectively. It is shown in Lemma B.2 that the functions f_k's that correspond to each node satisfy the PL condition, with $\mu_k = \frac{1}{8}$, for all k. We can also easily verify that the rest of the conditions in Assumption 6 – 7 are satisfied.

Then, we have the following cases:

- Case 1 (equal inputs): It holds that $u_1 = \ldots = u_K := u$. Thus, we have that $h(u_1) = \ldots = h(u_K) = h(u)$ and the minimum of (12) is 0, attained at some point θ^* such that $\theta^* = h(u)$.

  Also, protocol π returns an approximate minimum $f(\tilde{\theta})$ of (12), for which it holds that

$$f(\tilde{\theta}) \overset{(a)}{\leq} f(\theta^*) + \epsilon = \frac{1}{2}\sum_{k=1}^{K} \|\theta^* - h(u)\|_2^2 + \sin^2\left(\theta_k^* - (h(u))_k\right) + \epsilon \overset{(b)}{=} \epsilon,$$

  where in (a) we used (3); and in (b) we exploited the fact that the minimum is 0.

- Case 2 (inputs not equal): There exists a pair of nodes (i, j) such that $u_i \neq u_j$. As a result, $h(u_i) \neq h(u_j)$. Then, protocol π returns an approximate minimum $f(\tilde{\theta})$ of (12) for which it holds that

$$\begin{aligned} f(\tilde{\theta}) &\overset{(a)}{\geq} f(\theta^*) \\ &= f_i(\theta^*) + f_j(\theta^*) + \sum_{k=1, k\neq i,j}^{K} f_k(\theta^*) \\ &\geq \frac{1}{2}\|\theta^* - h_i(u_i)\|_2^2 + \frac{1}{2}\|\theta^* - h_j(u_j)\|_2^2 \\ &\overset{(b)}{>} \frac{1}{2}\left(\frac{\delta}{2}\right)^2 = \frac{\delta^2}{8} \overset{(c)}{=} \epsilon, \end{aligned}$$

  where in (a) we used (3), in (b) we used the characteristic property (i.e., the minimum distance between two points is $\|x - y\| > \delta, \forall x, y \in S(\delta)$) of set $|S(\delta)|$ (from Def. 6), and in (c) we use the quantity $\delta = 2\sqrt{2\epsilon}$.

From the above analysis we see that if $f(\tilde{\theta}) \leq \epsilon$ then $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 1$. Otherwise, if we assume that $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 0$, then the analysis of case 2 implies that $f(\tilde{\theta}) > \epsilon$, a contradiction. Similarly, we can claim that if $f(\tilde{\theta}) > \epsilon$, then $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 0$. In the opposite case (i.e., if $f(\tilde{\theta}) > \epsilon \implies \mathsf{EQUAL}_m(u_1, \ldots, u_K) = 1$) we see from case 1 that $f(\tilde{\theta}) \leq \epsilon$, that is we reach a contradiction. In summary, we have that

$$\mathsf{EQUAL}_m(u_1, \ldots, u_K) = \begin{cases} 1, & \text{if } f(\tilde{\theta}) \leq \epsilon \\ 0, & \text{if } f(\tilde{\theta}) > \epsilon. \end{cases}$$

Finally, the fact that the communication complexity of $\mathsf{EQUAL}_m(u_1, \ldots, u_K)$ is $\Omega(Km)$, and the above reduction imply that $\Omega(Km) = \Omega(KD \log\left(\frac{D}{\epsilon}\right))$ is a lower bound for the communication complexity of $\mathcal{C}_{\text{pl}}$. □

Proof of Theorem 3.2. To begin with, let $\pi \in \Pi(\epsilon)$ be an arbitrary protocol that solves the problems in $\mathcal{C}_{\text{op}}$. Then, for any input (i.e., for any specific problem within the class $\mathcal{C}_{\text{op}}$) the protocol returns an ϵ-approximate minimum $\tilde{\theta}$, as described in Def. 3.

Moreover, consider the set $S(\delta) \subseteq [0, 1]^N$ introduced in Def. 6 (where in place of n we have N) with $\delta = 2\sqrt{2\epsilon}$. Then, Lemma B.1 implies that

$$|S(\delta)| \geq \left(\frac{\sqrt{2N}}{2\sqrt{2\pi e\epsilon}}\right)^N = \left(\frac{1}{2\sqrt{\pi e}}\right)^N \left(\frac{N}{\epsilon}\right)^{N/2}.$$

Then, we set $m = \mathbf{\Theta}\left(\log_2 |S(\delta)|\right) = \mathbf{\Theta}\left(N \log\left(\frac{N}{\epsilon}\right)\right)$, where the exact value of m is selected such that $|S(\delta)| > 2^m$ holds. As a result, under the assumption that $\frac{N}{\epsilon} = \Omega(1)$, it holds that $|S(\delta)| \geq 2$, and $m \geq 1$.

Now, we are going to show that protocol π also solves the EQUAL problem. That is, we are going to reduce every instance of $\mathsf{EQUAL}_m(u_1, \ldots, u_K)$ to a problem in $\mathcal{C}_{\text{op}}$. To be more precise, for an

arbitrary input $(u_1, \ldots, u_K) \in \{0,1\}^{Km}$ we select the following instance from $\mathcal{C}_{\text{op}}$,

$$f(\theta) = \sum_{k=1}^{K} f_k(\theta) \text{ with } f_k(\theta) = \frac{1}{2}\|G(\theta) - h(u_k)\|_2^2, \tag{13}$$

where $h(\cdot)$ is introduced in Def. 6; and with $h(u_k) \in S \subseteq [0,1]^N$.

Then, we have the following cases:

- Case 1 (equal inputs): It holds that $u_1 = \ldots = u_K := u$. Thus, we have that $h(u_1) = \ldots = h(u_K) = h(u)$ and the minimum of (13) is 0, attained at some point θ^* such that $G(\theta^*) = h(u)$.

  Also, protocol π returns an approximate minimum $f(\tilde{\theta})$ of (13) for which it holds that

  $$f(\tilde{\theta}) \overset{(a)}{\leq} f(\theta^*) + \epsilon = \frac{1}{2}\sum_{k=1}^{K} \|G(\theta^*) - h(u)\|_2^2 + \epsilon \overset{(b)}{=} \epsilon,$$

  where in (a) we used (3), and (b) follows from the fact that the minimum of (13) is 0 in this case (i.e., equal inputs).

- Case 2 (inputs not equal): There exists a pair of nodes (i,j) such that $u_i \neq u_j$. As a result, $h(u_i) \neq h(u_j)$. Then, protocol π returns an approximate minimum $f(\tilde{\theta})$ of (13) for which it holds that:

  $$\begin{aligned} f(\tilde{\theta}) &\overset{(a)}{\geq} f(\theta^*) \\ &= \frac{1}{2}\|G(\theta^*) - h(u_i)\|_2^2 + \frac{1}{2}\|G(\theta^*) - h(u_j)\|_2^2 + \frac{1}{2}\sum_{k=1, k\neq i,j}^{K} \|G(\theta^*) - h(u_k)\|_2^2 \\ &\geq \frac{1}{2}\|G(\theta^*) - h(u_i)\|_2^2 + \frac{1}{2}\|G(\theta^*) - h(u_j)\|_2^2 \\ &\overset{(b)}{>} \frac{1}{2}\left(\frac{\delta}{2}\right)^2 = \frac{\delta^2}{8} \overset{(c)}{=} \epsilon, \end{aligned}$$

  where in (a) expression (3) is used, in (b) we used the characteristic property (i.e., the minimum distance between two points is $\|x - y\| > \delta, \forall x, y \in S(\delta)$) of set $|S(\delta)|$ (from Def. 6), and in (c) we use the quantity $\delta = 2\sqrt{2\epsilon}$.

From the above analysis we see that if $f(\tilde{\theta}) \leq \epsilon$ then $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 1$. Otherwise, if $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 0$, then the analysis of case 2 implies that $f(\tilde{\theta}) > \epsilon$, a contradiction. Similarly, we can claim that if $f(\tilde{\theta}) > \epsilon$, then $\mathsf{EQUAL}_m(u_1, \ldots, u_K) = 0$. In the opposite case (i.e., if $f(\tilde{\theta}) > \epsilon \implies \mathsf{EQUAL}_m(u_1, \ldots, u_K) = 1$) we see from case 1 that $f(\tilde{\theta}) \leq \epsilon$, that is we reach a contradiction. In summary, we have that

$$\mathsf{EQUAL}_m(u_1, \ldots, u_K) = \begin{cases} 1, & \text{if } f(\tilde{\theta}) \leq \epsilon \\ 0, & \text{if } f(\tilde{\theta}) > \epsilon. \end{cases}$$

Finally, the fact that the communication complexity of $\mathsf{EQUAL}_m(u_1, \ldots, u_K)$ is $\Omega(Km)$, and the above reduction imply that $\Omega(Km) = \Omega(KN \log\left(\frac{N}{\epsilon}\right))$ is a lower bound for the communication complexity of $\mathcal{C}_{\text{op}}$. □

C Proof for Theorem 4.1

C.1 Proof of Lemma 4.2

Let us consider the function

$$\phi(\theta) = f(\theta) - \langle \nabla f(\theta^*), \theta \rangle.$$

Since $\nabla f(\theta^*) = 0$, it is easy to see $\phi(\theta) \geqslant 0$. It can be derived directly that $\phi(\theta^*) = 0$, which means $\phi(\cdot)$ can achieve the minimum at θ^*. So we know $\theta^* \in \operatorname{argmin} \phi(\theta)$. Then we have the following inequality holds:

$$\begin{aligned}
\phi(\theta^*) &\leqslant \phi\left(\theta - \frac{1}{L}\nabla\phi(\theta)\right) \\
&\overset{(i)}{=} f(\theta - \frac{1}{L}\nabla\phi(\theta)) \overset{(ii)}{=} f(\theta - \frac{1}{L}\nabla f(\theta)) \\
&\overset{(iii)}{\leqslant} f(\theta) - \frac{1}{2L}\|\nabla f(\theta)\|^2,
\end{aligned}$$

where (i) performed one step of gradient decent; (ii) uses the fact that $\nabla f(\theta^*) = 0$; (iii) is because the Lipshcitz gradient assumption. Then it follows directly

$$\|\nabla f(\theta)\|^2 \leqslant 2L \cdot f(\theta).$$

C.2 Proof for Theorem 4.1

First, let us state the sketch of the proof. Denote $g^t := \sum_{k=1}^{K} g_k^t, q^t := \sum_{k=1}^{K} q_k^t$.

Step1: We show that the loss function decreases linearly if all the agents update the averaged gradient in one step. That is, the following holds true.

$$f(\theta^t - \eta g^t) \leq (1 - \eta\mu) f(\theta^t).$$

Step2: We show by induction that for $t = 1, \cdots$, the following inequalities hold true:

$$(1) f(\theta^t) \leqslant (\alpha)^t f(\theta^0), \text{where } 0 < \alpha < 1, \tag{14}$$

$$(2) \|g^t - q^t\|_\infty \leqslant \frac{\gamma^{t-1}}{2^b - 1} = \tau\gamma^{t-1}, \text{for some } \gamma^{t-1} > 0, \tag{15}$$

where $\{\gamma^t\}_{t=1}^{\infty}$ is a sequence of positive numbers.

The proof of **Step 1** is straightforward:

$$\begin{aligned}
f(\theta^t - \eta g^t) &= f(\theta^t - \eta \sum_{k=1}^{K} g_k^t) \\
&\overset{(i)}{\leqslant} f(\theta^t) - \langle \nabla f(\theta^t), \eta \sum_{k=1}^{K} g_k^t \rangle + \frac{\eta^2 L}{2} \left\| \sum_{k=1}^{K} g_k^t \right\|^2 \\
&\overset{(ii)}{=} f(\theta^t) - \eta\|g^t\|^2 + \frac{\eta^2 L}{2}\|g^t\|^2 \\
&\overset{(iii)}{=} f(\theta^t) - \eta\|g^t\|^2 + \frac{\eta}{4L^2}\|g^t\|^2 \qquad (\eta = \frac{1}{2L^{\frac{3}{2}}}) \\
&\overset{(iv)}{\leqslant} f(\theta^t) - \frac{1}{2}\eta\|g^t\|^2 \\
&\overset{(v)}{\leqslant} f(\theta^t) - \eta\mu f(\theta^t) \\
&= (1 - \eta\mu) f(\theta^t),
\end{aligned} \tag{16}$$

where (i) is by Decent Lemma; (ii) uses the definition of g^t; (iii) uses the choice of η; (iv) is because $L > 1$; (v) uses Assumption 1.

Now we prove **Step 2**. To begin with, let us set $\eta = \frac{1}{2L^{\frac{3}{2}}}$, and set:

$$\alpha = 1 - \frac{\mu}{8L^{\frac{3}{2}}},\ \gamma^t = \sqrt{(\alpha)^{t+1} f(\theta^0)},\ \tau = \frac{1}{\sqrt{C}D}, \tag{17}$$

$$C = \max\left(\sqrt{\frac{16L}{\mu^2} + \frac{L^{\frac{3}{2}}}{\mu}}, 100\right),\ b = \max\left(\log(\frac{1}{\tau} + 1), b_0\right).$$

First, let us verify that (14) holds true when $t = 1$. We have the following series of inequalities:

$$
\begin{aligned}
f(\theta^1) &= f(\theta^0 - \eta q^0) - f(\theta^0 - \eta g^0) + f(\theta^0 - \eta g^0) \\
&\overset{(i)}{\leqslant} \eta \langle \nabla f(\theta^0 - \eta g^0), q^0 - g^0 \rangle + \frac{\eta^2 L}{2} \|q^0 - g^0\|^2 + f(\theta^0 - \eta g^0) \\
&\overset{(ii)}{\leqslant} \frac{\eta}{2\beta} \|\nabla f(\theta^0 - \eta g^0)\|^2 + 2\eta\beta \|q^0 - g^0\|^2 + \frac{\eta^2 L}{2} \|q^0 - g^0\|^2 + f(\theta^0 - \eta g^0) \\
&\overset{(iii)}{\leqslant} \frac{\eta L}{\beta} f(\theta^0 - \eta g^0) + (2\eta\beta + \frac{\eta^2 L}{2}) \|q^0 - g^0\|^2 + f(\theta^0 - \eta g^0) \\
&\overset{(iv)}{=} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2}) \|q^0 - g^0\|^2 \\
&\overset{(v)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2}) D^2 \|q^0 - g^0\|_\infty^2 \\
&\overset{(vi)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2}) D^2 (\sum_{k=1}^{K} \|q_k^0 - g_k^0\|_\infty)^2 \\
&\overset{(vii)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2}) D^2 \sum_{k=1}^{K} K \|q_k^0 - g_k^0\|_\infty^2 \\
&\overset{(viii)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2}) \frac{f(\theta^0)}{C^2 D^4} \\
&\overset{(ix)}{=} (1 - \frac{1}{2}\eta\mu) f(\theta^0) + \big(\frac{4\eta(1 - \eta\mu)L}{\mu} + \frac{\eta^2 L}{2}\big) \frac{f(\theta^0)}{C^2 D^4} \qquad \beta = \frac{2(1 - \eta\mu)L}{\mu} \\
&= \big(1 - \frac{1}{2}\eta\mu + (\frac{4\eta(1 - \eta\mu)L}{\mu} + \frac{\eta^2 L}{2})/C^2 D^4\big) f(\theta^0)
\end{aligned}
$$

where (i) comes from the Decent Lemma; (ii) uses the Young's inequality with constant β; (iii) uses Assumption 1; (iv) is from (16); (v) uses the relationship between ℓ_2 and ℓ_∞ norm; (vi) uses the triangle inequality; (vii) uses the Cauchy-Schwartz inequality; (viii) uses the initialization condition $\|q_k^0 - g_k^0\|_\infty \leqslant \frac{\sqrt{f_k(\theta^0)}}{CD^3\sqrt{K}}$; (ix) uses the choice of β.

Now let us define the quantization at initialization. For each entry in q_k^0, denoted as $(g_k^0)_j$, we consider the interval $\big[\lfloor (g_k^0)_j \rfloor, \lceil (g_k^0)_j \rceil\big]$, which is constructed by the closest integers. We use b_0 bits to make the grid, and quantize each element of the vector by the closest point on the grid. It is clear that the quantization error for each element is at most $\frac{1}{2^{b_0} - 1}$. So we set

$$
\frac{1}{2^{b_0} - 1} = \frac{\sqrt{f(\theta^0)}}{CD^3\sqrt{K}},
$$

or equivalently, $b_0 = \log\big(\frac{CD^3\sqrt{K}}{\sqrt{f(\theta^0)}} + 1\big)$.

Now since we have chosen $C \geqslant \frac{16L}{\mu^2} + \frac{L^{\frac{3}{2}}}{\mu}$, we can bound the coefficient in front of $f(\theta^0)$ as:

$$
\begin{aligned}
& 1 - \frac{1}{2}\eta\mu + \big(\frac{4\eta(1 - \eta\mu)L}{\mu} + \frac{\eta^2 L}{2}\big)/C^2 D^4 \\
& \overset{(i)}{\leqslant} 1 - \frac{1}{2}\eta\mu + \big(\frac{2(1 - \eta\mu)}{\sqrt{L}\mu} + \frac{1}{8L^2}\big)/C^2 \\
& \overset{(ii)}{\leqslant} 1 - \frac{1}{2}\eta\mu + (\frac{2}{\sqrt{L}\mu} + \frac{1}{8})/C^2 \\
& \overset{(iii)}{\leqslant} 1 - \frac{1}{4}\eta\mu
\end{aligned} \tag{18}
$$

where in (i) we plugged in the choice of η and $D \geqslant 1$; (ii) is because we have assumed that $L > 1$ and $D \geqslant 1$; (iii) uses the choice of C in (17). It follows $f(\theta^1) \leqslant (1 - \frac{\mu}{8L^{\frac{3}{2}}}) f(\theta^0)$. Thus, (14) holds for $t = 1$.

Second, let us analyze (15) for $t = 1$. Observe that:

$$\begin{aligned}
\|g^1 - q^0\|_\infty &\leqslant \|g^1 - g^0\| + \|g^0 - q^0\|_\infty \overset{(i)}{\leqslant} \eta L\|q^0\| + \|g^0 - q^0\|_\infty \\
&\overset{(ii)}{\leqslant} \eta L\|g^0\| + \eta L\|q^0 - g^0\| + \|q^0 - g^0\|_\infty \\
&\overset{(iii)}{\leqslant} \eta L\|g^0\| + (1 + \eta LD)\|q^0 - g^0\|_\infty \\
&\overset{(iv)}{\leqslant} \eta L\|g^0\| + (1 + \eta LD)\sum_{k=1}^{K} \frac{\sqrt{f_k(\theta^0)}}{CD^3\sqrt{K}} \\
&\overset{(v)}{\leqslant} \eta L\sqrt{2L \cdot f(\theta^0)} + (1 + \eta LD)\frac{\sqrt{f(\theta^0)}}{CD^3} \\
&\overset{(vi)}{\leqslant} \frac{\sqrt{2}}{2}\sqrt{f(\theta^0)} + 2 \cdot \frac{\sqrt{f(\theta^0)}}{100} \overset{(vii)}{\leqslant} \sqrt{\alpha f(\theta^0)},
\end{aligned}$$

where (i) uses Assumption 2; (ii) is from triangle inequality; (iii) uses the relationship between ℓ_2 and ℓ_∞ norm; (iv) uses the condition $\|q_k^0 - g_k^0\|_\infty \leqslant \frac{\sqrt{f_k(\theta^0)}}{CD^3\sqrt{K}}$; (v) uses Assumption 1 and the Cauchy-Schwartz inequality; (vi) plugs in the choice of stepsize and the fact that $C \geqslant 100$; (v) compares the left side with definition of α in (17). By Lemma 4.1, it follows that:

$$\|q^1 - g^1\|_\infty = \|\text{quant}(g^1, q^0, r, b) - g^1\|_\infty \leqslant \frac{\gamma^0}{2^b - 1}.$$

Next, let us use induction to prove (14) and (15). First, we analyze the decent of $f(\theta^t)$:

$$\begin{aligned}
f(\theta^{t+1}) &= f(\theta^t - \eta q^t) - f(\theta^t - \eta g^t) + f(\theta^t - \eta g^t) \\
&\overset{(i)}{\leqslant} \eta\langle\nabla f(\theta^t - \eta g^t), q^t - g^t\rangle + \frac{\eta^2 L}{2}\|q^t - g^t\|^2 + f(\theta^t - \eta g^t) \\
&\overset{(ii)}{\leqslant} \frac{\eta}{2\beta}\|\nabla f(\theta^t - \eta g^t)\|^2 + 2\eta\beta\|q^t - g^t\|^2 + \frac{\eta^2 L}{2}\|q^t - g^t\|^2 + f(\theta^t - \eta g^t) \\
&\overset{(iii)}{\leqslant} \frac{\eta L}{\beta} f(\theta^t - \eta g^t) + (2\eta\beta + \frac{\eta^2 L}{2})\|q^t - g^t\|^2 + f(\theta^t - \eta g^t) \\
&\overset{(iv)}{=} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu) f(\theta^t) + (2\eta\beta + \frac{\eta^2 L}{2})\tau^2(\gamma^{t-1})^2 \\
&\overset{(v)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu)(\alpha)^t f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2})\tau^2(\gamma^{t-1})^2 \\
&\overset{(vi)}{\leqslant} (\frac{\eta L}{\beta} + 1)(1 - \eta\mu)(\alpha)^t f(\theta^0) + (2\eta\beta + \frac{\eta^2 L}{2})\tau^2(\alpha)^t f(\theta^0) \\
&= \left(1 - \frac{1}{2}\eta\mu + (\frac{4\eta(1 - \eta\mu)L}{\mu} + \frac{\eta^2 L}{2})\tau^2\right)(\alpha)^t f(\theta^0),
\end{aligned} \tag{19}$$

where (i) comes from the Decent Lemma; (ii) uses the Young's inequality with constant β; (iii) uses Assumption 1; (iv) is from (16) and induction assumption (15); (v) uses the induction assumption (14); (vi) uses the definition of γ^{t-1}. Finally, set $\beta = \frac{2(1-\eta\mu)L}{\mu}$, we will get the last equality. Next, let us analyze the coefficient in the expression (19):

$$\begin{aligned}
1 - \frac{1}{2}\eta\mu + (\frac{4\eta(1 - \eta\mu)L}{\mu} + \frac{\eta^2 L}{2})\tau^2 &\overset{(i)}{=} 1 - \frac{1}{2}\eta\mu + (\frac{2(1 - \eta\mu)}{\sqrt{L}\mu} + \frac{1}{8L^2})\tau^2 \\
&\overset{(ii)}{\leqslant} 1 - \frac{1}{2}\eta\mu + (\frac{2}{\sqrt{L}\mu} + \frac{1}{8})\tau^2 \\
&\overset{(iii)}{=} 1 - \frac{\mu}{4L^{\frac{3}{2}}} + (\frac{2}{\sqrt{L}\mu} + \frac{1}{8})\tau^2 \\
&\overset{(iv)}{\leqslant} 1 - \frac{\mu}{8L^{\frac{3}{2}}} = \alpha,
\end{aligned} \tag{20}$$

where in (i) we plugged in the choice of η; (ii) is because we have assumed that $L > 1$; (iii) plugged in the choice of stepsize; (iv) uses the definition of τ in (17); and the last equality comes from the definition of α in (17). Plugging (20) to (19), we obtain $f(\theta^{t+1}) \leqslant (\alpha)^{t+1} f(\theta^0)$.

Second, we show that (15) holds for $t+1$. We have:

$$\begin{aligned}
\left\|g^{t+1} - q^t\right\|_\infty &\leqslant \|g^{t+1} - g^t\| + \left\|g^t - q^t\right\| \\
&\overset{(i)}{\leqslant} L\eta\|q^t\| + \tau\gamma^{t-1} \\
&\overset{(ii)}{\leqslant} L\eta(\|g^t\| + D\|g^t - q^t\|_\infty) + \tau\gamma^{t-1} \\
&\overset{(iii)}{\leqslant} L\eta\sqrt{2L}\sqrt{f(\theta^t)} + (1 + DL\eta)\tau\gamma^{t-1} \\
&\overset{(iv)}{\leqslant} L\eta\sqrt{L}\sqrt{(\alpha)^t f(\theta^0)} + (1 + DL\eta)\tau\gamma^{t-1} \\
&\overset{(v)}{=} (L\eta\sqrt{2L} + (1 + DL\eta)\tau)\gamma^{t-1} \\
&= (\frac{\sqrt{2}}{2} + (1 + \frac{D}{2})\tau)\gamma^{t-1} \\
&\overset{(vi)}{\leqslant} (\frac{\sqrt{2}}{2} + \frac{3}{2}D\tau)\gamma^{t-1} \overset{(vii)}{\leqslant} \frac{9}{10}\gamma^{t-1} \overset{(viii)}{\leqslant} \sqrt{\alpha}\gamma^t = \gamma^{t+1}
\end{aligned}$$

where (i) uses Assumption 2 and the induction assumption; (ii) uses the triangle inequality to decompose $\|q^t\|$ and uses the relationship between ℓ_∞ and ℓ_2 ; (iii) is from Assumption 1 and induction assumption (15); (iv) uses the induction assumption (14); (v) uses the definition of γ^t; (vi) is because $\mu < 1$; (vii) comes from the fact that $\tau < \frac{1}{10D}$ in (17); (viii) compares the choice of α in (17). Thus, we obtain $\left\|g^{t+1} - q^t\right\|_\infty \leqslant \sqrt{(\alpha)^{t+1} f(\theta^0)} = \gamma^t$. Then by Lemma 4.1, we have (15) holds for $t+1$.

Now we have proved by induction that (14) and (15) hold. So for $t > 0$, there is

$$f(\theta^t) \leqslant (\alpha)^t f(\theta^0), \text{where } \alpha = 1 - \frac{\mu}{8L^{\frac{3}{2}}}.$$

Thus, to compute an ϵ-optimal solution, the total number of iterations required is $\log(f(\theta^0)/\epsilon)/\log(1-\mu/8L^{\frac{3}{2}})$. Since in each iteration, each agent k transmits a length-D vector q_k^t, it follows that the total number of bits each agent needs to communicate is $D\log(f(\theta^0)/\epsilon)/\log(1/(1-\mu/8L^{\frac{3}{2}}))$bits. Notice that $\log(1/(1-\mu/8L^{\frac{3}{2}}) = -\log(1-\mu/8L^{\frac{3}{2}}) \sim 8L^{\frac{3}{2}}/\mu$, so we can derive the simplified total number of bits as $bD \cdot \frac{8L^{\frac{3}{2}}}{\mu} \log\left(f\left(\theta^0\right)/\epsilon\right)$.

D The Proof of Theorem 4.2

First, let us provide the sketch of the proof. Denote

$$g^t := \sum_{k=1}^K g_k^t, \quad q^t := \sum_{k=1}^K q_k^t, \quad \tilde{g}^t := B^\top g^t.$$

Using the above notation, the agents' local update step (i.e., the 'Update' step in Alg. 1) can be expressed as:

$$\theta_k^{t+1} = \theta_k^t - \eta B^\top q^t, \forall\, k. \tag{21}$$

Step 1: We show that the loss function decreases linearly if all the agents update parameters using the direction $\tilde{g}^t$, as follows:

$$f(\theta^t - \eta\tilde{g}^t) \leq \left(1 - \frac{1}{\kappa^2}\right) f(\theta^t).$$

Step 2: Let $\tau = \frac{1}{2^b - 1}$, we show by induction that for $t = 1, \cdots$, the following inequalities hold true:

$$(1)\ f(\theta^t) \leqslant (\alpha)^t f(\theta^0), \text{for some } 0 < \alpha < 1. \tag{22}$$

$$(2)\ \|g^t - q^t\|_\infty \leqslant \tau\gamma^{t-1},\ \|\tilde{g}^t - B^\top q^t\|_\infty \leqslant \tau H\|B^\top\|_\infty \gamma^{t-1}, \text{for some } \gamma^{t-1} > 0, \tag{23}$$

where $\{\gamma^t\}_{t=1}^\infty$ is a sequence of positive numbers.

We prove **Step 1** first. At t-the iteration, Let us first expand the objective function as:

$$f\left(\theta^t - \eta\tilde{g}^t\right) = \frac{1}{2}\|A(\theta^t - \eta\tilde{g}^t) - b\|^2 = f\left(\theta^t\right) - \left\langle A\theta^t - b, \eta A\tilde{g}^t\right\rangle + \frac{1}{2}\eta^2\left\|A\tilde{g}^t\right\|^2. \tag{24}$$

To proceed, let us provide explicit expressions for g^t and $\tilde{g}^t$:

$$g^t = B\nabla f(\theta^t) = BA^\top(A\theta^t - b) \tag{25}$$

$$\tilde{g}^t = B^\top B\nabla f(\theta^t) = B^\top BA^\top(A\theta^t - b). \tag{26}$$

By using the above, the inner product in (24) can be bounded as follows

$$-\left\langle A\theta^t - b, \eta A\tilde{g}^t\right\rangle = -\eta\langle A\theta^t - b, AB^\top BA^\top(A\theta^t - b)\rangle \leqslant -\eta\sigma_{\min}(Z)\left\|A\theta^t - b\right\|^2. \tag{27}$$

Using the above relations, we can further bound the descent of the objective function as

$$\begin{aligned}
f\left(\theta^t - \eta\tilde{g}^t\right) &\overset{(i)}{\leqslant} f(\theta^t) - \eta\sigma_{\min}(Z)\left\|A\theta^t - b\right\|^2 + \frac{1}{2}\eta^2\left\|AB^\top BA^\top\left(A\theta^t - b\right)\right\|^2 \\
&\overset{(ii)}{\leqslant} f(\theta^t) - \eta\sigma_{\min}(Z)\left\|A\theta^t - b\right\|^2 + \frac{1}{2}\eta^2\sigma_{\max}^2(Z)\left\|A\theta^t - b\right\|^2 \\
&\overset{(iii)}{=} f(\theta^t) - 2\eta\sigma_{\min}(Z)f(\theta^t) + \eta^2\sigma_{\max}^2(Z)f(\theta^t) \\
&= (1 - \eta\sigma_{\min}(Z))f(\theta^t) \\
&= (1 - \frac{1}{\kappa^2})f(\theta^t),
\end{aligned} \tag{28}$$

where (i) comes from plugging (27) and (26) into (24); (ii) extracts the largest eigenvalue of Z; (iii) is due to the definition of the objective function; the last two equalities hold due to the definition $\eta = \frac{\sigma_{\min}(Z)}{\sigma_{\max}^2(Z)}$.

Next, we prove **Step 2**. To begin with, let us define:

$$\alpha := 1 - \frac{1}{2\kappa^4}, \quad \lambda := 6\sqrt{2}\cdot\frac{\sigma_{\min}(Z)}{\sqrt{\sigma_{\max}(Z)}}, \quad \gamma^t := \lambda\sqrt{(\alpha)^{t+1}f(\theta^0)}. \tag{29}$$

Further let us set

$$b = \max\left(\log_2\left(\frac{1}{\tau} + 1\right), b_0\right), \quad C = \max(\frac{1}{\lambda^2\tau^2D^2}, \frac{12}{\lambda}) \quad \text{with} \quad \tau \quad \text{given by} \tag{30}$$

$$\tau := \min\left(\frac{1}{\sqrt{2\kappa^4\lambda^2D^2H^2\|B^\top\|_\infty^2 s_{\max}^2(A)\left(\frac{\sigma_{\min}^2(Z)}{2\sigma_{\max}^4(Z)} + \frac{1}{\sigma_{\max}^2(Z)}\right)}}, \frac{1}{6(1 + \frac{H}{\kappa})}\right).$$

First, we analyze (22) for $t = 1$. We have the following relations:

$$\begin{aligned}
f(\theta^1) &= f(\theta^0 - \eta B^\top q^0) - f(\theta^0 - \eta\tilde{g}^0) + f(\theta^0 - \eta\tilde{g}^0) \\
&\overset{(i)}{\leqslant} \frac{1}{2}\|\eta A(B^\top q^0 - \tilde{g}^0)\|^2 + \langle\eta A(B^\top q^0 - \tilde{g}^0), A(\theta^0 - \eta\tilde{g}^0) - b\rangle + (1 - \eta\sigma_{\min}(Z))f(\theta^0) \\
&\overset{(ii)}{\leqslant} \frac{1}{2}\|\eta A(B^\top q^0 - \tilde{g}^0)\|^2 + \eta\left(\beta\|A(B^\top q^0 - \tilde{g}^0)\|^2 + \frac{1}{2\beta}\|A(\theta^0 - \eta\tilde{g}^0) - b\|^2\right) + (1 - \eta\sigma_{\min}(Z))f(\theta^0) \\
&\overset{(iii)}{\leqslant} (\frac{1}{2}\eta^2 + \beta\eta)s_{\max}^2(A)\|B^\top q^0 - \tilde{g}^0\|^2 + \frac{\eta}{\beta}(1 - \eta\sigma_{\min}(Z))f(\theta^0) + (1 - \eta\sigma_{\min}(Z))f(\theta^0) \\
&= (\frac{1}{2}\eta^2 + \beta\eta)s_{\max}^2(A)\|B^\top q^0 - \tilde{g}^0\|^2 + (1 + \frac{\eta}{\beta})(1 - \eta\sigma_{\min}(Z))f(\theta^0) \\
&\overset{(iv)}{=} (\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s_{\max}^2(A)\|B^\top q^t - \tilde{g}^0\|^2 + (1 - \eta^2\sigma_{\min}^2(Z))f(\theta^0)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(v)}{\leqslant} D^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s_{\max}^2(A)\|B^\top q^0 - \tilde{g}^0\|_\infty^2 + (1-\eta^2\sigma_{\min}^2(Z))f(\theta^0) \\
&\overset{(vi)}{\leqslant} D^2H^2\|B^\top\|_\infty^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s_{\max}^2(A)\|q^0 - g^0\|_\infty^2 + (1-\eta^2\sigma_{\min}^2(Z))f(\theta^0) \\
&\overset{(vii)}{\leqslant} KD^2H^2\|B^\top\|_\infty^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s_{\max}^2(A)\sum_{k=1}^K\|q_k^0 - g_k^0\|_\infty^2 + (1-\eta^2\sigma_{\min}^2(Z))f(\theta^0) \\
&\overset{(viii)}{\leqslant} H^2\|B^\top\|_\infty^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s_{\max}^2(A)\frac{f(\theta^0)}{C^2D^4} + (1-\eta^2\sigma_{\min}^2(Z))f(\theta^0) \\
&= \left(1-\eta^2\sigma_{\min}^2(Z) + H^2\|B^\top\|_\infty^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})\frac{s_{\max}^2(A)}{C^2D^4}\right)f(\theta^0) \\
&\overset{(ix)}{=} \left(1 - \frac{1}{\kappa^4} + H^2\|B^\top\|_\infty^2\left(\frac{\sigma_{\min}^2(Z)}{2\sigma_{\max}^4(Z)} + \frac{1}{\sigma_{\max}^2(Z)}\right)s_{\max}^2(A)\lambda^2\tau^2/D^2\right)f(\theta^0) \\
&\overset{(x)}{\leqslant} (1 - \frac{1}{2\kappa^4})f(\theta^0) = \alpha f(\theta^0),
\end{aligned}
$$

where (i) explicitly expands the $f(\cdot)$ function, and uses (28); (ii) applies the Young inequality with constant β; (iii) extracts the largest singular value of A, and uses (28); (iv) set $\beta = \frac{1}{\sigma_{\min}(Z)}$; (v) uses the relationship between ℓ_2 and ℓ_∞ norm; (vi) uses the fact that $\|B^\top q^0 - \tilde{g}^0\| \leqslant H^2\|\tilde{B}\|_\infty\|q^0 - g^0\|_\infty$; (vii) uses the Cauchy-Schwartz inequality; (viii) uses the condition $\left\|q_k^0 - g_k^0\right\|_\infty \leqslant \frac{\sqrt{f(\theta^0)}}{CD\sqrt{K}}$ in Algorithm 1; (vii) uses the induction assumption (22) and the definition of γ^{t-1} in (29); (viii) plug in the choice of stepsize; (ix) comes from the choice of λ and τ. So we have showed (22) holds for $t = 1$; (ix) plug in the choice of stepsize and constant $C \geqslant \frac{1}{\lambda^2\tau^2}$; (x) uses the choice of λ and τ.

Next, let us analyze (23) for $t = 1$. The idea is that, if we can show that $\|g^1 - q^0\|_\infty \leqslant \gamma^0$, then we will be able to use Lemma 4.1 to show (23). More specifically, we have:

$$
\begin{aligned}
\|q^1 - g^1\| &\overset{(i)}{=} \|\text{quant}(g^1, q^0, \gamma^0, b) - g^1\|_\infty \overset{(ii)}{\leqslant} \tau\gamma^0, \\
\|B^\top q^1 - \tilde{g}^1\|_\infty &\overset{(iii)}{=} \|B^\top\text{quant}(g^1, q^0, \gamma^0, b) - B^\top g^1\|_\infty \overset{(iv)}{\leqslant} \tau H\|B^\top\|_\infty\gamma^0,
\end{aligned}
$$

where (i) and (iii) come from the 'Quantize' step in Algorithm 1; (ii) and (iv) are from the two inequalities in Lemma 4.1 (assuming that $\|g^1 - q^0\|_\infty \leqslant \gamma^0$ holds).

Next, we show $\|g^1 - q^0\|_\infty \leq \gamma^0$. We observe that:

$$
\begin{aligned}
\|g^1 - q^0\|_\infty &\overset{(i)}{\leqslant} \|g^1 - g^0\|_\infty + \|g^0 - q^0\|_\infty \\
&\overset{(ii)}{\leqslant} \eta\|BA^\top AB^\top q^0\| + \|g^0 - q^0\|_\infty \\
&\overset{(iii)}{\leqslant} \eta\sigma_{\max}(Z)\|q^0\| + \|g^0 - q^0\|_\infty \\
&\overset{(iv)}{\leqslant} \eta\sigma_{\max}(Z)(\|g^0\| + \|q^0 - g^0\|) + \|g^0 - q^0\|_\infty \\
&\overset{(v)}{\leqslant} \eta\sigma_{\max}(Z)(\|BA^\top(A\theta^0 - b)\| + H\|q^0 - g^0\|_\infty) + \|g^0 - q^0\|_\infty \\
&\overset{(vi)}{\leqslant} \eta\sigma_{\max}^{\frac{3}{2}}(Z)\|A\theta^0 - b\| + (1 + \eta H\sigma_{\max}(Z))\|g^0 - q^0\|_\infty \\
&\overset{(vii)}{\leqslant} \eta\sigma_{\max}^{\frac{3}{2}}(Z)\|A\theta^0 - b\| + (1 + \eta H\sigma_{\max}(Z))\sum_{k=1}^K\|g_k^0 - q_k^0\|_\infty \\
&\overset{(viii)}{\leqslant} \eta\sigma_{\max}^{\frac{3}{2}}(Z)\sqrt{2f(\theta^0)} + (1 + \eta H\sigma_{\max}(Z))\sum_{k=1}^K\frac{\sqrt{f_k(\theta^0)}}{CD^3\sqrt{K}} \\
&\overset{(ix)}{\leqslant} \eta\sigma_{\max}^{\frac{3}{2}}(Z)\sqrt{2f(\theta^0)} + (1 + \eta H\sigma_{\max}(Z))\frac{\sqrt{f(\theta^0)}}{CD^3}
\end{aligned}
$$

$$\overset{(x)}{=} \sqrt{f(\theta^0)}(\frac{1}{CD^3} + \frac{H}{CD^3\kappa} + \frac{\sqrt{2}\sigma_{\min}(Z)}{\sqrt{\sigma_{\max}(Z)}})$$
$$\overset{(xi)}{\leqslant} \lambda\sqrt{f(\theta^0)}(\frac{2}{\lambda C} + \frac{1}{6}) \overset{(xii)}{\leqslant} \frac{1}{3}\lambda\sqrt{f(\theta^0)} \leqslant \lambda\sqrt{\alpha f(\theta^0)} = \gamma^0$$

where (i) is due to the triangle inequality; (ii) expands the expression of g^1 and g^0 in (25), uses the relation between ℓ_2 norm and ℓ_∞ norm and uses the update rule (21); (iii) uses the fact that non-zero eigen values of $BA^\top AB^\top$ and Z are the same and extracts the largest eigen value of Z; (iv) uses triangle inequality ; (v) uses the relationship between ℓ_2 and ℓ_∞ norm; (vi) is because $s^2_{\max}(BA^\top) = \sigma_{\max}(Z)$ and extract the largest singular value of $BA^\top$; (vii) uses triangle inequality; (viii) uses the initial condition $\|q^0_k - g^0_k\|_\infty \leqslant \frac{\sqrt{f_k(\theta^0)}}{C\cdot D\sqrt{K}}$ in Algorithm 1; (ix) uses Cauchy-Schwartz inequality; (x) plug in the choice of stepsize; (xi) is because $\kappa \geqslant 1, H \leqslant D, D \geqslant 1$ and the choice of λ; (xii) is from $C \geqslant \frac{12}{\lambda}$; the last inequality comes from $\sqrt{\alpha} \geqslant \frac{\sqrt{2}}{2} > \frac{1}{3}$ since $\kappa \geqslant 1$. So we can show $\left\|g^1 - q^0\right\|_\infty \leq \gamma^0$.

Next, we will show (22) holds for $t+1$ by induction, based on the base assumption that (22) and (23) holds for t. We have the following series of relations:

$$\begin{aligned}
f(\theta^{t+1}) &= f(\theta^t - \eta B^\top q^t) - f(\theta^t - \eta\tilde{g}^t) + f(\theta^t - \eta\tilde{g}^t) \\
&\overset{(i)}{\leqslant} \frac{1}{2}\|\eta A(B^\top q^t - \tilde{g}^t)\|^2 + \langle \eta A(B^\top q^t - \tilde{g}^t), A(\theta^t - \eta\tilde{g}^t) - b\rangle + (1 - \eta\sigma_{\min}(Z))f(\theta^t) \\
&\overset{(ii)}{\leqslant} \frac{1}{2}\|\eta A(B^\top q^t - \tilde{g}^t)\|^2 + \eta\left(\beta\|A(B^\top q^t - \tilde{g}^t)\|^2 + \frac{1}{2\beta}\|A(\theta^t - \eta\tilde{g}^t) - b\|^2\right) + (1 - \eta\sigma_{\min}(Z))f(\theta^t) \\
&\overset{(iii)}{\leqslant} (\frac{1}{2}\eta^2 + \beta\eta)s^2_{\max}(A)\|B^\top q^t - \tilde{g}^t\|^2 + \frac{\eta}{\beta}(1 - \eta\sigma_{\min}(Z))f(\theta^t) + (1 - \eta\sigma_{\min}(Z))f(\theta^t) \\
&= (\frac{1}{2}\eta^2 + \beta\eta)s^2_{\max}(A)\|B^\top q^t - \tilde{g}^t\|^2 + (1 + \frac{\eta}{\beta})(1 - \eta\sigma_{\min}(Z))f(\theta^t) \\
&\overset{(iv)}{=} (\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s^2_{\max}(A)\|B^\top q^t - \tilde{g}^t\|^2 + (1 - \eta^2\sigma^2_{\min}(Z))f(\theta^t) \\
&\overset{(v)}{\leqslant} D^2(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s^2_{\max}(A)\|B^\top q^t - \tilde{g}^t\|^2_\infty + (1 - \eta^2\sigma^2_{\min}(Z))f(\theta^t) \\
&\overset{(vi)}{\leqslant} D^2H^2\|B^\top\|^2_\infty(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)})s^2_{\max}(A)(\tau\gamma^{t-1})^2 + (1 - \eta^2\sigma^2_{\min}(Z))f(\theta^t) \\
&\overset{(vii)}{\leqslant} \left(1 - \eta^2\sigma^2_{\min}(Z) + D^2H^2\|B^\top\|^2_\infty\left(\frac{1}{2}\eta^2 + \frac{\eta}{\sigma_{\min}(Z)}\right)s^2_{\max}(A)\lambda^2\tau^2\right)(\alpha)^t f(\theta^0) \\
&\overset{(viii)}{=} \left(1 - \frac{1}{\kappa^4} + D^2H^2\|B^\top\|^2_\infty\left(\frac{\sigma^2_{\min}(Z)}{2\sigma^4_{\max}(Z)} + \frac{1}{\sigma^2_{\max}(Z)}\right)s^2_{\max}(A)\lambda^2\tau^2\right)(\alpha)^t f(\theta^0) \\
&\overset{(ix)}{\leqslant} (1 - \frac{1}{2\kappa^4})(\alpha)^t f(\theta^0) = (\alpha)^{t+1} f(\theta^0),
\end{aligned}$$

where (i) we have explicitly expands the $f(\cdot)$ function, and have used (28); (ii) applies the Young inequality with constant β; (iii) extracts the largest singular value of A, and uses (28); (iv) set $\beta = \frac{1}{\sigma_{\min}(Z)}$; (v) uses the relationship between ℓ_2 and ℓ_∞ norm; (vi) uses the second inequality in induction assumption (23); (vii) uses the induction assumption (22) and the definition of γ^{t-1} in (29); (viii) plug in the choice of stepsize; (ix) comes from the choice of λ and τ.

Finally, we will show (23) holds for $t+1$ by induction, again base assumptions (22) holds for t and (23) holds for t. We have:

$$\begin{aligned}
\left\|g^{t+1} - q^t\right\|_\infty &\overset{(i)}{\leqslant} \|g^{t+1} - g^t\|_\infty + \|g^t - q^t\|_\infty \\
&\overset{(ii)}{\leqslant} \|BA^\top A(\theta^{t+1} - \theta^t)\| + \tau\gamma^{t-1} \\
&\overset{(iii)}{=} \eta\|BA^\top AB^\top q^t\| + \tau\gamma^{t-1}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{\leqslant} \eta\sigma_{\max}(Z)\|q^t\| + \tau\gamma^{t-1} \\
&\overset{(v)}{\leqslant} \eta\sigma_{\max}(Z)(\|q^t - g^t\| + \|g^t\|) + \tau\gamma^{t-1} \\
&\overset{(vi)}{\leqslant} \eta\sigma_{\max}(Z)(H\|q^t - g^t\|_\infty + \|g^t\|) + \tau\gamma^{t-1} \\
&\overset{(vii)}{\leqslant} \tau\gamma^{t-1}(1 + \eta H\sigma_{\max}(Z)) + \eta\sigma_{\max}(Z)\|g^t\| \\
&= \tau\gamma^{t-1}(1 + \eta H\sigma_{\max}(Z)) + \eta\sigma_{\max}(Z)\|BA^\top(A\theta^t - b)\| \\
&\overset{(viii)}{\leqslant} \tau\gamma^{t-1}(1 + \eta H\sigma_{\max}(Z)) + \eta\sigma_{\max}^{\frac{3}{2}}(Z)\sqrt{2f(\theta^t)} \\
&\overset{(ix)}{=} \tau\gamma^{t-1}(1 + \frac{H}{\kappa}) + \frac{\sqrt{2}\sigma_{\min}(Z)}{\sqrt{\sigma_{\max}(Z)}}\sqrt{f(\theta^t)} \\
&\overset{(x)}{\leqslant} \left(\tau\lambda(1 + \frac{H}{\kappa}) + \frac{\sqrt{2}\sigma_{\min}(Z)}{\sqrt{\sigma_{\max}(Z)}}\right)\sqrt{(\alpha)^t f(\theta^0)} \\
&= \lambda\left(\tau(1 + \frac{H}{\kappa}) + \frac{\sqrt{2}\sigma_{\min}(Z)}{\lambda\sqrt{\sigma_{\max}(Z)}})\right)\sqrt{(\alpha)^t f(\theta^0)} \\
&\overset{(xi)}{\leqslant} \frac{1}{3}\lambda\sqrt{(\alpha)^t f(\theta^0)} \overset{(xii)}{\leqslant} \lambda\sqrt{(\alpha)^{t+1} f(\theta^0)} = \gamma^t,
\end{aligned}
$$

where (i) uses the triangle inequality; (ii) expands the first term by (25), uses the relationship between ℓ_2 and ℓ_∞ norm, and and uses the first inequality in induction assumption (23); (iii) plug in the update of parameter: $\theta^{t+1} = \theta^t - \eta B^\top q^t$; (iv) comes from the fact that non-zero singular values of Z and $BA^\top AB^\top$ are the same and extracts the largest singular value of Z; (v) uses the triangle inequality; (vi) is due to the relationship between ℓ_2 and ℓ_∞ norm; (vii) uses the first inequality in induction assumption (23); (viii) uses the fact that $s_{\max}^2(BA^\top) = \sigma_{\max}(Z)$ and extracts $s_{\max}(BA^\top)$; (ix) plug in the choice of stepsize; (x) uses the definition of γ^{t-1} in (29) and induction assumption in (22); (xi) is because the choice of τ and λ; (xii) is because $\sqrt{\alpha} = \sqrt{1 - \frac{1}{2\kappa^2}} \geqslant \frac{\sqrt{2}}{2} > \frac{1}{3}$ since $\kappa \geqslant 1$. Thus, we obtain $\left\|g^{t+1} - q^t\right\|_\infty \leqslant \lambda\sqrt{(\alpha)^{t+1} f(\theta^0)} = \gamma^t$. Then by Lemma 4.1, with the correspondence that $c = g^{t+1}, p = q^t, r = \gamma^t$, we can obtain

$$
\|q^{t+1} - g^{t+1}\| \overset{(i)}{=} \|\text{quant}(g^{t+1}, q^t, \gamma^t, b) - g^{t+1}\|_\infty \overset{(ii)}{\leqslant} \tau\gamma^t,
$$

$$
\|B^\top q^{t+1} - \tilde{g}^{t+1}\|_\infty \overset{(iii)}{=} \|B^\top \text{quant}(g^{t+1}, q^t, \gamma^0, b) - B^\top g^{t+1}\|_\infty \overset{(iv)}{\leqslant} \tau D\|B^\top\|_\infty \gamma^t,
$$

where (i) and (iii) come from the 'Quantize' step in Algorithm 1; (ii) and (iv) are from the two relations in Lemma 4.1 (since we have proved that $\|g^{t+1} - q^t\|_\infty \leqslant \gamma^t$ holds).

Now we have proved that (14) and (15) hold. So for $t > 0$, there is

$$
f(\theta^t) \leqslant (\alpha)^t f(\theta^0), \quad \text{where} \quad \alpha = 1 - \frac{1}{2\kappa^4}.
$$

Thus, if we want the objective function to compute an ϵ-optimal solution, the total number of iterations is $\log(f(\theta^0)/\epsilon)/\log(1/(1 - \frac{1}{2\kappa^4}))$. Since in each iteration, each agent k transmits a length-H vector q_k^t, so we conclude that the total number of bits each node needs to communicate is $\log(f(\theta^0)/\epsilon)/\log(1/(1 - \frac{1}{2\kappa^4}))$ bits. Notice that $\log(1/(1 - \frac{1}{2\kappa^4})) = -\log(1 - \frac{1}{2\kappa^4}) \sim 2\kappa^4$, so we can derive the simplified total number of bits as $2\kappa^4 \cdot \log(f(\theta^0)/\epsilon)$.

E Proof for Proposition 1

Now we prove Proposition 1. We first state two lemmas that will be used.

Lemma E.1. *Rudelson and Vershynin [2010] Let X be a $H \times N$ matrix whose entries are independent standard normal random variables. Then*

$$
\mathbb{P}\left(\sqrt{H} - \sqrt{N} - t \leq s_{\min}(X) \leq s_{\max}(X) \leq \sqrt{H} + \sqrt{N} + t\right) \geq 1 - 2e^{-t^2/2}, \quad t \geq 0.
$$

Lemma E.2. *Suppose $A_i^\top \in \mathbb{R}^D$ follows $N(\mu, \mathbf{\Sigma})$. If $\mathbf{\Sigma}$ is a diagonal matrix, then each element of A_i is independent.*

We first write down the SVD decomposition of $A^\top$:

$$A^\top = V\Sigma W,$$

where $V \in \mathbb{R}^{D\times N}, \Sigma \in \mathbb{R}^{N\times N}, W \in \mathbb{R}^{N\times N}$. Since A is full rank, denote $\hat{\kappa} := \frac{s_{\max}(A^\top)}{s_{\min}(A^\top)} > 0$. Consider each entry in $BA^\top = BV\Sigma W$. It is clear that each entry of BV follows normal distribution because $(BV)_{hi}$ is linear combination of variables that are standard normal. Then we aim to show that each entry in BV follows standard normal distribution. Notice that $\text{vec}(BV) = (B_1V, B_2V, \cdots, B_HV)^\top \in \mathbb{R}^{HN}$, where B_h is each row of B. Denote $U_{ij} = \mathbf{Cov}(B_iV, B_jV) \in \mathbb{R}^{N\times N}$, we can obtain that

$$\mathbb{E}[\text{vec}(BV)] = (\mathbb{E}[B_1]V; \mathbb{E}[B_2]V; \cdots, \mathbb{E}[B_H]V)^\top = \mathbf{0},$$

$$\mathbf{Cov}(\text{vec}(BV)) = \begin{pmatrix} U_{11} & U_{12} & U_{13} & \cdots & U_{1N} \\ U_{21} & U_{22} & U_{23} & \cdots & U_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ U_{N1} & U_{N2} & U_{N3} & \cdots & U_{NN} \end{pmatrix}$$

Notice that $U_{ij} = \mathbf{Cov}(X^\top B_i^\top, X^\top B_j^\top) = X^\top \mathbf{Cov}(B_i^\top, B_j^\top)X$. Since $B_i^\top$ and $B_j^\top$ are independent, so $V_{ij} = \mathbf{0}, i \neq j$. Since $X^\top X = \mathbf{I}_N$, we have $U_{ij} = \mathbf{Cov}(B_i^\top, B_i^\top) = \mathbf{I}_N$ for $i = 1, 2, \cdots, N$. Then we obtain $\mathbf{Cov}(\text{vec}(BV)) = \mathbf{I}_{HN}$. From Lemma E.2, each entry in BV are independent. By Lemma E.1, we know the condition number of BV, which is $\kappa(BV)$ is independent of D, and we have

$$\mathbb{P}\left(\kappa(BV) \leqslant \frac{\sqrt{H} + \sqrt{N} + t}{\sqrt{H} - \sqrt{N} - t}\right) \geqslant 1 - 2e^{-t^2/2}, \quad t \geqslant 0.$$

Now we consider $Z = AB^\top BA^\top$. Notice the non-zero eigen values of Z are the same as $BA^\top AB^\top$, then we consider $BA^\top AB^\top$,

$$BA^\top AB^\top = BV\Sigma WW^\top \Sigma^\top V^\top B^\top = BV\Sigma^2 V^\top B^\top.$$

Notice the non-zero eigen values of $BV\Sigma^2V^\top B^\top$ are the same as $\Sigma^2 BVV^\top B^\top$, we can derive

$$\kappa(Z) = \kappa(\Sigma^2 BVV^\top B^\top) \leqslant \hat{\kappa}^2 \cdot \kappa^2(BV),$$

where $\hat{\kappa}$ is condition number of A and the inequality is because the property of square and invertible matrix. So we have with probability at least $1 - 2e^{-t^2/2}$,

$$\kappa(Z) \leqslant \hat{\kappa}^2 \left(\frac{\sqrt{H} + \sqrt{N} + t}{\sqrt{H} - \sqrt{N} - t}\right)^2.$$

F The Proof of Theorem 4.3 and Details about Algorithm 2

F.1 The CHOCO-GOSSIP Protocol

For completeness, we describe the CHOCO-GOSSIP protocol [Koloskova et al., 2019] for decentralized average consensus with compressed communication as follows. Notice the number of gossip rounds T_g is dependent to the error ϵ as discussed in Fact 1. Note that the protocol uses the compressor $Q(\cdot)$ for compressed communication, for example, this can be the randomized quantizer, see [Koloskova et al., 2019] for other examples. The protocol is summarized below:

Algorithm 3 CHOCO-GOSSIP

Input: step size γ; initial vectors $g_1^0, \dots, g_K^0$; gossip rounds T_g; compressor $Q(.)$; mixing matrix W; neighbor sets $\mathcal{N}_1, \dots, \mathcal{N}_K$.
Initialize: $\hat{g}_i^0 = \mathbf{0}$, $\forall i \in [K]$.
for t **in** $0, \dots, T_g - 1$ **do**
  Compress: Each agent i compress the difference, $q_i^t = Q(g_i^t - \hat{g}_i^t)$
  Communicate: Each agent i receives q_j^t from neighbor $j \in \mathcal{N}_i$ and update $\hat{g}_j^{t+1} = \hat{g}_j^t + q_j^t$
  Aggregation: Each agent i combines received vectors, i.e.,

$$g_i^{t+1} = g_i^t + \gamma \sum_{j \in \mathcal{N}_i} w_{ij}(\hat{g}_j^{t+1} - \hat{g}_i^{t+1})$$

end for
Output: $g_i^{T_g}$ at each agent i.

To derive the number of bits required in Theorem 4.3, we focus on using the random quantizer for $Q(.)$, i.e., for $x \in \mathbb{R}^d$, $s \in \mathbb{N}_+$ and $\tau = (1 + \min\{d/s^2, \sqrt{d}/s\})$, we have

$$Q(x) = \frac{\text{sign}(x) \cdot \|x\|}{s\tau} \cdot \left\lfloor s\frac{|x|}{\|x\|} + \xi \right\rfloor,$$

where $\xi \sim \text{Uniform}[0,1]^d$ and sending $Q(x)$ across the network requires $d\log(s+1) + d + 64$ bits of communication.

F.2 The Proof of Theorem 4.3

Since $F(x) = Bx$, $\tilde{F}(y) = B^\top y$, the consensus error can be expressed as $E_i^t := \bar{g}_i^t - \frac{1}{n}\sum_{j=1}^n B\nabla f_j(\theta_j^t)$. We also denote the deviation of locally computed gradient as $\Delta G_i^t := \frac{1}{n}\sum_{j=1}^n B(\nabla f_j(\theta_j^t) - \nabla f_j(\theta_i^t))$. Note that $\|E_i^t\| \le \bar{\epsilon}/(t+1)$. We first observe the updated iterate,

$$\begin{aligned}
\theta_i^{t+1} &= \theta_i^t - \eta B^\top \bar{g}_i^t - \frac{1}{n}\sum_{j=1}^n B^\top B \nabla f_j(\theta_j^t) + \frac{1}{n}\sum_{j=1}^n B^\top B \nabla f_j(\theta_j^t) \\
&= \theta_i^t - \eta \left[B^\top E_i^t + \frac{1}{n} B^\top B \sum_{j=1}^n \left(\nabla f_j(\theta_j^t) + \nabla f_j(\theta_i^t) - \nabla f_j(\theta_i^t)\right) \right] \\
&= \theta_i^t - \eta \left[B^\top E_i^t + B^\top \Delta G_i^t + B^\top B \nabla f(\theta_i^t) \right]
\end{aligned}$$

Next, we observe that the objective function value evolves as,

$$\begin{aligned}
f(\theta_i^{t+1}) &= \frac{1}{2}\sum_{j=1}^n \|A_j \theta_i^{t+1} - b_j\|^2 \\
&= \frac{1}{2}\sum_{j=1}^n \Big\{ \|A_j\theta_i^t - b_j\|^2 - 2\eta \left\langle A_j^\top (A_j\theta_i^t - b_j) \,\middle|\, B^\top E_i^t + B^\top \Delta G_i^t + B^\top B \nabla f(\theta_i^t) \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ \eta^2 \|A_j \left[B^\top E_i^t + B^\top \Delta G_i^t + B^\top B \nabla f(\theta_i^t)\right]\|^2 \Big\} \\
&\le f(\theta_i^t) - \eta \left\langle \nabla f(\theta_i^t) \,\middle|\, B^\top B \nabla f(\theta_i^t) \right\rangle + \frac{1}{2}\sum_{j=1}^n 2\eta^2 \|A_j B^\top B \nabla f(\theta_i^t)\|^2 \\
&+ \frac{1}{2}\sum_{j=1}^n \Big\{ -2\eta \left\langle A_j\theta_i^t - \boldsymbol{b}_j \,\middle|\, A_j B^\top (E_i^t + \Delta G_i^t) \right\rangle + 2\eta^2 \|A_j B^\top (E_i^t + \Delta G_i^t)\|^2 \Big\} \\
&\le (1 - 2\sigma_{\min}(Z)\eta + 2\sigma_{\max}^2(Z)\eta^2) f(\theta_i^t) \\
&+ \sum_{j=1}^n \Big\{ -\eta \left\langle A_j\theta_i^t - \boldsymbol{b}_j \,\middle|\, A_j B^\top (E_i^t + \Delta G_i^t) \right\rangle + \eta^2 \|A_j B^\top (E_i^t + \Delta G_i^t)\|^2 \Big\} \qquad (31)
\end{aligned}
$$

where the last inequality requires the observations

$$
\eta^2 \sum_{j=1}^n \|A_j B^\top B \nabla f(\theta_i^t)\|^2 = \eta^2 (A\theta_i^t - \boldsymbol{b})^\top (AB^\top BA^\top)^\top AB^\top BA^\top (A\theta_i^t - \boldsymbol{b}) \le 2\sigma_{\max}^2(Z)\eta^2 f(\theta_i^t),
$$

$$
\left\langle \nabla f(\theta_i^t) \,\middle|\, B^\top B \nabla f(\theta_i^t) \right\rangle \ge \sigma_{\min}(Z) f(\theta_i^t).
$$

It remains to deal with the error terms separately. Observe that

$$
\begin{aligned}
&\sum_{j=1}^n \eta \left\langle A_j\theta_i^t - b_j \,\middle|\, A_j B^\top (E_i^t + \Delta G_i^t) \right\rangle \\
&\le \eta \sum_{j=1}^n \left\{ \frac{\sigma_{\min}(Z)}{2} \|A_j\theta_i^t - b_j\|^2 + \frac{1}{2\sigma_{\min}(Z)} \|A_j B^\top (E_i^t + \Delta G_i^t)\|^2 \right\} \\
&= \sigma_{\min}(Z)\eta f(\theta_i^t) + \frac{\eta}{2\sigma_{\min}(Z)} \sum_{j=1}^n \|A_j B^\top (E_i^t + \Delta G_i^t)\|^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{j=1}^n \|A_j B^\top (E_i^t + \Delta G_i^t)\|^2 \\
&= \|AB^\top (E_i^t + \Delta G_i^t)\|^2 = (E_i^t + \Delta G_i^t)^\top BA^\top AB^\top (E_i^t + \Delta G_i^t) \\
&\le \sigma_{\max}(Z) \|E_i^t + \Delta G_i^t\|^2 \le 2\sigma_{\max}(Z)(\|E_i^t\|^2 + \|\Delta G_i^t\|^2)
\end{aligned}
$$

Putting together into (31) and setting $\eta = \frac{\sigma_{\min}(Z)}{4\sigma_{\max}^2(Z)}$ yields

$$
f(\theta_i^{t+1}) \le \left(1 - \frac{\sigma_{\min}(Z)}{8\sigma_{\max}(Z)}\right) f(\theta_i^t) + (\eta^2 + \frac{\eta}{2\sigma_{\min}(Z)}) \cdot 2\sigma_{\max}(Z)(\|E_i^t\|^2 + \|\Delta G_i^t\|^2). \qquad (32)
$$

To bound $\|\Delta G_i^t\|^2$, we observe that by the algorithm and the initial condition $\theta_i^0 = \theta_j^0$, for any $t \ge 0$,

$$
\begin{aligned}
\theta_j^{t+1} - \theta_i^{t+1} &= \theta_j^t - \theta_i^t - \eta(E_j^t - E_i^t) = -\eta \sum_{s=0}^t (E_j^s - E_i^s) \\
\|\theta_j^{t+1} - \theta_i^{t+1}\|^2 &\le 2\eta^2 \sum_{s=0}^t [\|E_j^s\|^2 + \|E_i^s\|^2] \le 4\eta^2 \sum_{s=0}^t (\frac{\bar{\epsilon}}{s+1})^2 \le \frac{2\eta^2\pi^2\bar{\epsilon}^2}{3}, \qquad (33)
\end{aligned}
$$

where the last inequality applied $\sum_{s=1}^\infty s^{-2} = \pi^2/6$, together with Fact 1 and the assumption $\epsilon_t = \bar{\epsilon}/(t+1)$. Then, the error ΔG_i^t can be bounded by

$$
\begin{aligned}
\|\Delta G_i^t\|^2 &= \frac{1}{n^2} \|\sum_{j=1}^n BA_j^\top A_j (\theta_j^t - \theta_i^t)\|^2 \\
&\le \frac{1}{n^2} \sum_{j=1}^n \|BA_j^\top A_j\|^2 \|\theta_j^t - \theta_i^t\|^2 \overset{(33)}{\le} \frac{2\eta^2\pi^2\bar{\epsilon}^2}{3n^2} \sum_{j=1}^n \|BA_j^\top A_j\|^2. \qquad (34)
\end{aligned}
$$

With the notation $\sigma_{BA^\top A} := \frac{1}{n}\sum_{j=1}^n \|BA_j^\top A_j\|^2$, (32) gives us

$$
\begin{aligned}
f(\theta_i^{t+1}) &\overset{(34)}{\leq} (1-\frac{1}{8\kappa^2})f(\theta_i^t) + \left(\eta^2 + \frac{\eta}{2\sigma_{\min}(Z)}\right)\cdot 2\sigma_{\max}^2(Z)\left(\epsilon_t^2 + \frac{2\eta^2\pi^2\bar{\epsilon}^2}{3n}\sigma_{BA^\top A}\right)\\
&\leq (1-\frac{1}{8\kappa^2})^{t+1}f(\theta_i^0) + \sum_{s=0}^{t}\left(1-\frac{\sigma_{\min}(Z)\eta}{2}\right)^{t-s}\left(\eta^2 + \frac{\eta}{2\sigma_{\min}(Z)}\right)\cdot 2\sigma_{\max}^2(Z)\left(\epsilon_s^2 + \frac{2\eta^2\pi^2\bar{\epsilon}^2}{3n}\sigma_{BA^\top A}\right)\\
&\leq (1-\frac{1}{8\kappa^2})^{t+1}f(\theta_i^0) + \frac{1}{6\sigma_{\min}^2(Z)}\left(1+\frac{\sigma_{\min}^2(Z)}{4n\sigma_{\max}^4(Z)}\sigma_{BA^\top A}\right)(\sigma_{\min}^2(Z)+2\sigma_{\max}^2(Z))\pi^2\bar{\epsilon}^2,
\end{aligned}
$$

where we have simplified notations by setting $\kappa = \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)}$. By adjusting $\bar{\epsilon}$, we can achieve dimension-independent linear convergence.

Communication Complexity Let $\xi_t = \sum_{i=1}^n \|g_i^t - \frac{1}{n}\sum_{j=1}^n g_j^t\|^2$. We recall from Fact 1 that achieving a consensus error of $\bar{\epsilon}/(t+1)$ at iteration t via `CHOCO-GOSSIP` requires

$$
\frac{82}{\delta^2\omega}\log\frac{(t+1)\xi_t}{\bar{\epsilon}} \quad \text{rounds of communication.}
$$

Applying b-bits random quantization to our projected H-dimensional vector, we have

$$
\frac{1}{\omega} = 1 + \min\{H/(2^b-1)^2, \sqrt{H}/(2^b-1)\}.
$$

Fix $\epsilon > 0$, to find an ϵ-optimal solution, we set

$$
\bar{\epsilon} := \sqrt{\epsilon\,\frac{3}{\pi^2(1+\kappa^2)}\left(1+\frac{1}{4n\kappa^2\sigma_{\max}^2(Z)}\sigma_{BA^\top A}\right)^{-1}}
$$

and $t \geq T_\epsilon := 8\kappa^2\log(2f(\theta_i^0)/\epsilon)$. Altogether, the number of communication rounds for achieving an ϵ optimal solution is bounded by:

$$
\begin{aligned}
\frac{82}{\delta^2\omega}\sum_{t=1}^{T_\epsilon}\log\frac{(t+1)\xi_t}{\bar{\epsilon}} &\leq \frac{82T_\epsilon}{\delta^2\omega}\left(\max_{t\in[T_\epsilon]}\log(\xi_t) + \log(T_\epsilon+1) + \log(1/\bar{\epsilon})\right)\\
&= \mathcal{O}\left(\frac{\kappa^2}{\delta^2\omega}\big(\log\frac{1}{\epsilon} + \log(1+\frac{1}{n\sigma_{\max}^2(Z)}\sigma_{BA^\top A})\big)\log\frac{1}{\epsilon}\right)
\end{aligned}
$$

where we have assumed that $\max_{t\in[T_\epsilon]}\log(\xi_t)$ is dominated by $\max\{\log(1/\bar{\epsilon}), \log(T_\epsilon)\}$.

Each communication round requires to send $\mathcal{O}((b+1)H)$ bits per agent. We can optimize the choice of b by

$$
\begin{aligned}
b^\star = \arg\min_b\ bH\omega^{-1} &= \arg\min_b\left\{\min\left\{bH + bH^2/(2^b-1)^2, bH + bH^{3/2}/(2^b-1)\right\}\right\}\\
\Rightarrow\ b^\star H\omega^{-1} &= \mathcal{O}(H\log H), \quad \text{with } b^\star = \log(H^{1/2}+1)
\end{aligned}
$$

Under the properties of A, B described in Proposition 1, by Lemma E.1, simplifying $\|BA_j^\top A_j\|$ shows that with probability $1-\zeta$,

$$
\|BA_j^\top A_j\| \leq \|A_j\|^2\|B\| \leq \left(\sqrt{H} + \sqrt{D} + \sqrt{2\log\frac{2}{\zeta}}\right)^3
$$

As such, we have $\sigma_{BA^\top A} \leq \left(\sqrt{H} + \sqrt{D} + \sqrt{2\log\frac{2}{\zeta}}\right)^6$. Putting together gives the communication complexity upper bound of $\mathcal{C}_{\text{op}}$ and the proof is completed.

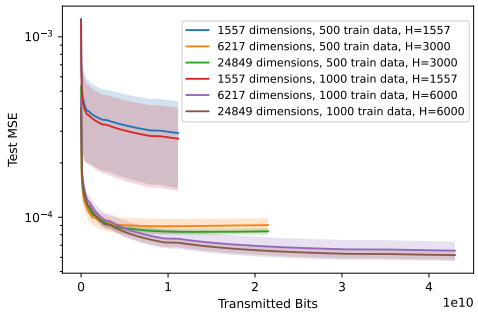

Figure 2: **Overparameterized Kernel Regression with Alg. 2.** Testing MSE against the number of bit transmitted on the whole network, averaged over 5 random seeded runs.

D	H	**Compression Ratio**
1557	1557	1.099
6217	3000	0.529
6217	6000	1.057
24849	3000	0.132
24849	6000	0.264

Table 1: **Communication Compression Ratio of Alg. 2** against vanilla DGD with double precision. 8-bits quantization is applied in `CHOCO-GOSSIP` and the average rounds of gossip T_g is 7.78.

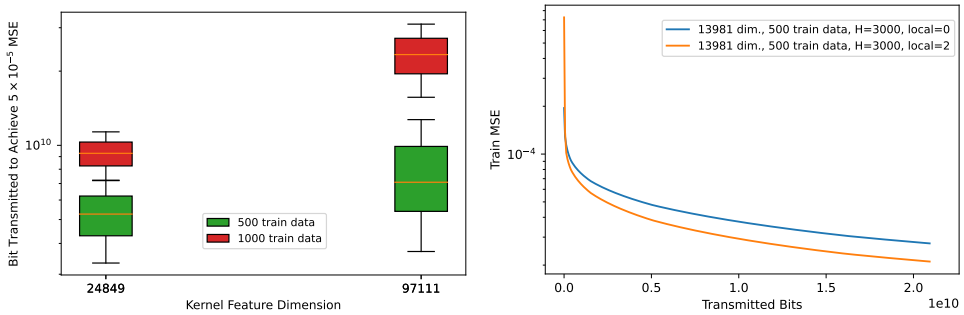

Figure 3: **Overparameterized Kernel Regression with Alg. 2.** (Left) Additional simulation results for large scale overparameterization (19×). (Right) Additional simulation results of Alg. 2 combining with local updates, compared to when no local updates are applied.

F.3 Additional Numerical Result

Fig. 2 shows the test MSE against communication cost for the cases of $N = 2500$ and $N = 5000$. The models are evaluated on the testing dataset of 23175 samples. Our result indicates that increasing the dimension of kernel features decreases the testing MSE. It suggests that overparameterization improves generalization on unseen data. The communication budget limits the number of training samples, thus bottlenecks the generalization accuracy. In the scenario where both computation and communication budgets are limited, Fig. 2 shows that the generalization power of Alg. 2 benefits more from spending into communication budget (thus more training samples) than from overparametrizing, under the situation where $H \ll D$ and more training samples are available with no cost. Table 1 shows the compression ratio for the schemes used in our experiments.

Lastly, Fig. 3 (left) provides additional results for the large-scale overparameterized scenario, where $\frac{D}{N}$ is as large as 19 which is close to the degree of over-parameterization in practice. Comparing the two cases with $D = 24,849$, $D = 97,111$ where the latter case uses 4 times more parameters, we observe that the bit complexity to reach an MSE of 10^{-5} only increases by 1.4 times. Fig. 3 (right) examines empirically the effects of running $L_{\text{local}} = 2$ local steps on the train MSE performance against the number of transmitted bits.

G The Proof of Theorem 4.4

G.1 Additional Notations

Let us further define some notations before we go into the details of the proof. Denote the parameter in t-th iteration as $\theta^t = (W_l^t)_{l=1}^L$; $P_l = \text{vec}(O_l)$ as the vectorized output from layer l; the pre-activation output from layer l as Q_l. Similarly, we can define $P_{l,k}, Q_{l,k}$ for each agent. Specifically, $O_L = P_L = Q_L$. Further, let us define some notations related to the singular values of the weight matrices.

$$\bar{\lambda}_l = \begin{cases} \frac{2}{3}\left(1 + s_{\max}(W_l^0)\right), & \text{for } l \in \{1,2\} \\ s_{\max}(W_l^0), & \text{for } l \in \{3,\dots,L\} \end{cases}, \ \underline{\lambda}_l = s_{\min}\left(W_l^0\right), \ \underline{\lambda}_{i\to j} = \prod_{l=i}^{j} \underline{\lambda}_l, \ \bar{\lambda}_{i\to j} = \prod_{l=i}^{j} \bar{\lambda}_l. \tag{35}$$

Specifically, $\bar{\lambda}_l$ represents the quantity related to the largest singular value of weight matrices for different layers; $\underline{\lambda}_l$ represents smallest singular value of weight matrices for different layers.

Further, denote $\lambda_O = s_{\min}\left(\tilde{B} \cdot a\left(XW_1^0\right)\right)$ as the smallest singular value of the output from first hidden layer at initialization multiplied by $\tilde{B}$. Let us define some notations related to gradient. Recall we have defined $u_l = \text{vec}(\nabla_{W_l} f(\theta))$. Let us define the per iteration gradient as: $g_k^t = (Bu_{2,k}^t, u_{3,k}^t; \cdots; u_{L,k}^t)$ that collects all the (compressed) gradient of each layer. Further, let us denote $\tilde{u}_{2,k}^t = B^\top B u_{2,k}^t$, $\tilde{g}_k^t = (\tilde{u}_{2,k}^t; u_{3,k}^t; \cdots; u_{L,k}^t)$, $q_k^t = (z_{2,k}^t; z_{3,k}^t \cdots; z_{L,k}^t)$, $\tilde{q}_k^t = (B^\top z_{2,k}^t; z_{3,k}^t \cdots; z_{L,k}^t)$, where $z_{2,k}^t$ quantizes $Bu_{2,k}^t$ and $z_{l,k}^t$ quantizes $u_{l,k}^t$ for $l \geqslant 3$; denote $\tilde{z}_{2,k}^t = B^\top z_{2,k}^t$. Note, that in the above notation, both quantities with $\tilde{\cdot}$ has the original dimension D. Further, $\tilde{q}_k^t$ is the actual vector that gets used at k-th agent, while $\tilde{g}_k^t$ represents a "virtual" vector which has not been quantized.

Now let us define:

$$\begin{aligned}
u_l^t &:= \sum_{k=1}^K u_{l,k}^t, \quad u^t = (u_2^t, u_3^t, \cdots, u_L^t), \\
g^t &:= \sum_{k=1}^K g_k^t = (Bu_2^t; u_3^t; \cdots; u_L^t), \ \tilde{g}^t := \sum_{k=1}^K \tilde{g}_k^t = (\tilde{u}_2^t; u_3^t; \cdots; u_L^t), \\
q^t &:= \sum_{k=1}^K q_k^t = (z_2^t; z_3^t; \cdots; z_L^t), \ \tilde{q}^t := \sum_{k=1}^K \tilde{q}_k^t = (\tilde{z}_2^t; z_3^t; \cdots, z_L^t).
\end{aligned}$$

Further, we define $\Delta W_l^t, \tilde{\Delta} W_l^t$, which are the unvectorized $\tilde{u}_2^t$ and $\tilde{q}_2^t$ for $l = 2$ and unvectorized $\tilde{u}_l^t$ and $\tilde{q}_l^t$ for $l \geqslant 3$:

$$\text{vec}(\Delta W_l^t) = \begin{cases} \tilde{u}_2^t & l = 2 \\ u_l^t & l \geqslant 3, \end{cases} \qquad \text{vec}(\tilde{\Delta} W_l^t) = \begin{cases} \tilde{z}_2^t & l = 2 \\ z_l^t & l \geqslant 3. \end{cases} \tag{36}$$

That is, $\Delta W_l^t, \tilde{\Delta} W_l^t \in \mathbb{R}^{n_{l-1} \times n_l}$.

Using the above notation, the 'Update' step in Algorithm 1, it can be expressed as

$$\theta^{t+1} = \theta^t - \eta \tilde{q}^t. \tag{37}$$

Denote $\Sigma_l^t = \text{diag}\left[\text{vec}\left(a'\left(Q_l^t\right)\right)\right] \in \mathbb{R}^{Nn_l \times Nn_l}$, which is a diagonal matrix, whose diagonal entries are the vectorized gradient of the activation function in each layer. Recall that we have defined $B = \mathbf{I}_{n_2} \otimes \tilde{B}$ in Section 4.2, where $\tilde{B}$ is a Gaussian random matrix of size $H \times n_1$. For convenience, let us assume $s_{\max}(B) = s_{\max}(\tilde{B}) \geqslant 1$.

G.2 Useful Lemma

Now we first state a collection of results from [Nguyen and Mondelli, 2020] that will be used in our proof. Notice that in the aforementioned work, the same pyramidal neural network structure and the l_2 loss function are used, so the loss function is the same as (6). It follows that all the properties of the loss function can be reused.

Lemma G.1. *Suppose Assumption 4 and 5 hold, for each θ^t, we have the following relations:*

1. $$u_l^t = \left(\mathbf{I}_{n_l} \otimes O_{l-1}^{t\top}\right) \prod_{p=l+1}^{L} \Sigma_{p-1}^t \left(W_p^t \otimes \mathbf{I}_N\right)(O_L^t - y), \tag{38}$$

2. $$\frac{\partial O_L^t}{\partial \operatorname{vec}(W_l)} = \prod_{p=0}^{L-l-1} \left(W_{L-p}^{t\top} \otimes \mathbf{I}_N\right) \Sigma_{L-p-1}^t \left(\mathbf{I}_{n_l} \otimes F_{l-1}^t\right), \tag{39}$$

3. $$\left\|u_2^t\right\| \geqslant s_{\min}\left(O_1^{t\top}\right) \prod_{p=3}^{L} s_{\min}\left(\Sigma_{p-1}^t\right) s_{\min}\left(W_p^t\right) \left\|O_L^t - y\right\|, \tag{40}$$

4. $$\left\|u_l^t\right\| \leqslant \|X\|_F \prod_{\substack{p=1\\p\neq l}}^{L} s_{\max}(W_p^t) \left\|O_L^t - y\right\|, \tag{41}$$

5. $$\|g^t\| \leqslant s_{\max}(\tilde{B}) L \|X\|_F \frac{\prod_{l=1}^{L} s_{\max}(W_l^t)}{\min_{l\in[L]} s_{\max}(W_l^t)} \|O_L^t - y\|. \tag{42}$$

Furthermore, given with θ^a and θ^b, if $\bar{\Lambda}_l \geq \max\left(s_{\max}\left(W_l^a\right), s_{\max}\left(W_l^b\right)\right)$ for some scalars $\bar{\Lambda}_l$. Let $\tilde{R} := \prod_{p=1}^{L} \max\left(1, \bar{\Lambda}_p\right)$. Then, for $l \in [L]$,

6. $$\left\|O_L^a - O_L^b\right\|_F \leqslant \sqrt{L}\|X\|_F \frac{\prod_{l=1}^{L} \bar{\Lambda}_l}{\min_{l\in[L]} \bar{\Lambda}_l} \left\|\theta^a - \theta^b\right\|, \tag{43}$$

7. $$\left\|\frac{\partial \operatorname{vec}\left(O_L(\theta^a)\right)}{\partial \operatorname{vec}\left(W_l^a\right)} - \frac{\partial \operatorname{vec}(O_L\left(\theta^b\right))}{\partial \operatorname{vec}\left(W_l^b\right)}\right\|_2 \leq \sqrt{L}\|X\|_F \tilde{R}\left(1 + L\rho\|X\|_F\tilde{R}\right)\left\|\theta^a - \theta^b\right\|. \tag{44}$$

8. $$\begin{aligned}&\left\|\operatorname{vec}\left(\nabla f(\theta^a)\right) - \operatorname{vec}\left(\nabla f(\theta^b)\right)\right\| \\ &\leqslant \left(L\sqrt{L}\|X\|_F^2 \frac{\prod_{l=1}^{L} \bar{\Lambda}_l^2}{\min_{l\in[L]} \bar{\Lambda}_l^2} + L\sqrt{L}\|X\|_F\tilde{R}(1 + L\rho\|X\|_F\tilde{R})\right)\|\theta^a - \theta^b\|,\end{aligned} \tag{45}$$

Now let us discuss the properties above one by one. The relations (38) and (39) show how the vectorized gradients of each layer, as well as the vectorized gradients of the output over each layer are computed, which are true regardless of the algorithm; see the first and the second equalities in [Nguyen and Mondelli, 2020, Lemma 4.1], respectively; (40) is the lower bound of the norm of the vectorized gradient over W_2, which is true as long as the network has the pyramidal structure (and regardless of the algorithm); see the third relation in [Nguyen and Mondelli, 2020, Lemma 4.1]; (41) is the upper bound of the norm of vectorized gradient over W_l, which is true regardless of the algorithm; see the third relation in [Nguyen and Mondelli, 2020, Lemma 4.2]; (42) can be derived by summing over l in (41). To be specific, we have

$$\begin{aligned}\|g^t\| &\leqslant \|Bu_2^t\| + \sum_{l=3}^{L}\|u_l^t\| \leqslant s_{\max}(\tilde{B})\|u_2^t\| + \sum_{l=3}^{L}\|u_l^t\| \\ &\leqslant s_{\max}(\tilde{B})\sum_{l=2}^{L}\|u_l^t\| \\ &\leqslant s_{\max}(\tilde{B})\sum_{l=2}^{L}\|X\|_F \prod_{\substack{p=1\\p\neq l}}^{L} s_{\max}(W_p)\|O_L - y\|\end{aligned}$$

$$\leqslant s_{\max}(\tilde{B})L\|X\|_F \frac{\prod_{l=1}^{L} s_{\max}(W_l^t)}{\min_{l\in[L]} s_{\max}(W_l^t)} \|O_L^t - y\|.$$

The relation in (43) gives the upper bound of the gap between output layer with two sets of parameters θ^a and θ^b, and it is regardless of the algorithm; see [Nguyen and Mondelli, 2020, Eq. (19)]; Similarly, (44) states the gap between the vectorized Jacobian matrix over W_l; see [Nguyen and Mondelli, 2020, Eq. (20)]; Finally, (45) computes the Lipschitz constants for the gradient, it is independent of the algorithm, and comes from plugging in (42), (43) and (44) into [Nguyen and Mondelli, 2020, Eq. (22)].

G.3 Initialization Strategy

As stated in Theorem 4.4, special initialization strategy is required. Now let us describe the initialization in detail. Recall the definition of $\bar{\lambda}_l$ and $\underline{\lambda}_l$ in (35) and $\lambda_O = s_{\min}\big(\tilde{B} \cdot a(XW_1^0)\big)$. Initialize θ^0 such that the following holds:

$$\lambda_O^2 \geqslant \frac{48 \cdot 6^{L-2}\|X\|_F \bar{\lambda}_{1\to L} s_{\max}^2(\tilde{B})\|O_L^0 - y\|}{(\nu^2)^{L-2}\underline{\lambda}_{3\to L}^2} \max\left(\frac{1}{\bar{\lambda}_2}, \max_{l\geqslant 3} \frac{2}{\bar{\lambda}_l \underline{\lambda}_l}\right). \tag{46}$$

To satisfy the above relation, it requires that $\|O_L^0 - y\|$ cannot be large, which means the θ^0 should not be far from the optimal solution so that the initial loss is small. Further, it requires the gap between $\bar{\lambda}_l$ and $\underline{\lambda}_l$ is small. From (45), we know that small $\bar{\lambda}_l$ induces small Lipschitz constant for gradient, so the condition requires that the Lipschitz constant is not very large, which further implies that the optimization landscape is smooth. On the other hand, from (40), we know the lower bound of the norm of u_2^t is related to $\underline{\lambda}_l$. So the initialization condition guarantees the lower bound of $\|u_2^t\|$ is not too small, which avoids vanished gradient.

The initialization can be realized by using the procedure suggested [Nguyen and Mondelli, 2020, Sec. 3.1]. The idea is that we can scale up W_1^0 to make λ_O not too small, and then randomly choose W_2^0 with small entries. Then for all $l \geqslant 3$, set W_l^0 as scaled identity matrices (top block as scaled identities) with large entries.

G.4 Formal Statement and Proof of Theorem 4.4

First, let us state the formal Theorem 4.4:

Theorem 4.4 *Consider using Alg. 1 to solve the problem (6), with X being full row rank. Suppose θ^0 is initialized as (46). Choose $\tilde{B}$ such as $\operatorname{rank}(\tilde{B}O_1^{0\top}) = N$; choose $\tilde{F}(\cdot)$ and $F(\cdot)$ as in (7). Set stepsize η and bits number b as following:*

$$\eta = \frac{\phi}{Q_0\Lambda^2},\ b = \max\big(\log(\frac{1}{\tau}+1), b_0\big),$$

where ϕ is defined in (47), τ is defined in (50), Λ is defined in (48), Q_0 is defined in (49). The following hold true:

$$f(\theta^{t+1}) \leqslant (1 - \frac{1}{4}\eta \cdot \phi) f(\theta^t).$$

To compute an ϵ-optimal solution (6), each agent is required to transmit:

$$\frac{4b}{\eta\phi} \cdot (Hn_2 + \sum_{l=2}^{L-1} n_l n_{l+1}) \log(f(\theta^0)/\epsilon) \quad \textit{bits/agent}.$$

Proof. To begin with, let us set the following constants:

$$\alpha = 1 - \frac{1}{4}\eta\phi, \quad \gamma^t = \lambda\sqrt{(\alpha)^{t+1} f(\theta^0)}, \quad \eta = \frac{\phi}{Q_0\Lambda^2}, \quad \lambda = 6\sqrt{2}s_{\max}(\tilde{B})\eta Q_0 \Lambda, \tag{47}$$

$$C_1 := \frac{\phi}{16\sqrt{2}Hn_1n_2^2\|\tilde{B}\|_\infty\lambda\|O_L^0 - y\|}, \quad C_2 := \frac{\phi \min\limits_{l\geqslant 3}\underline{\lambda}_l}{32\sqrt{2}\lambda n_2 n_3\|O_L^0 - y\|},$$

$$C_3 := \frac{1}{\sqrt{2}\eta Hn_1n_2^2\|\tilde{B}\|_\infty\lambda\|O_L^0 - y\|}, \quad C_4 := \frac{\min\limits_{l\geqslant 3}\underline{\lambda}_l}{2\sqrt{2}\eta\lambda n_2 n_3\|O_L^0 - y\|}$$

$$C_5 := \sqrt{\frac{\phi}{4\eta\left(\frac{1}{2}+\frac{1}{\beta}\right)\left(Hn_1n_2^2\|\tilde{B}\|_\infty^2 + \sum\limits_{l=3}^{L} n_{l-1}^2 n_l^2\right)\Lambda^2\lambda^2}},$$

$$C_6 := \frac{\sqrt{1-\frac{1}{4}\eta\phi} - \frac{1}{6}}{1+\eta s_{\max}(\tilde{B})Q_0\left(Hn_2 + \sum\limits_{l=3}^{L} n_{l-1}n_l\right)},$$

where

$$\phi = \left(\frac{1}{2}\nu\right)^{L-2}\lambda_{O\underline{\lambda}_{3\to L}}^2, \quad \Lambda = \left(\frac{3}{2}\right)^{L-1}\cdot Ls_{\max}^2(\tilde{B})\|X\|_F\frac{\bar{\lambda}_{1\to L}}{\min\limits_{l\geqslant 2}\bar{\lambda}_l} \tag{48}$$

$$Q_0 = L\sqrt{L}\left(\frac{3}{2}\right)^{2(L-1)}\|X\|_F^2\frac{\prod_{l=1}^{L}\bar{\lambda}_l^2}{\min\limits_{l\in[L]}\bar{\lambda}_l^2} + L\sqrt{L}\|X\|_F R\left(1+L\rho\|X\|_F R\right), \tag{49}$$

$$R = s_{\max}^2(\tilde{B})\prod_{p=1}^{L}\max\left(1, \frac{3}{2}\bar{\lambda}_p\right).$$

Define

$$\tau := \min(C_1, C_2, C_3, C_4, C_5, C_6). \tag{50}$$

Further, let us define

$$\Lambda_l := \left(\frac{3}{2}\right)^{L-1}\|X\|_F\frac{\bar{\lambda}_{1\to L}}{\bar{\lambda}_l}. \tag{51}$$

For convenience, let us assume $Q_0 > 1$. Let us explain the above constants: α is the constants based on which the objective function contracts; η is the stepsize; ϕ is related to the lower bound of $\|u_2^t\|$; Λ is related to the upper bound of $\|\tilde{g}^t\|$; Q_0 is the Lipschitz constant for the full gradient; τ is the parameter related to quantization; Λ_l is related to the upper bound for $\|u_l^t\|$.

The majority of the proof consists of showing the following relations by induction:

$$\begin{cases} s_{\max}\left(W_l^t\right) \overset{(i)}{\leqslant} \frac{3}{2}\bar{\lambda}_l, l = \{2, 3, \cdots, L\}, \\ s_{\min}\left(W_l^t\right) \overset{(ii)}{\geqslant} \frac{1}{2}\underline{\lambda}_l,\ l \in \{3, \ldots, L\}, \\ s_{\max}(W_2^t + \eta\tilde{\Delta}W_2^{t-1} - \eta\Delta W_2^{t-1}) \overset{(iii)}{\leqslant} \frac{3}{2}\bar{\lambda}_l, l = \{2, 3, \cdots, L\}, \\ s_{\min}(W_2^t + \eta\tilde{\Delta}W_2^{t-1} - \eta\Delta W_2^{t-1}) \overset{(iv)}{\geqslant} \frac{1}{2}\underline{\lambda}_l, l = \{3, \cdots, L\}, \\ f(\theta^{t-1} - \eta\tilde{g}^{t-1}) \overset{(v)}{\leqslant} (1-\eta\phi)f(\theta^{t-1}), \\ f(\theta^t) \overset{(vi)}{\leqslant} (\alpha)^t f(\theta^0), \\ \|Bu_2^t - z_2^t\|_\infty \overset{(vii)}{\leqslant} \tau\gamma^{t-1} = \tau\lambda\sqrt{(\alpha)^t f(\theta^0)}, \\ \|u_l^t - z_l^t\|_\infty \overset{(viii)}{\leqslant} \tau\gamma^{t-1} = \tau\lambda\sqrt{(\alpha)^t f(\theta^0)}. \end{cases} \tag{52}$$

Let us explain the meanings of the above relations. Relations (i) and (ii) provide the upper and lower bounds of the singular values of weight matrices in each iteration; (iii) and (iv) give the upper and lower bound of the singular values of weight matrices after one step update without quantization; (v) shows the decrease of loss function after one step of update without quantization; (vi) shows the

linear decrease of loss function in each iteration; (vii) and (viii) provide the error bound of gradient after quantization.

Compared to the induction proof in [Nguyen and Mondelli, 2020], the key challenges are: (1) Our analysis includes the quantization of gradient; (2) The update of parameter is not the simple gradient but a function of gradient.

Our proof consists of two steps:

Step 1: We show the above relations hold for $t = 1$.

Step 2: We show that if the above relations hold for t, then they hold for $t + 1$.

We show Step 1 first. For simplicity, we set $g^0 = q^0$, it is easy to verify the relations hold if we choose q^0 in Algorithm 1 with large enough C. Step 1 will be shown in the following five substeps: (a) Show (i) and (ii) in (52); (b) Show (iii) and (iv) in (52); (c) Show (v) in (52); (d) Show (vi) in (52); (e) Show (vii) and (viii) in (52).

(Step 1.a) We will show

$$\begin{cases} s_{\max}\left(W_l^1\right) \leqslant \frac{3}{2}\bar{\lambda}_l \;\; l \in [L],\; l \in \{2, 3, \ldots, L\} \\ s_{\min}\left(W_l^1\right) \geqslant \frac{1}{2}\underline{\lambda}_l,\; l \in \{3, \ldots, L\}. \end{cases}$$

We will use the fact that, according to the update rule in (37), we have $q^0 = g^0$.

For $l = 2$, we have:

$$\begin{aligned} &\|W_2^1 - W_2^0\|_F \overset{(i)}{=} \eta\|\tilde{z}_2^0\| = \eta\|B^\top B u_2^0\| \overset{(ii)}{\leqslant} \eta s_{\max}^2(\tilde{B})\|u_2^0\| \\ &\overset{(iii)}{\leqslant} \eta s_{\max}^2(\tilde{B})\Lambda_2\|O_L^0 - y\| \\ &\overset{(iv)}{\leqslant} \sum_{t=0}^{\infty} \eta s_{\max}^2(\tilde{B})\Lambda_2\|O_L^0 - y\| \cdot \sqrt{\alpha}^t \\ &\leqslant \eta s_{\max}^2(\tilde{B})\Lambda_2\|O_L^0 - y\|\frac{1}{1-\alpha}(1 + \alpha^{\frac{1}{2}}) \\ &\overset{(v)}{\leqslant} \frac{8 s_{\max}^2(\tilde{B})\Lambda_2}{\phi}\|O_L^0 - y\| \\ &\overset{(vi)}{\leqslant} \frac{1}{2}, \end{aligned}$$

where (i) is from the update rule in (37); (ii) extracts the largest singular value of B; (iii) uses the upper bound of the gradient norm in (41) and the definition of Λ_2 in (51) ; (iv) uses the fact that $\sum_{t=0}^{\infty} \sqrt{\alpha} > 1$; (v) plugs in the definition of α and $1 + \alpha^{\frac{1}{2}} < 2$; (vi) is because the initialization strategy in (46).

Next, we show the case where $l \geqslant 3$. We have

$$\begin{aligned} &\|W_l^1 - W_l^0\|_F \overset{(i)}{=} \eta\|u_l^0\| \overset{(ii)}{\leqslant} \eta\Lambda_l\|u_l^0\| \\ &\overset{(iii)}{\leqslant} \sum_{t=0}^{\infty} \eta\Lambda_l\|O_L^0 - y\| \cdot \sqrt{\alpha}^t \\ &= \eta\Lambda_l\|O_L^0 - y\|\frac{1}{1-\alpha^{\frac{1}{2}}} \\ &\leqslant \eta\Lambda_l\|O_L^0 - y\|\frac{1}{1-\alpha}(1 + \alpha^{\frac{1}{2}}) \\ &\overset{(iv)}{\leqslant} \frac{8\Lambda_l}{\phi}\|O_L^0 - y\| \\ &\overset{(v)}{\leqslant} \frac{1}{4}\underline{\lambda}_l < \frac{1}{2}\underline{\lambda}_l, \end{aligned}$$

where (i) is because the update of parameter in (37); (ii) uses the upper bound of the gradient norm in (41) and definition of Λ_l in (51); (iii) uses the fact that $\sum_{t=0}^{\infty} \sqrt{\alpha} > 1$; (iv) plugs in the choice of α and $1 + \alpha^{\frac{1}{2}} < 2$; (v) comes from the initialization strategy in (46).

Applying Weyl' inequality to the matrices W_l^0 and $(W_l^1 - W_l^0)$, we have

$$\begin{cases} s_{\max}(W_l^1) \leqslant \bar{\lambda}_l + \frac{1}{2}\bar{\lambda}_l = \frac{3}{2}\bar{\lambda}_l, \ l \in \{3, \cdots, L\}, \\ s_{\max}(W_l^1) \leqslant \bar{\lambda}_l + 1 = \frac{3}{2}\bar{\lambda}_l, \ l = 2, \\ s_{\min}(W_l^1) \geqslant \underline{\lambda}_l - \frac{1}{2}\bar{\lambda}_l = \frac{1}{2}\underline{\lambda}_l, \ l \in \{3, \ldots, L\}. \end{cases}$$

This concludes the proof of this substep.

(Step 1.b) We will show

$$\begin{cases} s_{\max}(W_l^1 + \eta\tilde{\Delta}W_l^0 - \eta\Delta W_l^0) \leqslant \frac{3}{2}\bar{\lambda}_l, l = \{2, 3, \cdots, L\}, \\ s_{\min}(W_l^1 + \eta\tilde{\Delta}W_l^0 - \eta\Delta W_l^0) \geqslant \frac{1}{2}\underline{\lambda}_l, l = \{3, \cdots, L\}. \end{cases}$$

Since we have $\tilde{q}^0 = \tilde{g}^0$, it is easy to derive that

$$\|\eta\tilde{\Delta}W_l^0 - \eta\Delta W_l^0\|_F = \eta\|\tilde{\Delta}W_l^0 - \eta\Delta W_l^0\| = \eta\|\tilde{q}^0 - \tilde{g}^0\| = 0.$$

Thus we can conclude

$$\begin{cases} s_{\max}(W_l^1 + \eta\tilde{\Delta}W_l^0 - \eta\Delta W_l^0) = s_{\max}(W_l^1) \leqslant \frac{3}{2}\bar{\lambda}_l, l = \{2, 3, \cdots, L\} \\ s_{\min}(W_l^1 + \eta\tilde{\Delta}W_l^0 - \eta\Delta W_l^0) = s_{\max}(W_l^1) \geqslant \frac{1}{2}\underline{\lambda}_l, l = \{3, \cdots, L\} \end{cases}$$

This concludes the proof for this substep.

(Step 1.c) We will show

$$f(\theta^0 - \eta\tilde{g}^0) \leqslant (1 - \eta\alpha_0)f(\theta^0).$$

From (Step 1.a) and (Step 1.b) we know

$$\max\left(s_{\max}(W_l^0), s_{\max}(W_l^1)\right) \leqslant \frac{3}{2}\bar{\lambda}_l.$$

Using the above relation, we can upper bound the differences of the gradients by using (45). More specifically, using the the definition of Q_0 in (49), we have

$$\|\text{vec}(\nabla f(\theta^1)) - \text{vec}(\nabla f(\theta^0))\| \leqslant Q_0\|\theta^1 - \theta^0\|, \tag{53}$$

where we repeat the definition of Q_0 below (where R is defined in (49)):

$$Q_0 = L\sqrt{L}\left(\frac{3}{2}\right)^{2(L-1)} \|X\|_F^2 \frac{\prod_{l=1}^{L} \bar{\lambda}_l^2}{\min_{l\in[L]} \bar{\lambda}_l^2} + L\sqrt{L}\|X\|_F R\left(1 + L\rho\|X\|_F R\right),$$

Further, it is easy to verify that for any $\hat{\theta}^0$ between θ^0 and θ^1, we still have $s_{\max}(W_l(\hat{\theta}^0)) \leqslant \frac{3}{2}\bar{\lambda}_l$. So we can apply the same argument leading to (53), and obtain:

$$\|\text{vec}(\nabla f(\hat{\theta}^0)) - \text{vec}(\nabla f(\theta^0))\| \leqslant Q_0\|\hat{\theta}^0 - \theta^0\|, \forall\hat{\theta}^0 = \theta^0 + \delta(\theta^1 - \theta^0), \delta \in [0, 1]. \tag{54}$$

As long as for any $\hat{\theta}^0$ (54) holds, we can apply Decent Lemma. More specifically, we have

$$\begin{aligned} f(\theta^0 - \eta\tilde{g}^0) &\overset{(i)}{\leqslant} f(\theta^0) - \eta\langle u^0, \tilde{g}^0\rangle + \frac{Q_0}{2}\eta^2\|\tilde{g}^0\|^2 \\ &\overset{(ii)}{=} f(\theta^0) - \eta\langle u_2^0, \tilde{u}_2^0\rangle - \eta\sum_{l=3}^{L}\|u_l^0\|^2 + \frac{Q_0}{2}\eta^2\|\tilde{g}^0\|^2 \\ &\leqslant f(\theta^0) - \eta\langle u_2^0, \tilde{u}_2^0\rangle + \frac{Q_0}{2}\eta^2\|\tilde{g}^0\|^2, \end{aligned} \tag{55}$$

where (i) uses Decent Lemma, (ii) uses the fact that u^0 is the stacked version of $\{u_i^0\}_{i\geq 2}$, and $\tilde{g}^0$ are stacked versions of $\{u_i^0\}_{i\geq 3}$ and $\tilde{u}_2^0$.

Next, we will bound the inner product on the right hand side of the above relation:

$$
\begin{aligned}
\langle u_2^0, \tilde{u}_2^0 \rangle &\overset{(i)}{=} \left\langle \left(\mathbf{I}_{n_2} \otimes O_1^{0\top}\right) \prod_{l=3}^{L} \Sigma_{l-1}^0 \left(W_l^0 \otimes \mathbf{I}_N\right) \left(O_L^{0\top} - y\right), B^\top B \left(\mathbf{I}_{n_2} \otimes O_1^{0\top}\right) \prod_{l=3}^{L} \Sigma_{l-1}^0 \left(W_l^0 \otimes \mathbf{I}_N\right) \left(O_L^{0\top} - y\right) \right\rangle \\
&\overset{(ii)}{=} \left\langle \left(\mathbf{I}_{n_2} \otimes O_1^{0\top}\right) \prod_{l=3}^{L} \Sigma_{l-1}^0 \left(W_l^0 \otimes \mathbf{I}_N\right) \left(O_L^{0\top} - y\right), \left(\mathbf{I}_{n_2} \otimes \tilde{B}^\top \tilde{B} O_1^{0\top}\right) \prod_{l=3}^{L} \Sigma_{l-1}^0 \left(W_l^0 \otimes \mathbf{I}_N\right) \left(O_L^{0\top} - y\right) \right\rangle \\
&\overset{(iii)}{=} \left\| \left(\mathbf{I}_{n_2} \otimes \tilde{B} O_1^{0\top}\right) \prod_{l=3}^{L} \Sigma_{l-1}^0 \left(W_l^0 \otimes \mathbf{I}_N\right) \left(O_L^0 - y\right) \right\|^2 \\
&\geqslant s_{\min}^2(\tilde{B} O_1^{0\top}) \prod_{l=3}^{L} s_{\min}^2 \left(\Sigma_{l-1}^0\right) s_{\min}^2 \left(W_l^0\right) \left\| O_L^0 - y \right\|^2 \\
&\overset{(iv)}{\geqslant} (\frac{1}{2}\nu)^{2(L-2)} \lambda_O^2 \underline{\lambda}_{3\to L}^2 \|O_L^0 - y\|^2,
\end{aligned} \tag{56}
$$

where (i) uses the expression of u_2^0 by (38) and $\tilde{u}_2^0 = B^\top B u_2^0$; (ii) and (iii) are from the property of Kronecker product; (iv) is from Assumption 5 and fact that $s_{\max}(W_l^0) \geqslant \frac{1}{2}\underline{\lambda}_l$.

Next, we upper bound the term $\|\tilde{g}^0\|^2$ in (55). We have

$$
\begin{aligned}
\|\tilde{g}^0\|^2 &\overset{(i)}{=} \|B^\top B u_2^0\|^2 + \sum_{l=3}^{L} \|u_l^0\|^2 \overset{(ii)}{\leqslant} s_{\max}^4(\tilde{B}) \|u_2^0\|^2 + \sum_{l=3}^{L} \|u_l^0\|^2 \\
&\overset{(iii)}{\leqslant} s_{\max}^2(\tilde{B}) (\|B u_2^0\|^2 + \sum_{l=3}^{L} \|u_l^0\|^2) = s_{\max}^2(\tilde{B}) \|g^0\|^2 \\
&\overset{(iv)}{\leqslant} s_{\max}^2(\tilde{B}) \left(L s_{\max}\left(\tilde{B}\right) \|X\|_F \frac{\prod_{l=1}^{L} s_{\max}\left(W_l^0\right)}{\min_l s_{\max}\left(W_l^0\right)} \right)^2 \left\| O_L^0 - y \right\|_2^2 \\
&\overset{(v)}{\leqslant} (\frac{3}{2})^{2(L-1)} L^2 s_{\max}^4\left(\tilde{B}\right) \|X\|_F^2 \frac{\bar{\lambda}_{1\to L}^2}{\min_l \bar{\lambda}_l^2} \|O_L^0 - y\|^2, \\
&\overset{(vi)}{=} \Lambda^2 \|O_L^0 - y\|^2
\end{aligned} \tag{57}
$$

where (i) uses the definition of $\tilde{g}^0$; (ii) extracts the largest singular value of B; (iii) is because we assume $s_{\max}(\tilde{B}) \geqslant 1$; (iv) uses the upper bound of gradient in (42), (v) comes from the fact that $s_{\max}(W_l^0) \geqslant \frac{1}{2}\bar{\lambda}_l$; (iii) is because the definition of Λ in (48).

Recall that we have defined

$$
\phi := (\frac{1}{2}\nu)^{2(L-2)} \lambda_O^2 \underline{\lambda}_{3\to L}^2.
$$

If we choose $\eta = \frac{\phi}{Q_0 \Lambda^2}$, then plug (56) and (57) into (55), we obtain

$$
f(\theta^0 - \eta \tilde{g}^0) = f(\theta^0) - 2\eta\phi f(\theta^0) + \eta\phi f(\theta^0) = (1 - \eta\phi) f(\theta^0). \tag{58}
$$

(Step 1.d) We will show

$$
f(\theta^1) \leqslant \alpha f(\theta^0).
$$

Since $\tilde{g}^0 = \tilde{q}^0$, we have $f(\theta^1) = f(\theta^0 - \eta\tilde{g}^0) \overset{(i)}{\leqslant} (1 - \eta\phi) f(\theta^0) < \alpha f(\theta^0)$, where (i) is because (58). This completes the proof of this step.

(Step 1.e) We will show that:

$$
\begin{cases}
\|B u_l^1 - v_l^1\|_\infty \leqslant \tau\gamma^0, \ l = 2 \\
\|u_l^1 - v_l^1\|_\infty \leqslant \tau\gamma^0, \ l = \{3, \cdots, L\}.
\end{cases}
$$

Notice that for $l = 2$, we have

$$\begin{aligned}\|Bu_2^1 - z_2^0\|_\infty &\overset{(i)}{\leqslant} \|Bu_2^1 - Bu_2^0\| + \|Bu_2^0 - z_2^0\|_\infty \overset{(ii)}{\leqslant} s_{\max}(\tilde{B})\|u_2^1 - u_2^0\| \\ &\overset{(iii)}{\leqslant} s_{\max}(\tilde{B})\|\text{vec}(\nabla f(\theta^1)) - \text{vec}(\nabla f(\theta^0))\| \\ &\overset{(iv)}{\leqslant} s_{\max}(\tilde{B})Q_0\|\theta^1 - \theta^0\| \overset{(v)}{=} \eta s_{\max}(\tilde{B})Q_0\|\tilde{q}^0\| \\ &\overset{(vi)}{=} \eta s_{\max}(\tilde{B})Q_0\|\tilde{g}^0\| \overset{(vii)}{\leqslant} \eta s_{\max}(\tilde{B})Q_0\Lambda\|O_L^0 - y\| \\ &\overset{(viii)}{\leqslant} \lambda\sqrt{\alpha f(\theta^0)} = \gamma^0,\end{aligned}$$

where (i) uses triangle inequality; (ii) extracts the largest singular value of B and uses $g^0 = q^0$; (iii) is because $\|u_2^1 - u_2^0\| \leqslant \|u^1 - u^0\|$; (iv) is from (53), which uses (45) and definition of Q_0 in (49); (v) uses the update rule in (37); (vi) is because $g^0 = q^0$; (vii) uses the definition of Λ in (48); (viii) uses the definition of λ in (47) and the fact $\alpha = 1 - \frac{\phi^2}{4Q_0\Lambda^2} > 1 - \frac{1}{4} = \frac{3}{4}$.

Then by Lemma 4.1, with the correspondence that $c = Bu_2^1, p = z_2^0, r = \gamma^0$, we can obtain

$$\|z_2^1 - Bu_2^1\| \overset{(i)}{=} \|\text{quant}(Bu_2^1, z_2^0, \gamma^0, b) - Bu_2^1\|_\infty \overset{(ii)}{\leqslant} \tau\gamma^0,$$

where (i) comes from the 'Quantize' step in Algorithm 1; (ii) is from the first relation in Lemma 4.1.

For $l \geqslant 3$, we have

$$\begin{aligned}\|u_l^1 - v_l^0\|_\infty &\overset{(i)}{\leqslant} \|u_l^1 - u_l^0\| + \|u_l^0 - z_l^0\|_\infty \overset{(ii)}{\leqslant} \|\text{vec}(\nabla f(\theta^1)) - \text{vec}(\nabla f(\theta^0))\| \\ &\overset{(iii)}{\leqslant} Q_0\|\theta^1 - \theta^0\| \overset{(iv)}{=} \eta Q_0\|\tilde{q}^0\| \overset{(v)}{=} \eta Q_0\|\tilde{g}^0\| \overset{(vi)}{\leqslant} \eta Q_0\Lambda\|O_L^0 - y\| \\ &\overset{(vii)}{\leqslant} \lambda\sqrt{\alpha f(\theta^0)} = \gamma_0,\end{aligned}$$

where (i) uses triangle inequality; (ii) is because $g^0 = q^0$ and uses the fact $\|u_l^0 - z_l^0\| \leqslant \|g^0 - q^0\|$; (iii) is from (53); (iv) uses the update rule in (37); (v) is because $\tilde{g}^0 = \tilde{q}^0$; (vi) uses the upper bound of $\|\tilde{g}^0\|$ in (42) and the definition of Λ in (48); (vii) comes from the choice of λ in (47) and the fact $\alpha = 1 - \frac{\phi^2}{4Q_0\Lambda^2} > 1 - \frac{1}{4} = \frac{3}{4}$.

Then by Lemma 4.1, with the correspondence that $c = u_l^1, p = z_l^0, r = \gamma^0$, we can obtain

$$\|z_l^{t+1} - u_l^{t+1}\| \overset{(i)}{=} \|\text{quant}(u_l^1, z_l^0, \gamma^0, b) - u_l^1\|_\infty \overset{(ii)}{\leqslant} \tau\gamma^0.$$

where (i) comes from the 'Quantize' step in Algorithm 1; (ii) is from the first relation in Lemma 4.1. Now we have showed all the induction assumptions hold for $t = 1$.

Next, we show Step 2. Given these inequalities in (52) hold for iteration t, we aim to show that they hold for $t + 1$. We prove Step 2 in five substeps similarly as in the proof of Step 1.

(Step 2.a) We will show that

$$\begin{cases} s_{\max}\left(W_l^{t+1}\right) \leq \frac{3}{2}\bar{\lambda}_l \;\; l \in [L],\; l \in \{2, 3, \ldots, L\} \\ s_{\min}\left(W_l^{t+1}\right) \geqslant \frac{1}{2}\underline{\lambda}_l,\; l \in \{3, \ldots, L\}. \end{cases}$$

For $l = 2$, we have:

$$\begin{aligned}&\|W_2^{t+1} - W_2^0\|_F \overset{(i)}{\leqslant} \sum_{r=0}^{t}\|W_2^{r+1} - W_2^r\|_F \overset{(ii)}{\leqslant} \sum_{r=0}^{t}\|\tilde{z}_2^r\| \\ &\overset{(iii)}{\leqslant} \eta\sum_{r=0}^{t}(\|\tilde{u}_2^r\| + \|\tilde{u}_2^r - \tilde{z}_2^r\|) \\ &\overset{(iv)}{\leqslant} \eta\sum_{r=0}^{t}(s_{\max}^2(\tilde{B})\Lambda_2\|O_L^r - y\| + n_1n_2\|B^\top(Bu_2^r - z_2^r)\|_\infty)\end{aligned}$$

$$
\begin{aligned}
&\overset{(v)}{\leqslant} \eta \sum_{r=0}^{t} (s_{\max}^2(\tilde{B})\Lambda_2 \|O_L^r - y\| + H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \|B u_2^r - z_2^r\|_\infty) \\
&\overset{(vi)}{\leqslant} \eta \sum_{r=0}^{t} \left(s_{\max}^2(\tilde{B})\Lambda_2 \|O_L^r - y\| + H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \tau\lambda \sqrt{(\alpha)^r f(\theta^0)}\right) \\
&\overset{(vii)}{=} (\eta s_{\max}^2(\tilde{B})\Lambda_2 + \frac{\sqrt{2}}{2}\eta H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \tau\lambda)\|O_L^0 - y\| \sum_{r=0}^{t} \alpha^{\frac{r}{2}} \\
&= (\eta s_{\max}^2(\tilde{B})\Lambda_2 + \frac{\sqrt{2}}{2}\eta H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \tau\lambda)\|O_L^0 - y\| \frac{1-\alpha^{\frac{r+1}{2}}}{1-\alpha^{\frac{1}{2}}} \\
&\overset{(viii)}{\leqslant} (\eta s_{\max}^2(\tilde{B})\Lambda_2 + \frac{\sqrt{2}}{2}\eta H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \tau\lambda)\|O_L^0 - y\| \frac{1}{1-\alpha}(1+\alpha^{\frac{1}{2}}) \\
&\overset{(ix)}{\leqslant} \frac{8(s_{\max}^2(\tilde{B})\Lambda_2 + \frac{\sqrt{2}}{2} H n_1 n_2^2 \|\tilde{B}^\top\|_\infty \tau\lambda)}{\phi}\|O_L^0 - y\| \\
&\overset{(x)}{\leqslant} \frac{1}{4} + \frac{1}{4} = \frac{1}{2} < 1,
\end{aligned} \tag{59}
$$

where (i) uses triangle inequality; (ii) uses the update rule in (37); (iii) uses triangle inequality; (iv) extracts the largest singular value of $B^\top B$, the upper bound of $\|u_2^t\|$ in (41), definition of Λ_2 in (51) and the relationship between l_2 and l_∞ norm; (v) uses the fact that $\|B^\top(Bu_2^r - z_2^r)\|_\infty \leqslant H n_2 \|\tilde{B}\|_\infty \|Bu_2^r - z_2^r\|_\infty$; (vi) uses the induction assumption $\|Bu_2^r - z_2^r\|_\infty \leqslant \tau\gamma^{r-1}$ and definition of γ^{r-1}; (vii) reorganizes the terms; (viii) is because $0 < \alpha < 1$; (ix) uses $1 + \alpha^{\frac{1}{2}} < 2$ and plugs in the choice of α; (x) is because the initialization strategy in (46) and choice of τ in (50).

So Applying Weyl' inequality to the matrices W_2^t and $(W_2^{t+1} - W_2^t)$ and (59), we have

$$
s_{\max}(W_2^{t+1}) \leqslant \bar{\lambda}_2 + 1 = \frac{3}{2}\bar{\lambda}_2.
$$

For $l \geqslant 3$, we have the following relation:

$$
\begin{aligned}
&\|W_l^{t+1} - W_l^0\|_F \overset{(i)}{\leqslant} \sum_{r=0}^{t} \|W_l^{r+1} - W_l^r\|_F \overset{(ii)}{\leqslant} \sum_{r=0}^{t} \|z_l^r\| \\
&\overset{(iii)}{\leqslant} \eta \sum_{r=0}^{t} (\|u_l^r\| + \|u_l^r - z_l^r\|) \\
&\overset{(iv)}{\leqslant} \eta \sum_{r=0}^{t} (\Lambda_l \|O_L^r - y\| + n_{l-1} n_l \|u_l^r - z_l^r\|_\infty) \\
&\overset{(v)}{\leqslant} \eta \sum_{r=0}^{t} \left(\Lambda_l \|O_L^r - y\| + n_{l-1} n_l \tau\lambda \sqrt{(\alpha)^r f(\theta^0)}\right) \\
&\overset{(vi)}{\leqslant} (\eta\Lambda_l + \frac{\sqrt{2}}{2}\eta\tau\lambda n_{l-1} n_l)\|O_L^0 - y\| \sum_{r=0}^{t} \alpha^{\frac{r}{2}} \\
&\overset{(vii)}{\leqslant} (\eta\Lambda_l + \frac{\sqrt{2}}{2}\eta\tau\lambda n_{l-1} n_l)\|O_L^0 - y\| \frac{1}{1-\alpha} \\
&= (\eta\Lambda_l + \frac{\sqrt{2}}{2}\eta\tau\lambda n_{l-1} n_l)\|O_L^0 - y\| \frac{1}{1-\alpha}(1+\alpha^{\frac{1}{2}}) \\
&\overset{(viii)}{\leqslant} \frac{8(\Lambda_l + \frac{\sqrt{2}}{2}\tau\lambda n_{l-1} n_l)}{\phi}\|O_L^0 - y\| \\
&\overset{(ix)}{\leqslant} \frac{1}{8}\underline{\lambda}_l + \frac{1}{8}\underline{\lambda}_l = \frac{1}{4}\underline{\lambda}_l < \frac{1}{2}\bar{\lambda}_l,
\end{aligned} \tag{60}
$$

where (i) uses triangle inequality; (ii) uses the update rule in (37); (iii) uses triangle inequality; (iv) uses the upper bound of $\|u_l^t\|$ in (41), the definition of Λ_l in (51) and the relationship between l_2 and

l_∞ norm; (v) uses the induction assumption $\|u_l^r - z_l^r\| \leqslant \tau\gamma^{r-1}$; (vi) uses reorganizes the terms; (vii) is because $0 < \alpha < 1$; (viii) uses the fact $1 + \alpha^{\frac{1}{2}} < 2$ and plugs in the choice of α; (ix) is from the initialization strategy in (46) and the choice of τ in (50).

Applying Weyl' inequality to the matrices W_l^{t+1} and $(W_l^{t+1} - W_l^t)$, we have

$$s_{\max}(W_l^{t+1}) \leqslant \bar{\lambda}_l + \frac{1}{4}\bar{\lambda}_l = \frac{5}{4}\bar{\lambda}_l < \frac{3}{2}\bar{\lambda}_l,\ s_{\min}(W_l^{t+1}) \geqslant \bar{\lambda}_l - \frac{1}{2}\underline{\lambda}_l = \frac{1}{2}\bar{\lambda}_l. \tag{61}$$

(Step 2.b) We will show that

$$\begin{cases} s_{\max}(W_l^{t+1} + \eta\tilde{\Delta}W_l^t - \eta\Delta W_l^t) \leqslant \frac{3}{2}\bar{\lambda}_l, l = \{2, 3, \cdots, L\} \\ s_{\min}(W_l^{t+1} + \eta\tilde{\Delta}W_l^t - \eta\Delta W_l^t) \geqslant \frac{1}{2}\underline{\lambda}_l, l = \{3, \cdots, L\}. \end{cases}$$

For $l = 2$, we have

$$\begin{aligned} & s_{\max}(W_2^{t+1} + \eta\tilde{\Delta}W_2^t - \eta\Delta W_2^t) \overset{(i)}{\leqslant} s_{\max}(W_2^{t+1}) + \eta\|\tilde{\Delta}W_2^t - \Delta W_2^t\|_F \\ & \overset{(ii)}{=} s_{\max}(W_2^{t+1}) + \eta\|\tilde{z}_2^t - \tilde{u}_2^t\| \\ & \overset{(iii)}{\leqslant} s_{\max}(W_2^{t+1}) + \eta n_1 n_2\|B^\top(Bu_2^r - z_2^r)\|_\infty \\ & \overset{(iv)}{\leqslant} s_{\max}(W_2^{t+1}) + Hn_1n_2^2\|\tilde{B}^\top\|_\infty\|Bu_2^r - z_2^r\|_\infty \\ & \overset{(v)}{\leqslant} s_{\max}(W_2^{t+1}) + Hn_1n_2^2\|\tilde{B}^\top\|_\infty\eta\tau\lambda\sqrt{(\alpha)^r f(\theta^0)} \\ & \overset{(vi)}{\leqslant} s_{\max}(W_2^{t+1}) + Hn_1n_2^2\|\tilde{B}^\top\|_\infty\eta\tau\lambda\sqrt{f(\theta^0)} \overset{(v)}{\leqslant} \bar{\lambda}_l + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}\bar{\lambda}_2, \end{aligned}$$

where (i) uses Weyl's inequality and the relationship between the Frobenius norm and l_2 norm of matrix; (ii) uses the definition in (36); (iii) uses the relationship between l_2 and l_∞ norm and the definition of $\tilde{u}_2^t$ and $\tilde{z}_2^t$; (iv) uses the fact that $\|B^\top(Bu_2^r - z_2^r)\|_\infty \leqslant Hn_2\|\tilde{B}\|_\infty\|Bu_2^r - z_2^r\|_\infty$ and the definition of γ^{r-1}; (v) is from the induction assumption $\|Bu_2^r - z_2^r\|_\infty \leqslant \tau\gamma^{r-1}$; (vi) is from the fact that $\alpha \leqslant 1$; (v) is from the (59) and choice of τ and λ in (50) and (47).

For $l \geqslant 3$, by Weyl's inequality, we have

$$\begin{aligned} & s_{\max}(W_l^{t+1} + \eta\tilde{\Delta}W_l^t - \eta\Delta W_l^t) \overset{(i)}{\leqslant} s_{\max}(W_l^{t+1}) + \eta\|\tilde{\Delta}W_l^t - \Delta W_l^t\|_F \\ & \overset{(ii)}{=} s_{\max}(W_l^{t+1}) + \eta\|z_l^t - u_l^t\| \overset{(iii)}{\leqslant} s_{\max}(W_l^{t+1}) + \eta n_{l-1}n_l\tau\lambda\sqrt{(\alpha)^t f(\theta^0)} \\ & \overset{(iv)}{\leqslant} s_{\max}(W_l^{t+1}) + \eta n_{l-1}n_l\tau\lambda\sqrt{f(\theta^0)} \overset{(v)}{\leqslant} \bar{\lambda}_l + \frac{1}{4}\underline{\lambda}_l + \frac{1}{4}\underline{\lambda}_l \leqslant \frac{3}{2}\bar{\lambda}_l, \end{aligned} \tag{62}$$

where (i) uses the Weyl's inequality on W_l^{t+1} and $\eta(\tilde{\Delta}W_l^t - \Delta W_l^t)$; (ii) uses the update rule in (37); (iii) uses the relationship between l_2 and l_∞ norm and the induction assumption; (iv) is because $\alpha < 1$; (v) uses (60) and choice of τ and λ in (50) and (47).

Similarly,

$$\begin{aligned} & s_{\min}(W_l^{t+1} + \eta\tilde{\Delta}W_l^t - \eta\Delta W_l^t) \overset{(i)}{\geqslant} s_{\min}(W_l^{t+1}) - \eta\|\tilde{\Delta}W_l^t - \Delta W_l^t\|_F \\ & \overset{(ii)}{=} s_{\min}(W_l^{t+1}) - \eta\|z_l^t - u_l^t\| \overset{(iii)}{\geqslant} s_{\min}(W_l^{t+1}) - \eta n_{l-1}n_l\tau\lambda\sqrt{(\alpha)^t f(\theta^0)} \\ & \overset{(iv)}{\geqslant} s_{\min}(W_l^{t+1}) - \eta n_{l-1}n_l\tau\lambda\sqrt{f(\theta^0)} \overset{(v)}{\geqslant} \frac{3}{4}\underline{\lambda}_l - \frac{1}{4}\underline{\lambda}_l = \frac{1}{2}\underline{\lambda}_l, \end{aligned}$$

where (i) uses Weyl's inequality on W_l^{t+1} and $\eta(\tilde{\Delta}W_l^t - \Delta W_l^t)$; (ii) plugs in the update rule in (37); (iii) uses the relationship between l_2 and l_∞ norm and the induction assumption; (iv) is because $\alpha < 1$; (v) comes from (60) and choice of τ and λ in (50) and (47).

(Step 2.c) We will show that:

$$f(\theta^t - \eta\tilde{g}^t) \leqslant (1 - \eta\phi)f(\theta^t).$$

From (1) and (2) we know

$$\max\left(s_{\max}(W_l^t), s_{\max}(W_l^t)\right) \leqslant \frac{3}{2}\bar{\lambda}_l.$$

Using the above relation, we can upper bound the differences of the gradients by using (45). More specifically, using the the definition of Q_0 in (49), we have

$$\|\text{vec}(\nabla f(\theta^{t+1})) - \text{vec}(\nabla f(\theta^t))\| \leqslant Q_0\|\theta^{t+1} - \theta^t\|,$$

where Q_0 is defined in (49).

Further, it is easy to verify that for any $\hat{\theta}^t$ between θ^t and θ^{t+1}, we still have $s_{\max}(W_l(\hat{\theta}^t)) \leqslant \frac{3}{2}\bar{\lambda}_l$. So we can apply the same argument leading to (53), and obtain:

$$\|\text{vec}(\nabla f(\hat{\theta}^t)) - \text{vec}(\nabla f(\theta^t))\| \leqslant Q_0\|\hat{\theta}^t - \theta^t\|, \forall \hat{\theta}^t = \theta^t + \delta(\theta^{t+1} - \theta^t), \delta \in [0, 1].$$

So we have

$$\begin{aligned}
f(\theta^t - \eta\tilde{g}^t) &\overset{(i)}{\leqslant} f(\theta^t) - \eta\langle u^t, \tilde{g}^t\rangle + \frac{Q_0}{2}\eta^2\|\tilde{g}^t\|^2 \\
&\overset{(ii)}{=} f(\theta^t) - \eta\langle u_2^t, \tilde{u}_2^t\rangle - \eta\sum_{l=3}^{L}\|u_l^t\|^2 + \frac{Q_0}{2}\eta^2\|\tilde{g}^t\|^2 \\
&\leqslant f(\theta^t) - \eta\langle u_2^t, \tilde{u}_2^t\rangle + \frac{Q_0}{2}\eta^2\|\tilde{g}^t\|^2,
\end{aligned}$$

where (i) uses Decent Lemma; (ii) expands $\langle u^t, \tilde{g}^t\rangle$ by utilizing the stacked structure. Similar to (56), with the induction assumption $s_{\min}(W_l^t) \geqslant \frac{1}{2}\underline{\lambda}_l, l \geqslant 3$, we have the following relation:

$$\langle u_2^t, \tilde{u}_2^t\rangle \geqslant \phi\|O_L^t - y\|^2.$$

Similar to (57), with the induction assumption $s_{\max}(W_l^t) \leqslant \frac{3}{2}\bar{\lambda}_l, l \geqslant 2$, we have

$$\|\tilde{g}^t\|^2 \leqslant \Lambda^2\|O_L^t - y\|^2.$$

If we choose $\eta \leqslant \frac{\phi}{Q_0\Lambda^2}$, we obtain

$$f(\theta^t - \eta\tilde{g}^t) \leqslant f(\theta^t) - 2\eta\phi f(\theta^t) + \eta\phi f(\theta^t) = (1 - \eta\phi)f(\theta^t).$$

(Step 2.d) We will show that:

$$f(\theta^{t+1}) \leqslant \alpha f(\theta^t).$$

We have the following relation:

$$\begin{aligned}
&f(\theta^{t+1}) = f(\theta^{t+1}) - f(\theta^t - \eta\tilde{g}^t) + f(\theta^t - \eta\tilde{g}^t) \\
&\overset{(i)}{\leqslant} \frac{1}{2}\|G(\theta^{t+1}) - G(\theta^t - \eta\tilde{g}^t)\|^2 + \left\langle G(\theta^{t+1}) - G(\theta^t - \eta\tilde{g}^t), G(\theta^t - \eta\tilde{g}^t) - y\right\rangle \\
&\quad + f(\theta^t + \eta\tilde{q}^t - \eta\tilde{g}^t) \\
&\overset{(ii)}{\leqslant} \frac{1}{2}\|G(\theta^{t+1}) - G(\theta^t - \eta\tilde{g}^t)\|^2 + \frac{1}{\beta}\|G(\theta^{t+1}) - G(\theta^t - \eta\tilde{g}^t)\|^2 + \frac{\beta}{2}\|G(\theta^t t - \eta\tilde{g}^t) - y\|^2 \\
&\quad + f(\theta^t t - \eta\tilde{g}^t) \\
&\overset{(iii)}{\leqslant} (\frac{1}{2} + \frac{1}{\beta})\|G(\theta^{t+1}) - G(\theta^t - \eta\tilde{g}^t)\|^2 + (1 + \beta)(1 - \eta\phi)f(\theta^t) \\
&\overset{(iv)}{=} (\frac{1}{2} + \frac{1}{\beta})\langle \text{vec}(\nabla G(\hat{\theta}^t)), -\eta\tilde{q}^t + \eta\tilde{g}^t\rangle^2 + (1 + \beta)(1 - \eta\phi)f(\theta^t) \\
&\overset{(v)}{\leqslant} \eta^2(\frac{1}{2} + \frac{1}{\beta})\|\text{vec}(\nabla G(\hat{\theta}^t))\|^2\|\tilde{q}^t - \tilde{g}^t\|^2 + (1 + \beta)(1 - \eta\phi)f(\theta^t),
\end{aligned}$$

where (i) expands and reorganizes the loss function; (ii) uses Young's inequality with constant β; (iii) uses the induction assumption $f(\theta^{t+1}) \leqslant (1 - \eta\phi)f(\theta^t)$; (iv) uses the mean value Theorem where $\hat{\theta}^t$ will be discussed in the next paragraph, the update rule in (37); (v) uses Cauchy-Schwartz inequality.

Now we discuss $\hat{\theta}^t$ and derive a bound for $\|\text{vec}(\nabla G(\hat{\theta}^t))\|$. By the Mean Value Theorem we have

$$\hat{\theta}^t = \theta^{t+1} + \delta(\eta\tilde{g}^t - \eta\tilde{q}^t),$$

for some $\delta \in [0, 1]$. By (61) and (62), we know that

$$s_{\max}(W_l(\theta^{t+1})) \leqslant \frac{5}{4}\bar{\lambda}_l,\ s_{\max}(W_l(\theta^t - \eta\tilde{g}^t)) \leqslant \frac{3}{2}\bar{\lambda}_l.$$

So it is easy to conclude that for $l \geqslant 3$,

$$\begin{aligned} s_{\max}(W_l(\hat{\theta}^t)) &\overset{(i)}{\leqslant} s_{\max}(W_l(\theta^{t+1})) + \delta\eta\|z_l^t - u_l^t\| \\ &\overset{(ii)}{\leqslant} s_{\max}(W_l^{t+1}) + \eta\|z_l^t - u_l^t\| \overset{(iii)}{\leqslant} \frac{3}{2}\bar{\lambda}_l, \end{aligned}$$

where (i) uses Wely's inequality on W_l^{t+1} and $\delta\eta\big(\tilde{\Delta}W_l^t - \Delta W_l^t\big)$; (ii) is because $\delta \in [0, 1]$; (iii) uses (62). We can derive the similar result for $l = 2$. By (42) without $\|O_L^t - y\|$, it is clear that

$$\|\text{vec}(\nabla G(\hat{\theta}^t))\| \leqslant \Lambda.$$

So we have

$$\begin{aligned} &\eta^2(\frac{1}{2} + \frac{1}{\beta})\|\text{vec}(\nabla G(\hat{\theta}^t))\|^2\|\tilde{q}^t - \tilde{g}^t\|^2 + (1+\beta)(1-\eta\phi)f(\theta^t) \\ &\overset{(i)}{\leqslant} \eta^2(\frac{1}{2} + \frac{1}{\beta})\Lambda^2(\|\tilde{u}_2^t - \tilde{v}_2^t\|^2 + \sum_{l=3}^{L}\|u_l^t - z_l^t\|^2) + (1+\beta)(1-\eta\phi)f(\theta^t) \\ &\overset{(ii)}{\leqslant} \eta^2(\frac{1}{2} + \frac{1}{\beta})\Lambda^2(n_1^2n_2^2\|\tilde{u}_2^t - \tilde{z}_2^t\|_\infty^2 + \sum_{l=3}^{L} n_{l-1}^2n_l^2\|u_l^t - z_l^t\|_\infty^2) + (1+\beta)(1-\eta\phi)f(\theta^t) \\ &\overset{(iii)}{\leqslant} \eta^2(\frac{1}{2} + \frac{1}{\beta})\Lambda^2(H^2n_1^4n_2^2\|\tilde{B}^\top\|_\infty^2\|Bu_2^t - z_2^t\|_\infty^2 + \sum_{l=3}^{L} n_{l-1}^2n_l^2\|u_l^t - z_l^t\|_\infty^2) + (1+\beta)(1-\eta\phi)f(\theta^t) \\ &\overset{(iv)}{\leqslant} \eta^2(\frac{1}{2} + \frac{1}{\beta})\Lambda^2\left(H^2n_1^4n_2^2\|\tilde{B}^\top\|_\infty^2\tau^2\lambda^2(\alpha)^tf(\theta^0) + \sum_{l=3}^{L} n_{l-1}^2n_l^2\tau^2\lambda^2(\alpha)^tf(\theta^0)\right) + (1+\beta)(1-\eta\phi)f(\theta^t) \\ &= \left((1+\beta)(1-\eta\phi) + \eta^2(\frac{1}{2} + \frac{1}{\beta})(H^2n_1^4n_2^2\|\tilde{B}^\top\|_\infty^2 + \sum_{l=3}^{L} n_{l-1}^2n_l^2)\Lambda^2\tau^2\lambda^2\right)(\alpha)^tf(\theta^0), \end{aligned}$$

(i) uses (42) with the fact that $s_{\max}(W_l(\hat{\theta}^t)) \leqslant \frac{3}{2}\bar{\lambda}_l$, and expands $\|\tilde{q}^t - \tilde{g}^t\|^2$; (ii) uses the relationship between l_2 and l_∞ norm; (iii) uses the fact that $\|B^\top(Bu_2^r - z_2^r)\|_\infty \leqslant Hn_2\|\tilde{B}\|_\infty\|Bu_2^r - z_2^r\|_\infty$; (iv) uses the induction assumption $\|Bu_2^t - z_2^t\| \leqslant \tau\gamma^{t-1}$, $f(\theta^t) \leqslant (\alpha)^tf(\theta^0)$ and the definition of γ^{t-1}.

Let $\beta = \frac{\frac{1}{2}\eta\phi}{1-\eta\phi}$, we have

$$\begin{aligned} &\left((1+\beta)(1-\eta\phi) + \eta^2(\frac{1}{2} + \frac{1}{\beta})(Hn_1n_2^2\|\tilde{B}^\top\|_\infty^2 + \sum_{l=3}^{L} n_{l-1}^2n_l^2)\Lambda^2\tau^2\lambda^2\right)(\alpha)^tf(\theta^0) \\ &\overset{(i)}{\leqslant} \left(1 - \frac{1}{2}\eta\phi + \eta^2(\frac{1}{2} + \frac{2(1-\eta\phi)}{\eta\phi})(Hn_1n_2^2\|\tilde{B}^\top\|_\infty^2 + \sum_{l=3}^{L} n_{l-1}^2n_l^2)\Lambda^2\tau^2\lambda^2\right)(\alpha)^tf(\theta^0) \\ &\overset{(ii)}{\leqslant} (1 - \frac{1}{4}\eta\phi)(\alpha)^tf(\theta^0) = (\alpha)^{t+1}f(\theta^0), \end{aligned}$$

where (i) plugs in the choice of β; (ii) is because the choice of τ and λ in (50) and (47).

(Step 2.e) We will show that:

$$\begin{cases} \|Bu_l^{t+1} - z_l^{t+1}\|_\infty \leqslant \tau\gamma^t,\ l = 2 \\ \|u_l^{t+1} - z_l^{t+1}\|_\infty \leqslant \tau\gamma^t,\ l = \{3, \cdots, L\}. \end{cases}$$

For $l = 2$, we have

$$
\begin{aligned}
\|Bu_2^{t+1} - z_2^t\|_\infty &\overset{(i)}{\leqslant} \|Bu_2^{t+1} - Bu_2^t\| + \|Bu_2^t - z_2^t\|_\infty \overset{(ii)}{\leqslant} s_{\max}(\tilde{B})\|u_2^{t+1} - u_2^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(iii)}{\leqslant} s_{\max}(\tilde{B})\|u^{t+1} - u^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(iv)}{\leqslant} s_{\max}(\tilde{B})\|\text{vec}(\nabla f(\theta^{t+1})) - \text{vec}(\nabla f(\theta^t))\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(v)}{\leqslant} s_{\max}(\tilde{B})Q_0\|\theta^{t+1} - \theta^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(vi)}{=} \eta s_{\max}(\tilde{B})Q_0\|\tilde{q}^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(vii)}{\leqslant} \eta s_{\max}(\tilde{B})Q_0(\|\tilde{g}^t\| + \|\tilde{g}^t - \tilde{q}^t\|) + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(viii)}{\leqslant} \eta s_{\max}(\tilde{B})Q_0\Lambda\|O_L^t - y\| + \eta s_{\max}^2(\tilde{B})Q_0\|g^t - q^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(ix)}{\leqslant} \eta s_{\max}(\tilde{B})Q_0\Lambda\|O_L^t - y\| + \eta s_{\max}^2(\tilde{B})Q_0\sum_{l=3}^{L}\|u_l^t - z_l^t\| + \eta s_{\max}^2(\tilde{B})Q_0\|Bu_2^t - z_2^t\| + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(x)}{\leqslant} \eta s_{\max}(\tilde{B})Q_0\Lambda\|O_L^t - y\| + \eta s_{\max}^2(\tilde{B})Q_0Hn_2\|u_2^t - z_2^t\|_\infty + \eta s_{\max}^2(\tilde{B})Q_0\sum_{l=3}^{L}n_{l-1}n_l\|u_l^t - z_l^t\|_\infty + \|Bu_2^t - z_2^t\|_\infty \\
&\overset{(xi)}{\leqslant} \left(1 + \eta s_{\max}^2(\tilde{B})Q_0\big(Hn_2 + \sum_{l=3}^{L}n_{l-1}n_l\big)\right)\tau\gamma^{t-1} + \eta s_{\max}(\tilde{B})Q_0\Lambda\|O_L^t - y\| \\
&\overset{(xii)}{\leqslant} \lambda\sqrt{(\alpha)^t f(\theta^0)}\left(\tau + \tau\eta s_{\max}^2(\tilde{B})Q_0\big(Hn_2 + \sum_{l=3}^{L}n_{l-1}n_l\big) + \frac{\sqrt{2}\eta s_{\max}(\tilde{B})Q_0\Lambda}{\lambda}\right) \\
&\overset{(xiii)}{\leqslant} \lambda\sqrt{\alpha}\sqrt{(\alpha)^t f(\theta^0)} = \lambda\gamma^t,
\end{aligned}
$$

where (i) uses triangle inequality; (ii) extracts the largest singular value of B, (iii) uses the fact $\|u^{t+1} - u^t\| \geqslant \|u_2^{t+1} - u_2^t\|$; (iv) uses the definition of u^t; (v) upper bounds the differences of the gradients by using (45) (vi) uses the update rule in (37); (vii) uses triangle inequality; (viii) uses the upper bound of gradient norm in (42), and it extracts the largest singular value of B and definition of Λ in (48); (ix) uses the stacked structure of g^t and q^t to expand $\|g^t - q^t\|$; (x) uses the relationship between l_2 and l_∞ norm; (xi) reorganizes the terms and uses induction assumption; (xii) uses the definition of γ^{t-1}; (xiii) is from the choice of τ and λ in (50) and (47).

Then by Lemma 4.1, with the correspondence that $c = Bu_2^{t+1}, p = z_2^t, r = \gamma^t$, we can obtain

$$
\|z_2^{t+1} - Bu_2^{t+1}\| \overset{(i)}{=} \|\text{quant}(Bu_2^{t+1}, z_2^t, \gamma^t, b) - Bu_2^{t+1}\|_\infty \overset{(ii)}{\leqslant} \tau\gamma^t.
$$

where (i) comes from the 'Quantize' step in Algorithm 1; (ii) is from the first relation in Lemma 4.1.

For $l \geqslant 3$, we have

$$
\begin{aligned}
\|u_l^{t+1} - z_l^t\|_\infty &\overset{(i)}{\leqslant} \|u_l^{t+1} - u_l^t\| + \|u_l^t - z_l^t\|_\infty \overset{(ii)}{\leqslant} \|u^{t+1} - u^t\| + \|u_l^t - z_l^t\|_\infty \\
&\overset{(iii)}{=} \|\text{vec}(\nabla f(\theta^{t+1})) - \text{vec}(\nabla f(\theta^t))\| + \|u_l^t - z_l^t\|_\infty \\
&\overset{(iv)}{\leqslant} \eta Q_0\|\tilde{q}^t\| + \|u_l^t - z_l^t\|_\infty \\
&\overset{(v)}{\leqslant} \eta Q_0(\|\tilde{g}^t\| + \|\tilde{g}^t - \tilde{q}^t\|) + \|u_l^t - z_l^t\|_\infty \\
&\overset{(vi)}{\leqslant} \eta Q_0\Lambda\|O_L^t - y\| + \eta s_{\max}(\tilde{B})Q_0\|g^t - q^t\| + \|u_l^t - z_l^t\|_\infty \\
&\overset{(vii)}{\leqslant} \eta Q_0\Lambda\|O_L^t - y\| + \eta s_{\max}(\tilde{B})Q_0\sum_{l=3}^{L}\|u_l^t - z_l^t\| + \eta s_{\max}(\tilde{B})Q_0\|Bu_2^t - z_2^t\| + \|u_2^t - z_2^t\|_\infty
\end{aligned}
$$

$$\overset{(viii)}{\leqslant} \eta Q_0 \Lambda \|O_L^t - y\| + \eta s_{\max}(\tilde{B}) Q_0 H n_2 \|B u_2^t - z_2^t\|_\infty + \eta s_{\max}(\tilde{B}) Q_0 \sum_{l=3}^{L} n_{l-1} n_l \|u_l^t - z_l^t\|_\infty + \|u_2^t - z_2^t\|_\infty$$

$$\overset{(ix)}{\leqslant} \left(1 + \eta s_{\max}(\tilde{B}) Q_0 \big(H n_2 + \sum_{l=3}^{L} n_{l-1} n_l\big)\right) \tau \gamma^{t-1} + \eta Q_0 \Lambda \|O_L^t - y\|$$

$$\overset{(x)}{\leqslant} \lambda \sqrt{f(\theta^t)} \left(\tau + \tau \eta s_{\max}(\tilde{B}) Q_0 \big(H n_2 + \sum_{l=3}^{L} n_{l-1} n_l\big) + \frac{\sqrt{2} \eta Q_0 \Lambda}{\lambda}\right)$$

$$\overset{(xi)}{\leqslant} \lambda \sqrt{\alpha} \sqrt{f(\theta^t)} = \lambda \gamma^t,$$

where (i) uses triangle inequality; (ii) uses the fact $\|u^{t+1} - u^t\| \geqslant \|u_2^{t+1} - u_2^t\|$; (iii) uses the definition of u^t;(iv) uses the update rule in (37); (v) uses triangle inequality; (vi) uses the upper bound of gradient norm in (42), extracts the largest singular value of B and uses the definition of Λ in (48); (vii) uses the stacked structure of g^t and q^t to expand $\|g^t - q^t\|$; (viii) uses the relationship between l_2 and l_∞ norm; (ix) uses the induction assumption; (x) uses the definition of γ^{t-1}; (xi) is because the choice of τ and λ in (50) and (47).

Then by Lemma 4.1, with the correspondence that $c = B u_2^{t+1}, p = z_2^t, r = \gamma^t$, we can obtain

$$\|z_l^{t+1} - u_l^{t+1}\| \overset{(i)}{=} \|\text{quant}(u_l^{t+1}, z_l^t, \gamma^t, b) - B u_l^{t+1}\|_\infty \overset{(ii)}{\leqslant} \tau \gamma^t,$$

where (i) comes from the 'Quantize' step in Algorithm 1; (ii) is from the first relation in Lemma 4.1.

Now we have proved that (52) holds. So for $t > 0$, there is

$$f(\theta^t) \leqslant (\alpha)^t f(\theta^0), \quad \text{where} \quad \alpha = 1 - \frac{1}{4}\eta\phi = 1 - \frac{\phi^2}{4 Q_0 \Lambda^2}.$$

Thus, if we want the objective function to compute an ϵ-optimal solution, the total number of iterations is $\log(f(\theta^0)/\epsilon)/\log(1/(1 - \frac{\phi^2}{4Q_0\Lambda^2}))$. Since in each iteration, each agent k transmits a length-H vector q_k^t, so we conclude that the total number of bits each node needs to communicate is $\log(f(\theta^0)/\epsilon)/\log(1/(1 - \frac{\phi^2}{4Q_0\Lambda^2}))$ bits. Notice that $\log(1/(1 - \frac{\phi^2}{4Q_0\Lambda^2})) = -\log(1 - \frac{\phi^2}{4Q_0\Lambda^2}) \sim \frac{4Q_0\Lambda^2}{\phi^2}$, so we can derive the simplified total number of bits as $b \cdot \frac{4Q_0\Lambda^2}{\phi^2} \cdot \log(f(\theta^0)/\epsilon)$. □