Journal of Machine Learning Research 25 (2024) 1-53 Submitted 7/24; Revised 11/24; Published 12/24

Countering the Communication Bottleneck in Federated Learning: A Highly Efficient Zero-Order Optimization Technique

Elissa Mhanna ELISSA.MHANNA@CENTRALESUPELEC.FR
Mohamad Assaad MOHAMAD.ASSAAD@CENTRALESUPELEC.FR
Laboratoire des Signaux & Systèmes (L2S)
Université Paris-Saclay, CNRS, CentraleSupélec
3 rue Joliot Curie
91190 Gif-sur-Yvette, France

Editor: Peter Richtárik

Abstract

Federated learning (FL) is a creative technique that enables multiple edge devices to train a model without revealing raw data. However, several issues hinder the practical implementation of FL, especially in wireless environments. These issues comprise the limited capacity of the upload transmission link between the edge devices and the aggregator, as well as the wireless disturbances. To address these challenges, we develop a zero-order (ZO) communication-efficient framework for FL. While in standard FL, each device must upload a long vector containing the gradient or the model per communication round, our novel ZO method incorporates a two-point gradient estimator, which requires uploading only two scalars. What also sets our approach apart is that it directly incorporates wireless perturbations into the learning, eliminating the need for additional computational resources to remove their impact. In this work, we overcome the technical and analytical challenges associated with FL problems and ZO methods, comprehensively study our algorithm, and prove it converges almost surely under different conditions, convexity and non-convexity, noise-free and noisy environments. We then find theoretical bounds on the convergence rate when the objective is strongly convex, non-convex, and κ-gradient-dominated that compete with first-order (FO) or centralized methods under the same settings. Finally, we provide experimental results demonstrating the effectiveness of our algorithm, considering relevant examples. We provide an example illustrating the amount of communication saved due to its efficiency compared to its FO counterpart.

Keywords: Federated learning, zero-order, gradient estimate.

1 Introduction

Federated learning (FL) has emerged as an innovative solution for distributed machine learning, as indicated by McMahan et al. (2017). This paradigm has been adopted by major technology companies to implement at scale (Bonawitz et al., 2019) as it addresses the challenge of training models without requiring users to transmit their private data to a central server. Instead, data remains on the users' devices, and model training occurs through collaborative interactions between these devices and the server: The devices receive the model from the server, utilize their data to update gradients, and then transmit these gradients/updated models back to the server. Subsequently, the server refines the model using the accumulated and averaged gradients, and this iterative process repeats.

FL aims to leverage the rich network of devices such as mobile phones, wearable devices, and tablets, creating more intelligent systems with applications spanning from predicting health issues and promoting traffic safety through location-based analytics to customized experiences in digital services. This inspires the growing research on FL (Kairouz et al., 2021), encompassing both first-order (FO) (McMahan et al., 2017; Zhang et al., 2021; Wang et al., 2021) and second-order (Elgabli et al., 2022; Li et al., 2019) methods. However, as these techniques rely on exchanging gradients and Hessians of local objective functions, this introduces various issues, including high communication and computation costs, systems heterogeneity due to the varying capabilities of each device, as well as privacy threats (Li et al., 2020). Successfully addressing these challenges is essential for realizing the full potential of federated learning in real-world applications.

On the other hand, such first and second-order information may not be available in scenarios where the closed form of the loss function is unavailable or when exact gradient calculation may be expensive or difficult. Zero-order (ZO) methods were thus designed to deal with these problems. ZO optimization belongs to the field of gradient-free optimization and relies on estimating the gradient through differences between function evaluations queried at specific points (Duchi et al., 2015; Agarwal et al., 2010). In this study, our focus centers on two-point gradient estimations, particularly for functions represented as $\theta \mapsto f(\theta, \xi)$, where the function is subject to a stochastic perturbation denoted as ξ. These estimations take the form:

$$g = d\frac{f(\theta + \gamma\omega, \xi) - f(\theta - \gamma\omega, \xi)}{2\gamma}\omega,$$

with $\theta \in \mathbb{R}^d$ the optimization variable, $\gamma > 0$ a small value, and ω a random vector with a symmetric distribution. ZO optimization has made its mark in the machine learning community, particularly in conjunction with optimizers that rely on gradient-based techniques. This adoption is notable across diverse domains, including reinforcement learning (Vemula et al., 2019; Malik et al., 2019), crafting contrastive explanations for black-box classification models (Dhurandhar et al., 2019), and the introduction of adversarial perturbations to manipulate these models (Ilyas et al., 2018; Chen et al., 2019).

Another noticeable surge in interest surrounds the optimization and learning processes within wireless environments, particularly in light of the growing number of devices connected to servers via cellular networks (Yang et al., 2020; Amiri and Gündüz, 2020; Sery and Cohen, 2020; Guo et al., 2021; Sery et al., 2021; Sun et al., 2022). Our research is concentrated on a particular scenario, depicted in Figure 1, which pertains to the application of FL in wireless environments. Much like the previous-referenced works, we are examining the scenario of analog communications between the devices and the server.

1.1 Related Work

Despite the benefits of FL, several challenges impede its practical implementation, which extensive research efforts aim to address.

Communication Bottleneck. Generally, in wireless systems, the link from the edge devices to the aggregator, denoted as the uplink, has limited capacity due to the fact that

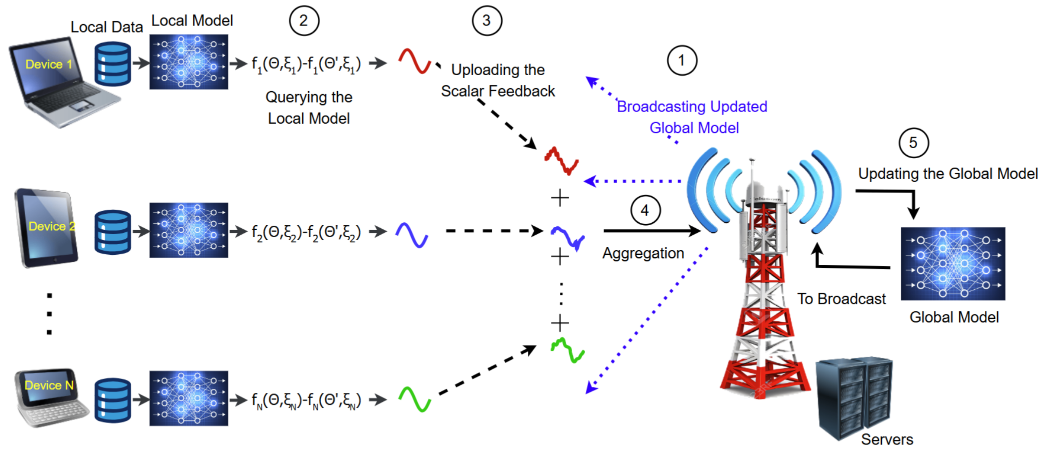
Local Data
Local Model
Device 1
2
$f_1(\Theta,\xi_1)-f_1(\Theta',\xi_1)$
Querying the Local Model
3
Uploading the Scalar Feedback
1
Broadcasting Updated Global Model
5
Updating the Global Model
Device 2
$f_2(\Theta,\xi_2)-f_2(\Theta',\xi_2)$
4
Aggregation
To Broadcast
Global Model
Device N
$f_N(\Theta,\xi_N)-f_N(\Theta',\xi_N)$
Servers

Figure 1: An overview of the proposed federated learning scheme over wireless networks.

these devices have limited transmission power. Thus, assuming that these devices can upload concurrently the information demanded by FL may be unrealistic. The downlink, on the other hand, is not affected by this problem, as the server can have much bigger transmission power. To address this uplink bottleneck, various strategies have been proposed. Some advocate for local multiple gradient descent steps to be carried out by the devices before sending their gradients back to the server, thereby conserving communication resources (Khaled et al., 2020). Others suggest enabling partial device participation in each iteration (Chen et al., 2018; Amiri et al., 2021), or a combination of both strategies (McMahan et al., 2017). Additionally, alternative approaches involve the utilization of lossy compression techniques on the gradients prior to uploading. For instance, in the works of Konečný et al. (2016), Khirirat et al. (2018), and Elgabli et al. (2020), stochastic unbiased quantization methods are recommended, where gradients are approximated using a finite set of discrete values to enhance efficiency. Mishchenko et al. (2019) introduce gradient quantization differences between the current and previous iterations, allowing the update to incorporate new information. Conversely, Chen et al. (2022) propose gradient sparsification of this difference, meaning that vector components below a certain threshold are not transmitted.

On the other hand, our proposed method counters the bottleneck by having every user upload only a scalar value as feedback instead of a long vector to the wireless medium, saving a factor of $O(d)$ of transmission size per communication round. The trick is to assemble the gradient estimator vector at the server's size instead of having each user estimate it. Users are alternatively expected to query their model for the loss value and return this scalar loss.

Wireless Disturbance. When data is transmitted through a wireless medium, it is vulnerable to distortions induced by the medium itself. These distortions go beyond simple additive noise and are, in fact, a consequence of thermal fluctuations occurring at the receiver. Essentially, the wireless channel operates as a filter for the transmitted signal (Tse and Viswanath, 2005; Björnson and Sanguinetti, 2020),

$$\hat{x} = \hat{H}x + \hat{n}. \tag{1}$$

x represents the signals sent while $\hat{x}$ represents the signals received, both belonging to $\mathbb{C}^d$. Additionally, we have stochastic and ever-changing entities: the channel matrix $\hat{H} \in \mathbb{C}^{d\times d}$, and the additive noise vector $\hat{n} \in \mathbb{C}^d$. In the context of FL, the variable x can represent either the model itself or its gradients transmitted through the channel. The challenge of FL involves mitigating the channel's impact, necessitating a detailed analysis and removal of each channel element to recover the transmitted information and entails the exchange of control or reference signals between the devices and the server at each iteration. This analytical process is resource-intensive in terms of computation and time. In various works, including that of Yang et al. (2020); Fang et al. (2022); Amiri and Gündüz (2020); Sery and Cohen (2020); Guo et al. (2021); Sery et al. (2021), and references in this domain, the assumption that instantaneous channel knowledge is available/obtainable prevails in an attempt to eliminate the channel's impact.

In contrast, our approach offers a much more straightforward solution. We forego the resource-intensive channel analysis and instead integrate the channel directly into the learning process. It becomes an inherent part of our implementation, allowing us to construct gradient estimates without the need to eliminate its influence. This approach not only saves computational resources but also reduces the communication overhead.

System Heterogeneity. System heterogeneity in FL networks, where devices have varied computational power, network bandwidth, battery life, and storage, poses significant challenges. Traditional algorithms struggle with delays caused by slow clients or straggling nodes, as every device participates in training regardless of capability. Consequently, the server must wait for the slowest nodes to finish their updates, significantly hindering the process. For instance, Gu et al. (2021) suggest using outdated model parameters for users who are delayed and Reisizadeh et al. (2022) initiate the training process with the faster nodes and progressively incorporate the slower ones into the model training once the current nodes' statistical accuracy is achieved.

Contrarily to conventional methods that place significant demands on the computational capabilities of participating devices, our approach is notably less taxing in this regard. When the devices receive the global model, their sole task is to query this model with their data and subsequently transmit the resulting scalar loss. As a result, the computational load associated with the ”backward pass” is eliminated, and only the ”forward pass” is carried out.

Black-Box Optimization. A key rationale for the application of ZO methods lies in addressing black-box problems within FL (Fang et al., 2022), where obtaining gradient information is either infeasible or computationally challenging. One such scenario is evident in hyperparameter tuning, where gradient calculations are unattainable due to the absence of an analytical relationship between the loss function and the hyperparameters (Dai et al., 2020).

1.2 Contribution

In this work, we offer an alternative framework with much less computational and communication expenses. By incorporating the wireless disturbance in the learning itself, we propose a novel ZO method optimizing resource utilization and overcoming the bottleneck caused by the limited capacity of uplink transmissions. Our method is not simple exten-

Table 1: Convergence rates for the various cases studied in this paper, classified according to the nature of the objective function, the environment, and the step sizes.

OBJECTIVE FUNCTION	**SETTING**	**STEP SIZES**	**CONVERGENCE RATE**
Strongly Convex & Smooth	Noise-Free	vanishing	$O(\frac{1}{K})$
	Noisy	vanishing	$O(\frac{1}{\sqrt{K}})$
	Noise-free & Noisy	constant	$O(\lambda^K)$
Non-Convex & Smooth	Noise-Free	vanishing	$O(\frac{1}{\sqrt{K}})$
	Noisy	vanishing	$O(\frac{1}{\sqrt[3]{K}})$
κ-Gradient Dominated & Smooth	Noise-Free	vanishing	$O(\frac{1}{K})$
	Noisy	vanishing	$O(\frac{1}{\sqrt{K}})$
	Noise-free & Noisy	constant	$O(\lambda^K)$

sion from FO to ZO, like in the work of Fang et al. (2022) where each device still has to upload the whole ZO vector. Instead, each device sends two scalar values, one of which represents the loss function at two perturbed points, resulting in a huge reduction in the amount of information sent in the uplink. The aim thus is to perturb the objective function to estimate the gradient, similarly to usual ZO estimation methods. This is done in our method by using the perturbation introduced by the wireless environment in a judicious way. We then provide a comprehensive analysis, proving the almost sure convergence of our method with convex and non-convex objective function. An important distinction here is that typical ZO methods focus on the expected convergence, but our approach goes further by demonstrating the almost sure convergence. The key to this proof lies in the application of Doob's martingale inequality to bound the stochastic error arising from gradient estimates. Afterwards, we provide theoretical bounds on the convergence rates for strongly convex, non-convex, and κ-gradient dominated non-convex objective functions, highlighting the effect of the noise of the environment on the performance of the algorithm. The presence of noise causes the increase of the gradient estimate variance which leads to the slowing down of convergence.

In our previous paper (2024), a single point gradient estimate is considered but under the restrictive assumption that the objective function is bounded everywhere. It is worth mentioning that single-point estimates are generally more limiting in terms of assumptions and the best achievable convergence rate due to their bigger variance. In this paper, we remove the restrictive boundedness assumption of the objective function and consider rather a standard Lipschitz continuity. We also adopt a two-point gradient estimate and modify the proposed algorithm accordingly while maintaining the same uplink communication efficiency. We further extend the analysis as compared to our previous work to encompass both convex and non-convex settings. We considerably improve the convergence rate, where even a linear rate is established, by examining cases with constant and vanishing step sizes and underscoring the impact of noise. It is worth noticing that proving a linear rate for ZO method is an interesting result in itself. We summarize the convergence rates of our method in Table 1. The bounds we find compete with optimal rates found in the literature, sometimes even with full gradient information. These are interesting findings, as ZO methods are known to have worse performance than their FO counterparts.

To elaborate, in the strongly convex and smooth case, the optimal convergence rate with *full gradient information* has been established to be $O(\frac{1}{K})$ with vanishing step-sizes and linear $O(\lambda^K)$ with constant step sizes (Pu and Nedić, 2018; Nemirovski et al., 2009) where K represents the number of iterations; These rates match that of our ZO method under vanishing (noise-free) and constant step sizes, respectively (the first and third entries in Table 1).

In the context of noisy function queries in a *centralized* setting with strongly convex and smooth objectives, it has been demonstrated that gradient-free methods cannot do better than $\Omega(\frac{1}{\sqrt{K}})$ (Duchi et al., 2015; Jamieson et al., 2012; Shamir, 2013; Akhavan et al., 2020). Our method, which is distributed, achieves the same convergence rate under the same setting, i.e., with noise and vanishing step sizes (the second entry in the table). For the bounded noise setting, interesting results exist in the literature (Gasnikov et al., 2023; Akhavan et al., 2021) where our bound matches the rate $O(\frac{1}{\sqrt{K}})$ established for twice differentiable functions with more function queries (2d-point estimate) in Akhavan et al. (2021)'s work; however, in this work, we consider a Gaussian additive noise and an additional non-additive disturbance exists due to the presence of transmission channel.

In the smooth non-convex case, a rate of $O(\frac{1}{\sqrt{K}})$ is shown in the noise-free ZO *centralized* setting (Nesterov and Spokoiny, 2015), which matches our method's rate in the fourth entry.

Whereas we are unaware of existing literature on centralized noisy non-convex ZO settings, the following rates exist for *centralized* non-convex ZO methods: $O(\sqrt{\frac{W_K}{K}})$ with a two-point gradient estimator and $O(\sqrt[3]{\frac{W_K}{K}})$ with a one-point estimator (Roy et al., 2022) with W_K a bound on the amount of nonstationarity that is allowed to increase with K. The other rate is $O(\frac{1}{\sqrt[4]{K}})$ for a two-point estimate (Balasubramanian and Ghadimi, 2018). All of which seem slightly worse than our fifth entry. A single-point-based ZO distributed method is explored in our previous work (2023; 2022), which was shown to converge with a rate of $O(\frac{1}{\sqrt[3]{K}})$ under the goal of achieving consensus in a noisy, non-convex setting. To achieve this goal, we adopted a gradient tracking technique. Gradient tracking is a distinct algorithm that involves sharing between neighbors and updating two variables: the optimization variable and an auxiliary one. The auxiliary variable has the role of tracking the average gradient between the agents without explicitly sharing the gradients. While this algorithm generally accelerates gradient-based techniques, it has a worse communication cost as each user must send a long vector of $2d$ values, whereas here, only 2 scalar values are sent per user. Furthermore, contrary to this work, the gradient tracking algorithm's performance is dependent on the network architecture, and channel is not included in the algorithm. This considerably simplifies the analysis regarding the gradient estimate and its bias as they're independent of the channel variables and the intertwined stochastic associations between them.

In addition to the general non-convex settings, we provide a study on the convergence rate under a non-convex κ-gradient dominated objective (Polyak, 1963; Lojasiewicz, 1963) with and without noise. The κ-gradient domination property is generally viewed as a non-convex analogy of strong convexity, and this explains the identical convergence rates to that in the strongly convex case (sixth, seventh, and eighth entries).

Finally, we provide numerical evidence of the efficiency of our algorithm. In the training example, we provide the margin by which our algorithm is more communicationaly efficient as compared with its FO counterparts (FedAvg, for example). It might even be infeasible for the standard method to achieve the same reduction in communication as our method.

2 Federated Learning Framework

We consider a federated setting where N agents (e.g., mobile devices) collaborate and communicate with a server over a wireless channel. Each agent has access to its local data and performs computations independently, optimizing its local objective function using its data and its computational resources. The next step is for each agent to communicate its local updates to the server. The server then aggregates the received information from all agents weighted by the channel coefficients. The server then takes the necessary measure to update the global model or parameters and communicates it back to all agents, and the process repeats. To that end, let $\mathcal{N} = \{1, ..., N\}$ be the set of agents in the network and $\theta \in \mathbb{R}^d$ denote the optimization parameters. The objective is to minimize a loss function $F : \mathbb{R}^d \to \mathbb{R}$ that is composed of the said agents' loss functions $F_i : \mathbb{R}^d \to \mathbb{R}$ for $i \in \mathcal{N}$, such that

$$\min_{\theta \in \mathbb{R}^d} F(\theta) := \frac{1}{N} \sum_{i=1}^{N} F_i(\theta) \tag{2}$$

where

$$F_i(\theta) = \mathbb{E}_\xi f_i(\theta, \xi) \tag{3}$$

with $\xi \in \mathcal{X}$ denoting an i.i.d. ergodic stochastic process describing uncertainties in the communication system or variations in the data distributions. We further consider the case where the devices do not have access to their gradients for computational and communication restraints, and they must estimate this gradient by querying their model only once per update. They obtain a scalar value from this query, that they must send back to the server.

Throughout this paper, we consider twice continuously differentiable local objective functions and we assume problem (2) has a solution $\theta^* \in \mathbb{R}^d$ where $\nabla F_i(\theta^*) = 0$, $\forall i \in \mathcal{N}$. We also consider functions that satisfy the following two conditions.

Assumption 1 *We assume that the local Hessian is bounded above by a constant $\beta_1 \in \mathbb{R}^+$,*

$$\|\nabla^2 F_i(\theta)\|_2 \leq \beta_1, \ \forall i \in \mathcal{N}. \tag{4}$$

This assumption implies that all local objective functions are L-smooth. In addition, all locally queried functions are Lipschitz continuous,

Assumption 2 *Let $L_\xi > 0$ be a Lipschitz constant. Then,*

$$|f_i(\theta_1, \xi) - f_i(\theta_2, \xi)| \leq L_\xi \|\theta_1 - \theta_2\|, \; \forall \theta_1, \theta_2 \in \mathbb{R}^d, \forall i \in \mathcal{N}. \tag{5}$$

Lemma 3 *Let Assumption 2 hold. By applying Jensen's inequality, we find that $F(\theta)$ is also Lipschitz continuous with some constant $L' = \mathbb{E}[L_\xi] > 0$.*

3 The 2P-ZOFL algorithm

We consider a scenario where there exists an intermediary wireless environment between a central server and multiple distributed agents denoted by $\mathcal{N}$ and indexed by $i \in \mathcal{N}$. The wireless channels introduce stochastic scaling on the signals transmitted from each agent to the server, as described by equation (1). Our algorithm is based on obtaining a gradient estimate of $F(\theta)$ by perturbing the function and via a smart exchange between the server and the users. In other words, the upload wireless channels in the network will be part of the function's perturbation. Before describing the algorithm, we stress that the exchange from the users to the server are real scalars and not vectors. The content of x will be explained later on in this section. At the receiver, we do not remove the impact of the channel in order to decode correctly. Instead, the receiver will directly use the real part of the received signal, as follows

$$\Re[\hat{x}] = \Re[\hat{H}]x + \Re[\hat{n}], \tag{6}$$

where $\Re[\cdot]$ denotes the real part. To better understand the wireless transmission modeling, we refer the interested reader to Appendix A of our work (2024). In all that follows, we use the notation H or h to specify the real part of the channel and take into account the real part of the received signal only, meaning the phase channel is already included in the perturbation. At time slot k, we denote by $h_{i,k}$ the real part of user i's channel coefficient. The channel coefficients are autocorrelated from one time slot to the next, with $\mathbb{E}[h_{i,k}h_{i,k+1}] = K_{hh}$ for all i and k.

Our proposed optimization method denoted as 2P-ZOFL and described in Algorithm 1, involves two communication steps. In the first step, each agent sends a predefined scalar value a to the server. According to the channel characteristics, the real part of the signal received by the server from all agents is $\sum_{j=1}^N h_{j,k}a + n_{j,k}$. After obtaining these values in step 2, the server utilizes them to adjust the optimization parameter vector twice adding and subtracting the previously aggregated received values (and averaged by N) multiplied by γ_k and ω_k. These two vectors are then broadcasted to all agents. Contrary to the uplink, the impact of the downlink channel is then removed. Upon receiving these new parameters, each agent i queries its local loss function f_i at both vectors to obtain the difference of the losses $f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N} \sum_{j=1}^N (h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big) - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N} \sum_{j=1}^N (h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big)$, which is again a scalar. Here, ξ is a stochastic process that denotes the data distribution, noise or errors at the receiver (agent). In the second communication step, the agents send this scalar difference back to the server. Again, the server takes the real part of the received signal which is indicated in step 5. Finally, the server assembles the gradient estimate g_k from the received information and updates the optimization parameter θ in step 7.

We define α_k and γ_k as two step-sizes, along with $\omega_k \in \mathbb{R}^d$, a perturbation vector generated by the server with the same dimension as that of θ_k.

It is important to note that in this learning method, the impact of the uplink wireless channel is included in the gradient estimate g_k and, therefore, influences the optimization technique. The main advantage of this algorithm is its communication efficiency. Each agent only needs to send two scalar values, which is a significant improvement over standard distributed algorithms that require transmitting the entire optimization vector or local gradient of dimension d from each entity. This communication efficiency is crucial, as it avoids excessive resource consumption and is more realistic in practical scenarios.

Algorithm 1 The 2P-ZOFL algorithm

```
Input: Initial model $\theta_0 \in \mathbb{R}^d$, the initial step-sizes $\alpha_0$ and $\gamma_0$, and the scalar value $a$
for $k = 0$ to $K - 1$ do
    The real part of the signal received by the server is $\frac{1}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})$
    The server performs the two actions
    $\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})$ and $\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})$
    The server broadcasts these actions to all devices under the same stochastic wireless environment
    The real part of the signal again received by the server is
    $\frac{1}{N}\sum_{i=1}^{N}\Bigg(h_{i,k+1}\bigg[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big) - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big)\bigg] + n_{i,k+1}\Bigg)$
    The server multiplies the received scalar sum by $\omega_k$ to assemble $g_k$ given in (7)
    The server updates $\theta_{k+1} = \theta_k - \alpha_k g_k$
end for
```

3.1 Two-Point Gradient Estimator

In this subsection, we present and analyze the following proposed gradient estimate influenced by the wireless channel,

$$g_k = \frac{\omega_k}{N}\Bigg(\sum_{i=1}^{N} h_{i,k+1}\bigg[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big) - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{i,k+1}\Big)\bigg] + n_{i,k+1}\Bigg). \tag{7}$$

In our proposed method, we acknowledge that the channel coefficients $h_{i,k}$ and $h_{i,k+1}$ may not be known explicitly. This approach offers significant advantages, as it reduces computation complexity and greatly improves communication efficiency. Unlike conventional methods that require continuous transmission of pilot signals to estimate the channel, our approach circumvents this need.

In some scenarios, the noise at the reception is negligible, the environment is said to be noise-free. Our gradient estimate can then be simplified to

$$g_k = \frac{\omega_k}{N}\Bigg(\sum_{i=1}^{N} h_{i,k+1}\Bigg[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a, \xi_{i,k+1}\Big) - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a, \xi_{i,k+1}\Big)\Bigg]\Bigg), \tag{8}$$

where the noise at both reception steps of the algorithm is neglected. This case is interesting to study as it has remarkable effect on the convergence rate.

We assume $h_{i,k}$ to be a zero-mean random variable with standard deviation σ_h, $\forall i \in \mathcal{N}, \forall k \in \mathbb{N}^+$ and we consider the following assumptions

Assumption 4 *(on the additive noise) The noise term $n_{i,k}$ is assumed to have a zero mean and be uncorrelated with bounded variance. This means that its expected value is zero, and its expected squared value is finite and bounded by σ_n^2. This holds true for all devices i and for any time slot k, and there is no correlation between the noise terms of different devices at the same time slot, i.e., $E(n_{i,k}n_{j,k}) = 0$ when $i \neq j$. Additionally, for any given device i, the noise terms at different time slots are uncorrelated, i.e., $E(n_{i,k}n_{i,k'}) = 0$ when $k \neq k'$.*

Assumption 5 *(on the random perturbation) Let $\omega_k = (\omega_k^1, \omega_k^2, \ldots, \omega_k^d)^T$. At each iteration i, the server generates its ω_k vector independently from other iterations. In addition, the elements of ω_k are assumed i.i.d with $\mathbb{E}(\omega_k^{d_1}\omega_k^{d_2}) = 0$ for $d_1 \neq d_2$ and there exists $\beta_2 > 0$ such that $\mathbb{E}(\omega_k^{d_j})^2 = \beta_2$ $\forall d_j$, $\forall i$. We further assume there exists a constant $\beta_3 > 0$ where $\|\omega_k\| \leq \beta_3$, $\forall k$.*

Example 1 *An example of a perturbation vector satisfying Assumption 5, is picking every dimension of ω_k from $\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}$ with equal probability. Then, $\beta_2 = \frac{1}{d}$ and $\beta_3 = 1$.*

We let $\mathcal{H}_k = \{\theta_0, \xi_0, \theta_1, \xi_1, ..., \theta_k, \xi_k\}$ designate the history sequence resulting from applying the algorithm and we denote by $\mathbb{E}[.|\mathcal{H}_k]$ the conditional expectation given $\mathcal{H}_k$.

Lemma 6 *Let Assumptions 4 (when there is noise) and 5 be satisfied and define the scalar $c_1 = \frac{2a\beta_2 K_{hh}}{N}$, then the proposed gradient estimator is biased w.r.t. the objective function's exact gradient $\nabla F(\theta)$. Concretely,*

- *When the environment is noise-free, we have $\mathbb{E}[g_k|\mathcal{H}_k] = c_1\gamma_k(\nabla F(\theta_k) + b_k)$,*
- *When the environment is noisy, similarly, we have $\mathbb{E}[g_k|\mathcal{H}_k] = c_1\gamma_k(\nabla F(\theta_k) + b'_k)$,*

$\forall k \in \mathbb{N}^+$, where b_k and b'_k are the bias terms differing slightly in form.

Proof: Refer to Appendices A.1.1 and A.1.2.

Lemma 7 *Let Assumptions 2, 4 (when appropriate), and 5 hold. There exists two bounded constants $c_2, c'_2 > 0$, such that*

- *For a noise-free environment, $\mathbb{E}[\|g_k\|^2|\mathcal{H}_k] \leq c_2\gamma_k^2$,*
- *For a noisy environment, $\mathbb{E}[\|g_k\|^2|\mathcal{H}_k] \leq c'_2$,*

where c_2 evolves as $O(\frac{1}{N})$ and c_2' evolves as $O(1)$ in terms of N and as $O(\sigma_n^2)$.
Proof: Refer to Appendices A.3.1 and A.3.2.

The previous lemma highlights the stark difference between noise-free and noisy environments in the analysis context, as the consequence is a much smaller convergence rate when the noise is present, as shown in the following parts.

Lemma 8 *By Assumptions 5 and 1, we can find two scalar values $c_3, c_3' > 0$ such that*

$$\|b_k\| \leq c_3\gamma_k \quad \text{and} \quad \|b_k'\| \leq c_3'\gamma_k, \tag{9}$$

where c_3 evolves as $O(dN)$ and c_3' evolves as $O(dN)$ and as $O(\sigma_n^2)$, for $\beta_2 = \frac{1}{d}$ and $\beta_3 = 1$.
Proof: Refer to Appendices A.2.1 and A.2.2.

4 Convergence Analysis

In this section, we analyze our algorithm under different settings. We prove it converges almost surely with a general non-convex objective function. We then derive its convergence rate and extended it to the case of κ-gradient dominated objectives with vanishing and constant step sizes. Afterwards, we prove the almost sure convergence with strictly convex objective. We then study the convergence rate under strong convexity. When we fix the step sizes, we prove that a linear rate towards a neighborhood of the optimum (local optimum in the nonconvex case) is possible.

For all what follows with vanishing step sizes, the following assumption is vital for convergence.

Assumption 9 *Both $\alpha_k \to 0$ and $\gamma_k \to 0$ as $k \to \infty$ and we assume $\sum_{k=0}^{\infty} \alpha_k\gamma_k = \infty$.*

- *When the environment is noiseless, we further assume $\sum_{k=0}^{\infty} \alpha_k^2\gamma_k^2 < \infty$.*
- *We replace the previous assumption by $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ for noisy environments.*

Example 2 *We consider the following form of the step sizes, $\alpha_k = \alpha_0(l+k)^{-\upsilon_1}$ and $\gamma_k = \gamma_0(l+k)^{-\upsilon_2}$ with $\alpha_0, \gamma_0, l, \upsilon_1, \upsilon_2 > 0$. To satisfy Assumption 9, it is sufficient to find υ_1 and υ_2 such that $0 < \upsilon_1 + \upsilon_2 \leq 1$ and $\upsilon_1 + \upsilon_2 > 0.5$ for noise-free environments ($0 < \upsilon_1 + \upsilon_2 \leq 1$ and $\upsilon_1 > 0.5$ for noisy environments).*

We then introduce the stochastic error denoted as e_k. It represents the difference between an individual realization of g_k and its expected value given the historical sequence, expressed as

$$e_k = g_k - \mathbb{E}[g_k|\mathcal{H}_k].$$

The examination of this noise and its evolution plays a crucial role in analyzing the algorithm. It allows us to access the exact gradient when studying the algorithm's convergence behavior. Furthermore, it enables us to demonstrate that the exact gradient indeed converges to zero, not just its expected value. This constitutes a more robust convergence property that, to the best of our knowledge, has not been previously explored in ZO non-convex optimization. The key insight lies in proving that e_k behaves as a martingale difference sequence and then applying Doob's martingale inequality to establish the following lemma.

Lemma 10 *If all Assumptions 2, 4 (when applicable), 5, and 9 hold, then for any constant $\nu > 0$, we have*

$$\mathbb{P}(\lim_{K\to\infty} \sup_{K'\geq K} \| \sum_{k=K}^{K'} \alpha_k e_k \| \geq \nu) = 0, \ \forall \nu > 0.$$

Proof: Refer to Appendix B.

A side note is that the previous lemma holds when the environment is noise-free and noisy.

4.1 Convergence Analysis with a Non-Convex Objective Function

The only constraints on the objective function are that of Lipschitz continuity and smoothness given in Assumptions 1 and 2 in the introduction.

In this part, we add the following assumption on the step-sizes.

Assumption 11 *α_k and γ_k further satisfy $\sum_{k=0}^{\infty} \alpha_k \gamma_k^3 < \infty$.*

Theorem 12 *When Assumptions 1, 2, 4 (when necessary),5, 9, and 11 hold, we have*

$$\sum_k \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 < +\infty \quad \textit{implying} \quad \lim_{k\to\infty} \|\nabla F(\theta_k)\| = 0, \textit{ almost surely.} \tag{10}$$

Proof: Refer to Appendix C.

Starting from the smoothness inequality, we substitute by the algorithm's updates and make use of the stochastic noise. We then perform a telescoping summation over the iterations $k > 0$ and use Doob's martingal inequality, the conditions on the step sizes, and the upper bound estimate's squared norm. This approach allows us to find an upper limit on the expression $\sum_k \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2$ when k grows to ∞. This result guarantees that the infimum of $\|\nabla F(\theta_k)\|$ converges to 0 as k approaches ∞. The subsequent step thus involves considering the hypothesis $\lim_{k\to\infty} \sup \|\nabla F(\theta_k)\| \geq \rho$, for $\rho > 0$, and prove that it contradicts with the initial result. This theorem holds true in both noiseless and noisy settings, as demonstrated in the proof.

4.2 Convergence Rate with a General Non-Convex Objective Function

Let $\delta_k = F(\theta_k) - F(\theta^*)$ be the function optimality gap. The subsequent theorem identifies the convergence rate of the algorithm under both settings, the noise-free and the noisy one.

Theorem 13 *In addition to the assumptions of Theorem 12, let the step sizes have the form of Example 2 with $l = 1$ and $\upsilon_3 = \upsilon_1 + \upsilon_2 < 1$. Then,*

- *When the channels are noise-free,*

$$\frac{\sum_k \alpha_k \gamma_k \mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]}{\sum_k \alpha_k \gamma_k} \leq \frac{(1-\upsilon_3)}{(K+2)^{1-\upsilon_3} - 1}\bigg(A_0 + \frac{A_1}{\upsilon_1 + 3\upsilon_2 - 1} + \frac{A_2}{2\upsilon_3 - 1}\bigg). \tag{11}$$

- *When the channels are noisy,*

$$\frac{\sum_k \alpha_k\gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2]}{\sum_k \alpha_k\gamma_k} \leq \frac{(1-\upsilon_3)}{(K+2)^{1-\upsilon_3}-1}\left(A_0 + \frac{A_1'}{\upsilon_1+3\upsilon_2-1} + \frac{A_2'}{2\upsilon_1-1}\right). \qquad (12)$$

where $A_0 = \frac{2\delta_0}{c_1\alpha_0\gamma_0}$*,* $A_1 = (\upsilon_1+3\upsilon_2)(c_3\gamma_0)^2$*,* $A_2 = \frac{2\upsilon_3 c_2\alpha_0\gamma_0 L}{c_1}$*,* $A_1' = (\upsilon_1+3\upsilon_2)(c_3'\gamma_0)^2$*, and* $A_2' = \frac{2\upsilon_1 c_2'\alpha_0 L}{c_1\gamma_0}$*.*

Proof: Refer to Appendices C.1.1 and C.1.2.

In Theorem 13, the optimal choice of exponents in equation (11) is $\upsilon_1 = \upsilon_2 = \frac{1}{4}$, resulting in a rate of $O\left(\frac{1}{\sqrt{K}}\right)$. However, to prevent the constant part from becoming excessively large, we identify a very small value $\epsilon > 0$ such that $\upsilon_1 = \upsilon_2 = \frac{1}{4} + \frac{\epsilon}{2}$, leading to a rate of $O\left(\frac{1}{K^{\frac{1}{2}-\epsilon}}\right)$. By substituting the values of the scalars c_1, c_2 and c_3, we note that the bound of (11) evolves as $O(d^2N^2)$ in terms of dimension of the problem and number of agents and as $O(L)$ in terms of the objective's smoothness.

Similarly, in equation (12), the optimal choice is $\upsilon_1 = \frac{1}{2}$ and $\upsilon_2 = \frac{1}{6}$, resulting in a rate of $O\left(\frac{1}{\sqrt[3]{K}}\right)$. A practical alternative is to select $\upsilon_1 = \frac{1}{2} + \frac{\epsilon'}{2}$ and $\upsilon_2 = \frac{1}{6} + \frac{\epsilon'}{2}$ for a rate of $O\left(\frac{1}{K^{\frac{1}{3}-\epsilon'}}\right)$, with $\epsilon' > 0$. In (12), the bound also evolves as $O(d^2N^2)$, $O(L)$, and $O(\sigma_n^4)$ in terms of noise variance.

We must remark that the higher dependence on N in all our bounds is due to the fact that we have double aggregation of received information from users. This is done to deal with the impact of the channel and it has definitely an impact on the bound and it can be seen as a cost to pay to deal with the impact of the channel. One can see that if we remove the interior summation (first transmission), the dependence on N would completely disappear. In addition, due to the fact that there is no explicit form of dependence between $\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k)$ and the other stochastic variables (i.e., $h_{i,k+1}$, $h_{j,k}$, and ω_k), it is difficult to compute the expectation in b_k and settle for bounding its norm, where we establish an upper bound that might be loose in terms of N. For quadratic objectives, for example, both c_3 and c_3' evolve as $O(\frac{1}{N})$, improving the bounds for all convergence rates.

4.3 Convergence Rate with a κ-Gradient Dominated Non-Convex Objective Function

In this subsection, we consider functions satisfying the following property alongside Assumptions 1 and 2.

Assumption 14 *We assume that the objective function is κ-gradient dominated non-convex, i.e., it admits a constant $\kappa > 0$ such that (Polyak, 1963; Lojasiewicz, 1963)*

$$2\kappa(F(\theta) - F(\theta^*)) \leq \|\nabla F(\theta)\|^2, \ \ \forall \theta \in \mathbb{R}^d. \qquad (13)$$

Theorem 15 *In addition to the assumptions of Theorem 12, let the step sizes have the form of Example 2 with $l = 1$ and $\upsilon_1 + \upsilon_2 = 1$. Let $\upsilon_4 = c_1\alpha_0\gamma_0\kappa$. Then, when the objective function satisfies Assumption 14, we get*

- *When the environment is noise-free*

$$\mathbb{E}[\delta_K] \leq \frac{1}{(K+2)^{\upsilon_4}}\left(B_0 + B_1\sum_{k=0}^{K-1}\frac{(k+3)^{\upsilon_4}}{(k+2)^{\upsilon_1+3\upsilon_2}} + B_2\sum_{k=0}^{K-1}\frac{(k+3)^{\upsilon_4}}{(k+2)^{2\upsilon_1+2\upsilon_2}}\right) \tag{14}$$

- *When the environment is noisy*

$$\mathbb{E}[\delta_K] \leq \frac{1}{(K+2)^{\upsilon_4}}\left(B_0 + B_1'\sum_{k=0}^{K-1}\frac{(k+3)^{\upsilon_4}}{(k+2)^{\upsilon_1+3\upsilon_2}} + B_2'\sum_{k=0}^{K-1}\frac{(k+3)^{\upsilon_4}}{(k+2)^{2\upsilon_1}}\right) \tag{15}$$

where $B_0 = 2^{\upsilon_4}\delta_0$*,* $B_1 = \frac{c_1c_3^2\alpha_0\gamma_0^3}{2}$*,* $B_1' = \frac{c_1c_3'^2\alpha_0\gamma_0^3}{2}$*,* $B_2 = \frac{c_2\alpha_0^2\gamma_0^2L}{2}$*,* $B_2' = \frac{c_2'\alpha_0^2L}{2}$*.*

Proof: Refer to Appendices C.2.1 and C.2.2.

In (14), for the bound to converge, we need $\upsilon_1+3\upsilon_2-\upsilon_4>1$ and $2\upsilon_1+2\upsilon_2-\upsilon_4>1$. Then, the optimal choice is $\upsilon_1=\upsilon_2=\frac{1}{2}$ and $\upsilon_4=1-\epsilon$ for a rate of $O(\frac{1}{K^{1-\epsilon}})$ and some $\epsilon>0$.

Similarly, in (15), the conditions of convergence are $\upsilon_1+3\upsilon_2-\upsilon_4>1$ and $2\upsilon_1-\upsilon_4>1$. The optimal choice of step sizes is $\upsilon_1=\frac{3}{4}$ and $\upsilon_2=\frac{1}{4}$ with $\upsilon_4=\frac{1}{2}-\epsilon'$ for a rate of $O(\frac{1}{K^{\frac{1}{2}-\epsilon'}})$ and some $\epsilon'>0$.

Assuming that α_0 and γ_0 evolve as $O(\sqrt{\frac{1}{c_1\kappa}})$ and thus as $O(\sqrt{\frac{dN}{\kappa}})$, in both (14) and (15), the bounds evolve as $O(\frac{d^3N^3}{\kappa^2})$ and $O(L)$, and the noise dependence in (15) is similar to the previous general non-convex case where it is $O(\sigma_n^4)$. This worse dependence on d and N is due to the conditions on α_0 and γ_0, whereby the dependence on the number of iterations K is improved but with the price of an additional $O(dN)$. The κ-gradient domination property, however, improves both the dependence on K and the constant terms.

4.4 Convergence Rate with a κ-Gradient Dominated Non-Convex Objective Function and Fixed Step Sizes

In this subsection, we fix the step sizes and consider the same assumptions on the objective function as in the previous one. We then study the convergence rate in the following theorem.

Theorem 16 *Let the assumptions of Theorem 12 hold and let the objective function satisfy Assumption 14. Fix* $\alpha_k=\alpha>0$ *and* $\gamma_k=\gamma>0$ *for all* $k\geq 0$*. Then, for* $\varsigma = 1-c_1\alpha\gamma\kappa$ *and* $\alpha\gamma<\frac{1}{c_1\kappa}$*,*

- *In a noise-free environment,*

$$\mathbb{E}[\delta_{K+1}] \leq \varsigma^{K+1}\delta_0 + \alpha\gamma\left(\frac{c_1c_3^2}{2}\gamma^2 + \frac{c_2L}{2}\alpha\gamma\right)\frac{1-\varsigma^{K+1}}{1-\varsigma}. \tag{16}$$

- *In a noisy environment,*

$$\mathbb{E}[\delta_{K+1}] \leq \varsigma^{K+1}\delta_0 + \alpha\left(\frac{c_1c_3'^2}{2}\gamma^3 + \frac{c_2'L}{2}\alpha\right)\frac{1-\varsigma^{K+1}}{1-\varsigma}. \tag{17}$$

Proof: Refer to Appendices C.3.1 and C.3.2.

Knowing that $\varsigma<1$, then for an arbitrarily small value of the step sizes α and γ, we can say that the algorithm converges to a neighborhood of the local optimum with a linear rate $O(\varsigma^K)$. Since $\alpha\gamma$ may be taken arbitrarily much smaller than $O(\frac{1}{c_1})$, i.e. $O(dN)$, an improved dependence on the dimension and the number of agents can be established, where the bounds now evolve as $O(\frac{dN}{\kappa^2})$ while the dependencies $O(L)$ and $O(\sigma_n^4)$ remain.

4.5 Convergence with Convex Objective Function

In this subsection, we analyze the asymptotic behavior of Algorithm 1 with objective functions satisfying Assumption 1 and the following assumptions.

Assumption 17 *α_k and γ_k further satisfy $\sum_{k=0}^{\infty} \alpha_k \gamma_k^2 < \infty$.*

Assumption 18 *Let the objective function be strictly convex, i.e., satisfying*

$$F(\theta_1) > F(\theta_2) + \langle \nabla F(\theta_2), \theta_1 - \theta_2 \rangle, \ \ \forall \theta_1, \theta_2 \in \mathbb{R}^d. \tag{18}$$

As we analyze the convergence rate in the subsequent parts for strongly convex functions, the Lipschitz continuity in Assumption 2 can no longer hold. We thus replace it by the local Lipschitz continuity in the following assumption. To guarantee that the users are able to compute and send feasible loss values in step 5 of Algorithm 1, the server must thus project the updated parameter vector onto a compact convex set $\mathcal{K}$, i.e., step 7 of Algorithm 1 becomes $\theta_{k+1} = \Pi_{\mathcal{K}}(\theta_k - \alpha_k g_k)$, where $\Pi_{\mathcal{K}}(\cdot)$ denotes the Euclidean projection of a vector on the set $\mathcal{K}$.

Assumption 19 *$\mathcal{K}$ is a compact convex set and $\theta^* \in \mathcal{K}$. All local functions $\theta \mapsto f_i(\theta, \xi)$ are locally Lipschitz continuous on the $\beta_3\gamma_0$-neighborhood of $\mathcal{K}$, i.e.,*

$$|f_i(\theta_1, \xi) - f_i(\theta_2, \xi)| < L_\xi \|\theta_1 - \theta_2\|, \ \ \forall \theta_1, \theta_2 \in N_{\beta_3\gamma_0}(\mathcal{K}), \forall \xi \in \mathcal{X}, \forall i \in \mathcal{N},$$

where $N_{\beta_3\gamma_0}(\mathcal{K}) = \{\theta \in \mathbb{R}^d | \inf_{a \in \mathcal{K}} \|\theta - a\| < \beta_3\gamma_0\}$ is the $\beta_3\gamma_0$-neighborhood of $\mathcal{K}$.

We further remark that the projection on a closed convex set $\mathcal{K}$ is -xpansive (Kinderlehrer and Stampacchia, 2000, Corollary 2.4), i.e.,

$$\|\Pi_{\mathcal{K}}(\theta) - \Pi_{\mathcal{K}}(\theta')\| \leq \|\theta - \theta'\|, \ \ \forall \theta, \theta' \in \mathbb{R}^d. \tag{19}$$

For any integer $k \geq 0$, we define the divergence, or the error between the model set by the server θ_k and the optimal solution θ^* as

$$d_k = \|\theta_k - \theta^*\|^2. \tag{20}$$

The following theorem describes the main convergence result.

Theorem 20 *Whenever Assumptions 1, 2, 4 (when applicable), 5, 9, and 17—19 hold, then as $k \to \infty$, $d_k \to 0$ and $\theta_k \to \theta^*$ for all $i \in \mathcal{N}$ almost surely by applying the Algorithm.*

Proof: Refer to Appendix D.

The main idea is to write d_k as a function of the previous iterations using the algorithm's descent update and the -xpansive property of the projection. We then replace the gradient estimate with its expectation and the stochastic noise. Making use of the derived properties of the gradient estimate in the previous section, we take the telescoping sum of the divergence. We thus employ Doob martingale's inequality to prove that the term comprising the stochastic error is bounded. Another term is bounded by the assumption we impose on the step sizes. The final term incorporates a product of the optimality gap between the model and the optimal solution by the exact gradient at that model. We use the strict convexity property to prove that this term is negative. We incur that the only two options for d_k are either 0 or $-\infty$, but since d_k is positive by definition, it must converge to 0. This theorem is valid for both noise-free and noisy environments, as shown in the proof.

4.6 Convergence Rate with a Strongly Convex Objective Function

We now consider the following additional assumption.

Assumption 21 *Let the objective function be μ-strongly convex, i.e.,*

$$\langle \nabla F(\theta), \theta - \theta^* \rangle \geq \mu \|\theta - \theta^*\|^2, \quad \forall \theta \in \mathbb{R}^d. \tag{21}$$

We define the expected divergence

$$D_k = \mathbb{E}[\|\theta_k - \theta^*\|^2]. \tag{22}$$

The following theorems identify the convergence rate of the algorithm for both the noise-free and noisy settings, respectively.

Theorem 22 *(In the noise-free setting) Let $\alpha_k = \gamma_k = \sqrt{\frac{\nu/c_1}{l+k}}$, $\forall k \geq 0$ for some $\nu, l > 0$. Then, if in addition to the conditions in Theorem 20, F satisfies Assumption 21 and the constant values satisfy $\nu\mu > 1$ and $l > \frac{\nu}{2}(\mu + L)$, then*

$$D_k \leq \frac{D}{l+k} \quad \textit{with} \quad D \geq \frac{c\nu^2}{\nu\mu - 1}. \tag{23}$$

Proof: Refer to Appendix D.1.1.

The main idea is again to write D_{k+1} in terms of D_k using the algorithm's decent step and the expansive property of the projection. Then, taking the expectation (by the definition of D_k), we substitute by the conditional expectation of g_k derived in Lemma 6 and the upper bound of its norm in Lemma 7. After some necessary technical steps, we derive the left-hand side of the following inequality (24) in the upper bound of D_{k+1}. The strong convexity and L-smoothness properties allows us to write (Qu and Li, 2018, Lemma 10)

$$\|\theta_k - c_1\alpha_k\gamma_k\nabla F(\theta_k) - \theta^*\|^2 \leq \lambda_k^2\|\theta_k - \theta^*\|^2. \tag{24}$$

with $\lambda_k = 1 - c_1\alpha_k\gamma_k\mu$. Again, after some technical manipulations, D_{k+1} becomes bounded from above by D_k multiplied by λ_k and other terms which are function of the step sizes once we substitute by the upper bound of the bias in Lemma 8. Considering the form of the step sizes in Theorem 22, we are able to prove (24) by hypothesis testing.

Theorem 22 implies that the algorithm evolves as $O(\frac{1}{k})$ which is an important rate for ZO optimization as it competes with FO rates. In addition, the bound in (23) evolves as $O(\frac{d^3N^3}{\mu^2})$ and as $O(\frac{1}{L})$. Similarly to the κ-gradient dominated non-convex case with vanishing step sizes, the additional $O(dN)$ dependence comes from the conditions on α_0 and γ_0.

Theorem 23 *(In the noisy setting) Let $\alpha_k = \alpha_0(1+k)^{-\frac{3}{4}}$ and $\gamma_k = \gamma_0(1+k)^{-\frac{1}{4}}$, $\forall k \geq 0$ for some $\alpha_0, \gamma_0 > 0$. Define the iteration K_0 such that*

$$K_0 = \underset{\alpha_k\gamma_k < \min\{\frac{2}{c_1(\mu+L)}, \frac{1}{2c_1\mu}\}}{\arg\min} \quad k$$

Then, if in addition to the conditions in Theorem 20, F satisfies Assumption 21 and the constant values satisfy $\alpha_0\gamma_0 \geq \frac{1}{4c_1\mu}$, then

$$D_k \leq \frac{D'}{\sqrt{1+k}}, \quad \forall k \geq K_0, \tag{25}$$

with D' some bounded constant.

Proof: Refer to Appendix D.1.2.

We start with similar steps to prove this theorem to those of the previous theorem. However, a term containing $\sqrt{D_k}$ appears in the upper bound of D_{k+1}. This complicates the subsequent steps. The following hypothesis comprises testing if the rate can be bounded from above by a decreasing sequence U_k, i.e., $D_k \leq U_k$. Based on the conditions obtained for U_k, we derive candidates for this sequence form. The candidates have the forms $D_k \leq \zeta_1^2\gamma_k^2$ and $D_k \leq \zeta_2^2\frac{\gamma_k}{\alpha_k}$. After long technical measures, we prove that ζ_1 and ζ_2 are bounded with conditioning on α_0 and γ_0 and the exponents υ_1 and υ_2 in Example 2. We then provide a detailed analysis to optimize these exponents and find the optimal rate for $\upsilon_1 = \frac{3}{4}$ and $\upsilon_2 = \frac{1}{4}$.

The bound (25) also evolves as $O(\frac{d^3N^3}{\mu^2})$, $O(\frac{1}{L})$, and $O(\sigma_n^4)$.

4.7 Convergence Rate with a Strongly Convex Objective Function and Fixed Step Sizes

For this subsection, we consider the same assumptions on the objective function as in the previous one. The following theorem describes the convergence rate.

Theorem 24 *Let the assumptions of Theorem 20 alongside Assumption 21 hold. Fix $\alpha_k = \alpha > 0$ and $\gamma_k = \gamma > 0$ for all $k \geq 0$. Then, for $\lambda = 1 - c_1\alpha\gamma\mu$ and $\alpha\gamma < \frac{2}{c_1(\mu+L)}$ (resulting in $c_1\alpha\gamma\mu < 1$ as $\mu \leq L$),*

- *In a noise-free environment,*

$$D_{K+1} \leq \lambda^{K+1}D_0 + \alpha\gamma\bigg(c_1c_3^2\gamma^2\bigg(\frac{1}{\mu} + c_1\alpha\gamma\bigg) + c_2\alpha\gamma\bigg)\frac{1-\lambda^{K+1}}{1-\lambda}. \tag{26}$$

- *In a noisy environment,*

$$D_{K+1} \leq \lambda^{K+1}D_0 + \alpha\bigg(c_1c_3'^2\gamma^3\bigg(\frac{1}{\mu} + c_1\alpha\gamma\bigg) + c_2'\alpha\bigg)\frac{1-\lambda^{K+1}}{1-\lambda}. \tag{27}$$

Proof: Refer to Appendices D.2.1 and D.2.2.

Knowing that $\lambda < 1$, then for an arbitrarily small value of the step sizes α and γ, we can say that the algorithm converges to a neighborhood of the optimum with a linear rate $O(\lambda^K)$. The bound (26) evolves as $O(\frac{dN}{(\mu+L)^2})$, while (27) evolves as $O(\frac{dN}{\mu+L})$ and as $O(\sigma_n^4)$.

5 Simulation Results

We conduct our experiments using university servers equipped with 32 CPUs and 80GB memory, managed by the Slurm workload manager. The simulations ran in a Conda virtual environment, utilizing PyTorch (Version 2.0.0) as the primary library and Torchvision for data set access. The resources were allocated from the cpu_long partition. We compare our algorithm to the original FL algorithm, FedAvg (McMahan et al., 2017), with exact gradient and one local update per round. It is important to note that we did not consider the impact of the channel or any noise/stochasticity in the FedAvg algorithm. In Figure 2, we also compare with FedZO (Fang et al., 2022) with $H = 50$ local iterations and $M = 10$ participating users and $b_1 = b_2 = 1$ (two-point gradient estimate) for fairness, where we assume that the uplink transmission is subject to the same channel/noise as our algorithm and the signals are decoded (impact of the channel is removed) using the transceiver designed by Fang et al. (2022). Each experiment involved data batches consisting of 10 images per user per round and all graphs are averaged over 30 simulations with different random model initializations testing the accuracy in every iteration against an independent test set. Each communication round depicted in the graphs encompasses all steps from 2 to 7 of the algorithm.

5.1 Non-Convex Objective Function

In the first example, we conduct image classification on "shirts" and "sneakers" from the FashionMNIST data set (Xiao et al., 2017) using a multilayer perceptron. The model has an input layer with 784 units and two hidden layers, each with 200 units and ReLU activations. The final layer employs a sigmoid activation, resulting in a total of $197,602$ parameters. In this experiment, we test our algorithm for indenpendent and identically distributed (IID) data among users and non-IID data. For the non-IID data distribution, we arrange the images based on their labels and then distribute them among 100 devices. Similar to the approach by McMahan et al. (2017), each curve in the plot represents the best test-set accuracy achieved over all previous rounds. The corresponding results are depicted in Figure 2.

Although the impact of noise on both the theoretical and experimental convergence rates is evident, our algorithm consistently performs well across various random variations in each simulation. Introducing a non-IID data distribution appears to have a marginal impact, slightly slowing down our algorithm without significantly affecting the final outcome.

The key observation in this experiment is that for convergence, FedAvg necessitates 300 communication rounds and FedZO necessitates 750, while 2P-ZOFL requires 2000. However, by the time 300 rounds are completed, each device in FedAvg will have uploaded a total of $197602 \times 300 = 59280600$ scalar values/symbols to the server and each device in FedZO will have uploaded $197602 \times 750 = 148201500$, compared to only $2000 \times 2 = 4000$ for 2P-ZOFL. 4000 symbols sent by one device during all 2P-ZOFL's iterations are still less than that sent by one device in the standard method per iteration. This means FedAvg will have 14820 times more data points per user (FedZO will have 37050 more per user). Thus, to have a similar efficiency to 2P-ZOFL, the FL method must compress its data with a ratio of 14820 (99.99% resource saving) with the same guarantees of convergence during 300 iterations; otherwise, this ratio increases (with quantization/compression, usually the

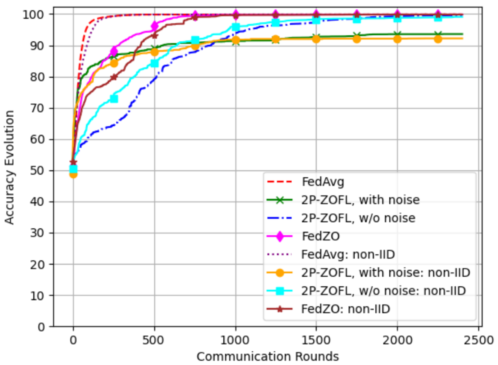

Figure 2: Accuracy evolution of Fashion-MNIST images classification via training using 2P-ZOFL with and without noise as compared with FedAvg and FedZO for IID and non-IID data distributions.

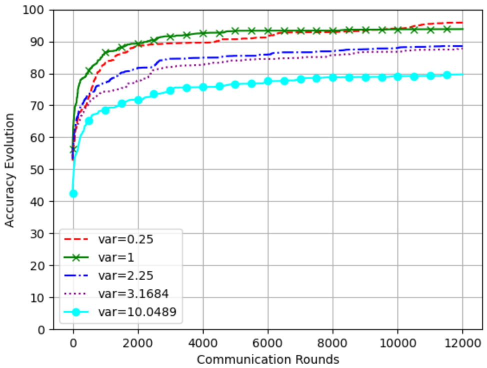

Figure 3: Accuracy evolution of MNIST images classification via non-convex logistic regression using 2P-ZOFL for different noise variance $\sigma_n^2 = \{0.25, 1, 2.25, 3.1684, 10.0489\}$.

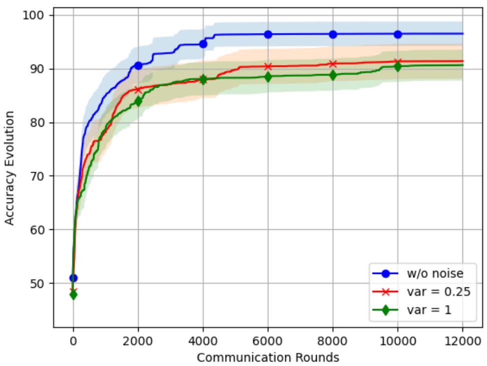

Figure 4: Accuracy evolution of MNIST images classification via non-convex logistic regression with 95% confidence intervals.

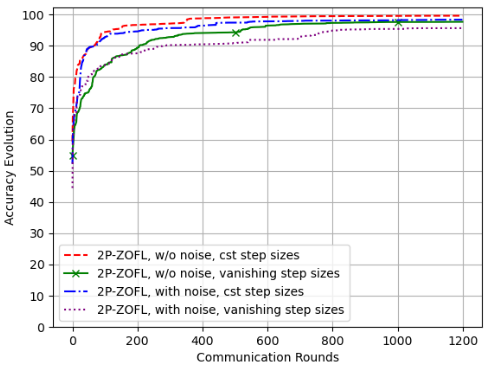

Figure 5: Accuracy evolution of MNIST images classification via gradient dominated non-convex function.

convergence becomes slower). The same concept applies to other efficiency strategies, e.g., with partial device participation. We also note that these numbers do not include the continuous channel knowledge information acquisition that is necessary for standard methods (e.g., FedZO) but not for 2P-ZOFL. We further remark that 2P-ZOFL's local computation takes even less time and consumes fewer resources (e.g., battery) as there's no "backward propagation" step, only querying of the model. In Appendix E.1, we provide a quantitative visualization of savings for 2P-ZOFL vs FedAvg and FedZO.

For the second example in Figure 3, we classify the images of the handwritten digits "0" and "1" taken from the MNIST data set (LeCun and Cortes, 2005) using the non-convex sigmoid function. All images are evenly distributed among 100 devices and undergo preprocessing, where they are compressed using a lossy autoencoder to reduce their dimensionality to $d = 10$. We then study in depth the effect of the noise on the performance of 2P-ZOFL. While the convergence remains intact, we notice a decrease in convergence speed with the increase of noise variance σ_n^2. This is natural as bigger noise variance results in bigger gradient estimation variance. However, 2P-ZOFL still performs well despite the degraded wireless conditions. In Figure 4, we plot the same accuracy evolution with 95% confidence interval, where at worst, there seems to be a loss in 2.5% accuracy.

5.2 κ-Gradient Dominated Non-Convex Objective Function

In Figure 5, we use the same classification example as in Figure 3, but employing the loss function $f_i(\theta; x_i, y_i) = (l(\theta^T x_i) - y_i)^2$, where l is the logistic sigmoid function $l(a) = (1 + \exp(-a))^{-1}$. It was shown that the corresponding objective function is gradient-dominated (Foster et al., 2018). With constant step sizes, the linear rate is evident in the faster convergence of accuracy.

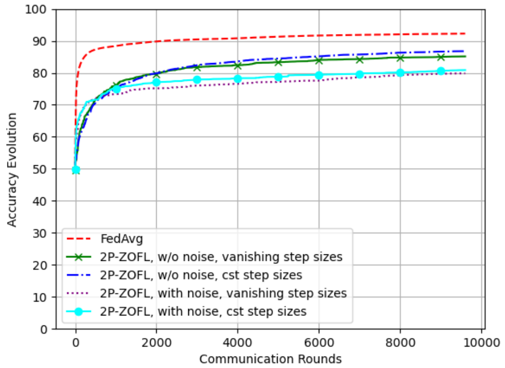

Figure 6: Accuracy evolution of mushroom classification using 2P-ZOFL with and without noise, with vanishing and constant step sizes, as compared with FedAvg.

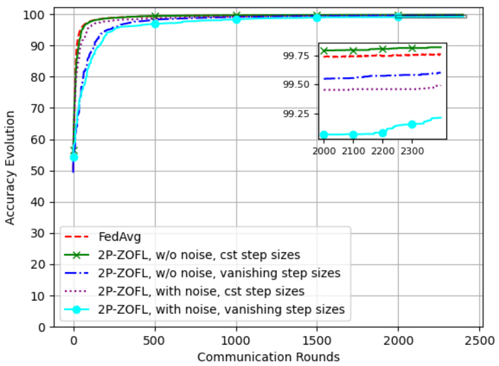

Figure 7: Accuracy evolution of MNIST images classification using 2P-ZOFL with and without noise, with vanishing and constant step sizes, as compared with FedAvg.

5.3 Strictly Convex Objective Function

For this subsection, we utilize the convex logistic regression model with regularization for binary classification with different data sets. In what follows, we set the regularization constant to 0.001 and we project the optimization variable into the set $[-10, 10]^d$.

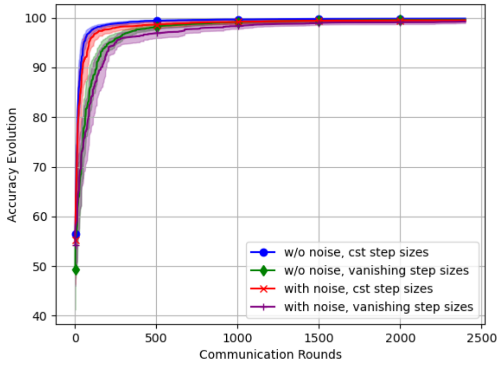

Figure 8: Accuracy evolution of MNIST images classification using 2P-ZOFL with and without noise, with vanishing and constant step sizes, with 95% confidence intervals.

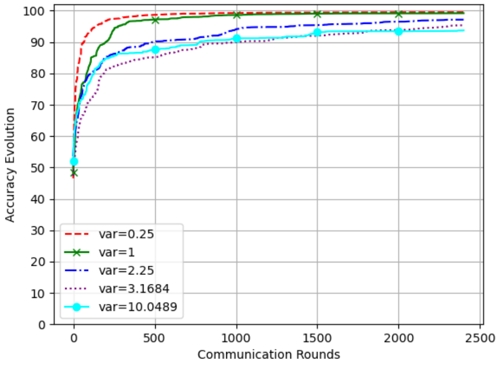

Figure 9: Accuracy evolution of MNIST images classification using 2P-ZOFL for different noise variance $\sigma_n^2 = \{0.25, 1, 2.25, 3.1684, 10.0489\}$ and constant step sizes.

In the third scenario, we classify data from the Mushroom data set (mus, 1987) as edible or not and we divide the data points equally among 20 users. For this example in Figure 6, we compare the effect of noise and the step sizes on 2P-ZOFL and we compare it with FedAvg. While the algorithm performs relatively well in all cases, we notice a slight gap in convergence speed due to the presence of noise. This is in line with our theoretical findings as the noise increases the upper bound for the estimation variance.

We also notice the increase of speed due to fixed step sizes irrelevant of the presence of noise.

In the final example of Figure 7, we use the same classication as in the second example of subsection 5.1, by replacing the non-convex sigmoid function by the regularized convex logistic loss. In Figure 8, we plot the 95% confidence intervals where the fluctuations seem negligible across the different experimental instances, especially near the end at convergence. We also notice the same effect of adding noise variance in Figure 9 to the slowing down of convergence as in the non-convex case.

For additional experimental details and parameter choices, please refer to Appendix E.

6 Conclusion

This study addresses a learning challenge in the context of wireless channels and introduces a novel two-point gradient estimator-based zero-order federated learning approach. Our method restricts communication to scalar-valued feedback from devices and integrates the wireless channel directly into the learning algorithm. We support our approach with both theoretical analyses and experimental validation, establishing convergence and deriving an upper bound on the convergence rate under various settings and conditions imposed on the objective function.

Acknowledgments and Disclosure of Funding

We sincerely thank the editor and the reviewers for their diligent work and insightful comments while overviewing the review of our paper.

Appendix A. Zero-Order Gradient Estimate

A.1 Biased Estimator

A.1.1 Without Noise

$$
\begin{aligned}
\mathbb{E}[g_k|\mathcal{H}_k] =& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[f_i\Big(\theta_k+\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a,\xi_{k+1}\Big) \\
&\qquad\qquad - f_i\Big(\theta_k-\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a,\xi_{k+1}\Big)\Big]\Big|\mathcal{H}_k\Big] \\
\overset{(a)}{=}& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[F_i\Big(\theta_k+\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a\Big)-F_i\Big(\theta_k-\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N} h_{j,k}a\Big)\Big]\Big|\mathcal{H}_k\Big] \\
\overset{(b)}{=}& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[F_i(\theta_k)+\gamma_k\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a\omega_k^T\nabla F_i(\theta_k)+\gamma_k^2(\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a)^2\omega_k^T\nabla^2 F_i(\acute{\theta}_k)\omega_k \\
&\quad -\Big(F_i(\theta_k)-\gamma_k\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a\omega_k^T\nabla F_i(\theta_k)+\gamma_k^2(\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a)^2\omega_k^T\nabla^2 F_i(\grave{\theta}_k)\omega_k\Big)\Big]\Big|\mathcal{H}_k\Big] \\
=& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big(2\gamma_k\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a\omega_k^T\nabla F_i(\theta_k) \\
&\qquad\qquad +\gamma_k^2(\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a)^2\omega_k^T(\nabla^2 F_i(\acute{\theta}_k)-\nabla^2 F_i(\grave{\theta}_k))\omega_k\Big)\Big|\mathcal{H}_k\Big] \\
\overset{(c)}{=}& \mathbb{E}\Big[\frac{\omega_k}{N^2}\Big(2a\gamma_k\sum_{i=1}^{N} h_{i,k+1}h_{i,k}\omega_k^T\nabla F_i(\theta_k)\Big|\mathcal{H}_k\Big] \\
&+\mathbb{E}\Big[\omega_k\Big(a^2\gamma_k^2\frac{1}{N^3}\sum_{i=1}^{N} h_{i,k+1}(\sum_{j=1}^{N} h_{j,k})^2\omega_k^T(\nabla^2 F_i(\acute{\theta}_k)-\nabla^2 F_i(\grave{\theta}_k))\omega_k\Big)\Big|\mathcal{H}_k\Big] \\
=& 2a\gamma_k\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}h_{i,k}\Big|\mathcal{H}_k\Big]\mathbb{E}\Big[\omega_k\omega_k^T\Big|\mathcal{H}_k\Big]\nabla F_i(\theta_k) \\
&+a^2\gamma_k^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}(\sum_{j=1}^{N} h_{j,k})^2\omega_k\omega_k^T(\nabla^2 F_i(\acute{\theta}_k)-\nabla^2 F_i(\grave{\theta}_k))\omega_k\Big|\mathcal{H}_k\Big] \\
\overset{(d)}{=}& 2a\beta_2 K_{hh}\gamma_k\frac{1}{N^2}\sum_{i=1}^{N}\nabla F_i(\theta_k) \\
&+a^2\gamma_k^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}(\sum_{j=1}^{N} h_{j,k})^2\omega_k\omega_k^T(\nabla^2 F_i(\acute{\theta}_k)-\nabla^2 F_i(\grave{\theta}_k))\omega_k\Big|\mathcal{H}_k\Big] \\
=& c_1\gamma_k(\nabla F(\theta_k)+b_k)
\end{aligned}
\tag{28}
$$

where (a) is by the definition in (3), (b) is by Taylor expansion and mean-valued theorem and considering $\acute{\theta}_k$ between θ_k and $\theta_k + \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k}a$, and $\grave{\theta}_k$ between θ_k and $\theta_k - \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k}a$. (c) is since $\mathbb{E}[h_{i,k+1}h_{j,k}] = 0$ when $i \neq j$ in the first term. (d) is due to Assumption 5. In (e), we let $c_1 = \frac{2a\beta_2 K_{hh}}{N}$.

A.1.2 With Noise

$$
\begin{aligned}
\mathbb{E}[g_k|\mathcal{H}_k] =& \mathbb{E}\Big[\frac{\omega_k}{N}\Big(\sum_{i=1}^{N} h_{i,k+1}\Big[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big) \\
&\qquad\qquad - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big)\Big] + n_{i,k+1}\Big)\Big|\mathcal{H}_k\Big] \\
\overset{(a)}{=}& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[F_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\Big) \\
&\qquad\qquad - F_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\Big)\Big]\Big|\mathcal{H}_k\Big] \\
\overset{(b)}{=}& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[F_i(\theta_k) + \gamma_k \frac{1}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\omega_k^T \nabla F_i(\theta_k) \\
&+ \gamma_k^2 (\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a + n_{j,k})^2 \omega_k^T \nabla^2 F_i(\acute{\theta}_k)\omega_k - F_i(\theta_k) + \gamma_k \frac{1}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\omega_k^T \nabla F_i(\theta_k) \\
&- \gamma_k^2 (\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a + n_{j,k})^2 \omega_k^T \nabla^2 F_i(\grave{\theta}_k)\omega_k\Big]\Big|\mathcal{H}_k\Big] \\
=& \mathbb{E}\Big[\frac{\omega_k}{N}\sum_{i=1}^{N} h_{i,k+1}\Big(2\gamma_k \frac{1}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\omega_k^T \nabla F_i(\theta_k) \\
&\qquad\qquad + \gamma_k^2 (\frac{1}{N}\sum_{j=1}^{N} h_{j,k}a + n_{j,k})^2 \omega_k^T (\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k))\omega_k\Big)\Big|\mathcal{H}_k\Big] \\
\overset{(c)}{=}& \mathbb{E}\Big[2a\gamma_k \frac{1}{N^2}\sum_{i=1}^{N} h_{i,k+1}h_{i,k}\omega_k\omega_k^T \nabla F_i(\theta_k)\Big|\mathcal{H}_k\Big] \\
&+ \mathbb{E}\Big[\omega_k\Big(\gamma_k^2 \frac{1}{N^3}\sum_{i=1}^{N} h_{i,k+1}(\sum_{j=1}^{N} h_{j,k}a + n_{j,k})^2 \omega_k^T (\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k))\omega_k\Big)\Big|\mathcal{H}_k\Big] \\
=& 2a\gamma_k \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{E}\Big[h_{i,k+1}h_{i,k}\Big|\mathcal{H}_k\Big]\mathbb{E}\Big[\omega_k\omega_k^T\Big|\mathcal{H}_k\Big]\nabla F_i(\theta_k) \\
&+ \gamma_k^2 \frac{1}{N^3}\sum_{i=1}^{N} \mathbb{E}\Big[h_{i,k+1}(\sum_{j=1}^{N} h_{j,k}a + n_{j,k})^2 \omega_k\omega_k^T (\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k))\omega_k\Big|\mathcal{H}_k\Big]
\end{aligned}
$$

$$\begin{aligned}&\overset{(d)}{=}2a\beta_2K_{hh}\gamma_k\frac{1}{N^2}\sum_{i=1}^{N}\nabla F_i(\theta_k)\\&\quad+\gamma_k^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}(\sum_{j=1}^{N}h_{j,k}a+n_{j,k})^2\omega_k\omega_k^T(\nabla^2F_i(\acute{\theta}_k)-\nabla^2F_i(\grave{\theta}_k))\omega_k\Big|\mathcal{H}_k\Big]\\&=c_1\gamma_k(\nabla F(\theta_k)+b'_k)\end{aligned}\tag{29}$$

where (a) is by the definition in (3) and the zero-mean noise in Assumption 4, (b) is by Taylor expansion and mean-valued theorem and considering $\acute{\theta}_k$ between θ_k and $\theta_k+\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a+n_{j,k})$, and $\grave{\theta}_k$ between θ_k and $\theta_k-\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a+n_{j,k})$. (c) is since $\mathbb{E}[h_{i,k+1}h_{j,k}]=0$ when $i\neq j$ in the first term. (d) is due to Assumption 5. In (e), we let $c_1=\frac{2a\beta_2K_{hh}}{N}$.

A.2 Bounding the Bias

A.2.1 Without Noise

From (28), we can see that the estimate bias has the form

$$b_k=\gamma_k\frac{a}{2\beta_2K_{hh}N^2}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}(\sum_{j=1}^{N}h_{j,k})^2\omega_k\omega_k^T(\nabla^2F_i(\acute{\theta}_k)-\nabla^2F_i(\grave{\theta}_k))\omega_k\Big|\mathcal{H}_k\Big].$$

This bias can be bounded from above using (4) and Assumption 5 as

$$\begin{aligned}\|b_k\|&\overset{(a)}{\leq}\gamma_k\frac{a}{2\beta_2K_{hh}N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\Big|h_{i,k+1}N(\sum_{j=1}^{N}h_{j,k}^2)\Big|\|\omega_k\|\|\omega_k^T\|\|\nabla^2F_i(\acute{\theta}_k)-\nabla^2F_i(\grave{\theta}_k)\|\|\omega_k\|\Big|\mathcal{H}_k\Big]\\&\leq 2\gamma_k\beta_1\beta_3^3\frac{a}{2\beta_2K_{hh}N}\sum_{i=1}^{N}\mathbb{E}\Big[\big|h_{i,k+1}(h_{i,k}^2+\sum_{j\neq i}h_{j,k}^2)\big|\Big|\mathcal{H}_k\Big]\\&\overset{(b)}{\leq}\gamma_k\beta_1\beta_3^3\frac{a}{\beta_2K_{hh}N}\sum_{i=1}^{N}\Big(\sqrt{\frac{2}{\pi}}\sigma_h\Big(2K_{hh}+\sqrt{\sigma_h^4-K_{hh}^2}\Big)+(N-1)\sigma_h^3\sqrt{\frac{2}{\pi}}\Big)\\&=\gamma_k\frac{a\beta_1\beta_3^3\sigma_h}{\beta_2K_{hh}}\sqrt{\frac{2}{\pi}}\Big(2K_{hh}+\sqrt{\sigma_h^4-K_{hh}^2}+(N-1)\sigma_h^2\Big)\\&:=c_3\gamma_k\end{aligned}$$

where (a) is due to Jensen's inequality and Cauchy-Schwarz, (b) is by using the half-normal distribution for normal random variables in absolute value explained in the following paragraph, and in (c), $c_3=\frac{a\beta_1\beta_3^3\sigma_h}{\beta_2K_{hh}}\sqrt{\frac{2}{\pi}}\Big(2K_{hh}+\sqrt{\sigma_h^4-K_{hh}^2}+(N-1)\sigma_h^2\Big)$.

Let X and Y be two random variables following the $\mathcal{N}(0,\sigma^2)$ distribution with correlation coefficient ϱ. Then, we can write $Y=\varrho X+\sqrt{1-\varrho^2}Z$, where Z is independent of X and following the same distribution $\mathcal{N}(0,\sigma^2)$. Then, $\mathbb{E}[|YX^2|]=\mathbb{E}[|(\varrho X+\sqrt{1-\varrho^2}Z)X^2|]=\mathbb{E}[|\varrho X^3+\sqrt{1-\varrho^2}ZX^2|]\leq\mathbb{E}[\varrho|X^3|+\sqrt{1-\varrho^2}|ZX^2|]=2\varrho\sqrt{\frac{2}{\pi}}\sigma^3+\sqrt{1-\varrho^2}\sqrt{\frac{2}{\pi}}\sigma\times\sigma^2=(2\varrho+\sqrt{1-\varrho^2})\sqrt{\frac{2}{\pi}}\sigma^3$. If we substitute $\sigma=\sigma_h$ and $\varrho=\frac{K_{hh}}{\sigma_h^2}$, we obtain the inequality above in (b).

A.2.2 With Noise

From (29), we can see that the estimate bias has the form

$$b_k' = \gamma_k \frac{1}{2a\beta_2 K_{hh} N^2} \sum_{i=1}^{N} \mathbb{E}\Big[h_{i,k+1} (\sum_{j=1}^{N} h_{j,k} a + n_{j,k})^2 \omega_k \omega_k^T (\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k)) \omega_k \Big| \mathcal{H}_k \Big]. \tag{30}$$

by Assumptions 1, 4, and 5,

$$\begin{aligned}
\|b_k'\| \overset{(a)}{\leq} & \gamma_k \frac{1}{2a\beta_2 K_{hh} N^2} \sum_{i=1}^{N} \mathbb{E}\Big[2N \Big| h_{i,k+1} \sum_{j=1}^{N} \Big(h_{j,k}^2 a^2 + n_{j,k}^2 \Big) \Big| \times \\
& \qquad \|\omega_k\| \|\omega_k^T\| \|\nabla^2 F_i(\acute{\theta}_k) - \nabla^2 F_i(\grave{\theta}_k)\| \|\omega_k\| \Big| \mathcal{H}_k \Big] \\
\leq & \gamma_k \frac{2\beta_1 \beta_3^3}{a\beta_2 K_{hh} N} \sum_{i=1}^{N} \mathbb{E}\Big[\Big| h_{i,k+1} \sum_{j=1}^{N} \Big(h_{j,k}^2 a^2 + n_{j,k}^2 \Big) \Big| \Big| \mathcal{H}_k \Big] \\
\overset{(b)}{\leq} & \gamma_k \frac{2\beta_1 \beta_3^3}{a\beta_2 K_{hh} N} \sum_{i=1}^{N} \Big[a^2 \sigma_h \sqrt{\frac{2}{\pi}} \Big(2K_{hh} + \sqrt{\sigma_h^4 - K_{hh}^2} \Big) + a^2 (N-1) \sigma_h^3 \sqrt{\frac{2}{\pi}} + N \sqrt{\frac{2}{\pi}} \sigma_h \sigma_n^2 \Big] \\
= & \gamma_k \frac{2a\beta_1 \beta_3^3 \sigma_h}{\beta_2 K_{hh}} \sqrt{\frac{2}{\pi}} \Big[2K_{hh} + \sqrt{\sigma_h^4 - K_{hh}^2} + (N-1)\sigma_h^2 + N \frac{\sigma_n^2}{a^2} \Big] \\
\overset{(c)}{=} & c_3' \gamma_k,
\end{aligned} \tag{31}$$

where (a) is due to Jensen's inequality and Cauchy-Schwarz, (b) is by using the half-normal distribution for normal random variables in absolute value explained at the end of subsection A.2.1, and in (c), $c_3' = \frac{2a\beta_1 \beta_3^3 \sigma_h}{\beta_2 K_{hh}} \sqrt{\frac{2}{\pi}} \Big[2K_{hh} + \sqrt{\sigma_h^4 - K_{hh}^2} + (N-1)\sigma_h^2 + N \frac{\sigma_n^2}{a^2} \Big]$.

A.3 Expected Norm Squared of the Gradient Estimate

A.3.1 Without Noise

Bounding the norm squared of the gradient estimate

$$\begin{aligned}
\mathbb{E}[\|g_k\|^2 | \mathcal{H}_k] = & \mathbb{E}\Big[\Big\| \frac{\omega_k}{N} \sum_{i=1}^{N} h_{i,k+1} \Big[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k} a, \xi_{k+1} \Big) \\
& \qquad - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k} a, \xi_{k+1} \Big) \Big] \Big\|^2 \Big| \mathcal{H}_k \Big] \\
= & \mathbb{E}\Big[\|\omega_k\|^2 \Big(\frac{1}{N} \sum_{i=1}^{N} h_{i,k+1} \Big[f_i\Big(\theta_k + \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k} a, \xi_{k+1} \Big) \\
& \qquad - f_i\Big(\theta_k - \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} h_{j,k} a, \xi_{k+1} \Big) \Big] \Big)^2 \Big| \mathcal{H}_k \Big]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq}\beta_3^2\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^{N}h_{i,k+1}\Big[f_i\Big(\theta_k+\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}h_{j,k}a,\xi_{k+1}\Big)\\
&\qquad\qquad\qquad\qquad-f_i\Big(\theta_k-\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}h_{j,k}a,\xi_{k+1}\Big)\Big]\Big)^2\Big|\mathcal{H}_k\Big]\\
&\overset{(b)}{\leq}\beta_3^2\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^{N}h_{i,k+1}L_{\xi_{k+1}}\Big\|2\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}h_{j,k}a\Big\|\Big)^2\Big|\mathcal{H}_k\Big]\\
&\overset{(c)}{\leq}\beta_3^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(h_{i,k+1}L_{\xi_{k+1}}\Big\|2\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}h_{j,k}a\Big\|\Big)^2\Big|\mathcal{H}_k\Big]\\
&\overset{(d)}{=}4\gamma_k^2L_\xi^2\beta_3^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[\|\omega_k\|^2h_{i,k+1}^2(\sum_{j=1}^{N}h_{j,k}a)^2\Big|\mathcal{H}_k\Big]\\
&\leq 4a^2\gamma_k^2L_\xi^2\beta_3^4\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}^2(\sum_{j=1}^{N}h_{j,k}^2+2\sum_{j<l}h_{j,k}h_{l,k})\Big|\mathcal{H}_k\Big]\\
&\overset{(e)}{=}4a^2\gamma_k^2L_\xi^2\beta_3^4\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}^2(h_{i,k}^2+\sum_{j\neq i}h_{j,k}^2)\Big|\mathcal{H}_k\Big]\\
&\overset{(f)}{=}4a^2\gamma_k^2L_\xi^2\beta_3^4\frac{1}{N^3}\sum_{i=1}^{N}\Big[\sigma_h^4+2K_{hh}^2+(N-1)\sigma_h^4\Big]\\
&=4a^2\gamma_k^2L_\xi^2\beta_3^4\frac{1}{N^2}(2K_{hh}^2+N\sigma_h^4)\\
&\overset{(g)}{=}c_2\gamma_k^2,
\end{aligned}
\tag{32}
$$

where (a) is by Assumption 5, (b) is by Assumption 2, and (c) is by Cauchy-Schwartz, $(\sum_{i=1}^{N}x_i)^2=(\sum_{i=1}^{N}1\cdot x_i)^2\leq N\sum_{i=1}^{N}x_i^2$. In (d), we let $L_\xi^2=\mathbb{E}[L_{\xi_{k+1}}^2|\mathcal{H}_k]$, in (e), the last term has a zero mean since one element of the zero-mean channels will always be independent of the others, and (f) is due to the *normally-distributed* channel random variables. In (g), $c_2=\frac{4a^2\gamma_k^2L_\xi^2\beta_3^4}{N^2}(2K_{hh}^2+N\sigma_h^4)$.

A.3.2 With Noise

$$
\begin{aligned}
\mathbb{E}[\|g_k\|^2|\mathcal{H}_k]=\mathbb{E}\Big[\Big\|\frac{\omega_k}{N}\Big(\sum_{i=1}^{N}h_{i,k+1}\Big[f_i\Big(\theta_k+\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a+n_{j,k}),\xi_{k+1}\Big)\\
-f_i\Big(\theta_k-\gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a+n_{j,k}),\xi_{k+1}\Big)\Big]+n_{i,k+1}\Big)\Big\|^2\Big|\mathcal{H}_k\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\|\omega_k\|^2\Big(\frac{1}{N}\sum_{i=1}^{N} h_{i,k+1}\Big[f_i\Big(\theta_k + \gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big) \\
&\qquad\qquad - f_i\Big(\theta_k - \gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big)\Big] + n_{i,k+1}\Big)^2\Big|\mathcal{H}_k\Big] \\
&\overset{(a)}{\leq} \beta_3^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(h_{i,k+1}\Big[f_i\Big(\theta_k + \gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big) \\
&\qquad\qquad - f_i\Big(\theta_k - \gamma_k\frac{\omega_k}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k}), \xi_{k+1}\Big)\Big] + n_{i,k+1}\Big)^2\Big|\mathcal{H}_k\Big] \\
&\overset{(b)}{\leq} \beta_3^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}^2 L_{\xi_{k+1}}^2\|2\gamma_k\omega_k\frac{1}{N}\sum_{j=1}^{N}(h_{j,k}a + n_{j,k})\|^2 + n_{i,k+1}^2\Big|\mathcal{H}_k\Big] \qquad (33) \\
&\overset{(c)}{\leq} 4\gamma_k^2\beta_3^4L_\xi^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}^2(\sum_{j=1}^{N}h_{j,k}a + n_{j,k})^2\Big|\mathcal{H}_k\Big] + \beta_3^2\sigma_n^2 \\
&= 4\gamma_k^2\beta_3^4L_\xi^2\frac{1}{N^3}\sum_{i=1}^{N}\mathbb{E}\Big[h_{i,k+1}^2(\sum_{j=1}^{N}(h_{j,k}^2a^2 + n_{j,k}^2) + \sum_{j<l}h_{j,k}h_{l,k}a^2)\Big|\mathcal{H}_k\Big] + \beta_3^2\sigma_n^2 \\
&\overset{(d)}{=} 4a^2\gamma_k^2\beta_3^4L_\xi^2\frac{1}{N^3}\Big[N\Big(\sigma_h^4 + 2K_{hh}^2 + (N-1)\sigma_h^4 + N\frac{\sigma_n^2\sigma_h^2}{a^2}\Big)\Big] + \beta_3^2\sigma_n^2 \\
&= 4a^2\gamma_k^2\beta_3^4L_\xi^2\frac{1}{N^2}\Big(2K_{hh}^2 + N\sigma_h^4 + N\frac{\sigma_n^2\sigma_h^2}{a^2}\Big) + \beta_3^2\sigma_n^2 \\
&\overset{(e)}{:=} c_2',
\end{aligned}
$$

where (a) is by Assumption 5 and Cauchy-Schwartz,i.e., $(\sum_{i=1}^N x_i)^2 = (\sum_{i=1}^N 1 \cdot x_i)^2 \leq N\sum_{i=1}^N x_i^2$. (b) is by (5) and the independence of the noise in Assumption 4. In (c), we let $L_\xi^2 = \mathbb{E}[L_{\xi_{k+1}}^2|\mathcal{H}_k]$, and (d) is due to the *normally-distributed* channel random variables. In (e), $c_2' = \frac{4a^2\gamma_k^2\beta_3^4L_\xi^2}{N^2}\Big(2K_{hh}^2 + N\sigma_h^4 + N\frac{\sigma_n^2\sigma_h^2}{a^2}\Big) + \beta_3^2\sigma_n^2$.

Appendix B. Stochastic Noise

To prove Lemma 10, we begin by demonstrating that the sequence $\{\sum_{k=K}^{K'}\alpha_k e_k\}_{K'\geq K}$ is a martingale. To do so, we have to prove that for all $K' \geq K$, $X_{K'} = \sum_{k=K}^{K'}\alpha_k e_k$ satisfies the following two conditions:

(i) $\mathbb{E}[X_{K'+1}|X_{K'}] = X_{K'}$

(ii) $\mathbb{E}[\|X_{K'}\|^2] < \infty$

We know that $\mathbb{E}[e_k] = \mathbb{E}[g_k - \mathbb{E}[g_k|\mathcal{H}_k]] = \mathbb{E}_{\mathcal{H}_k}\Big[\mathbb{E}\Big[g_k - \mathbb{E}[g_k|\mathcal{H}_k]\Big|\mathcal{H}_k\Big]\Big] = 0$ by the law of total expectation. Hence, $\mathbb{E}[X_{K'+1}|X_{K'}] = \mathbb{E}\Big[\alpha_{K'+1}e_{K'+1} + \sum_{k=K}^{K'}\alpha_k e_k\Big|\sum_{k=K}^{K'}\alpha_k e_k\Big] = 0 + \sum_{k=K}^{K'}\alpha_k e_k = X_{K'}$.

In addition, e_k and $e_{k'}$ are uncorrelated for any $k \neq k'$ since (assuming $k > k'$) $\mathbb{E}\big[e_k^T e_{k'}\big] = \mathbb{E}\big[\mathbb{E}[e_k^T e_{k'}|\mathcal{H}_k]\big] = \mathbb{E}\big[e_{k'}\mathbb{E}[e_k^T|\mathcal{H}_k]\big] = 0$. Thus,

$$\begin{aligned}\mathbb{E}(\|\sum_{k=K}^{K'}\alpha_k e_k\|^2) =& \mathbb{E}(\sum_{k=K}^{K'}\sum_{k'=K}^{K'}\alpha_k\alpha_{k'}\langle e_k, e_{k'}\rangle) \\ \overset{(a)}{=}& \mathbb{E}(\sum_{k=K}^{K'}\|\alpha_k e_k\|^2) \\ \leq& \sum_{k=K}^{\infty}\mathbb{E}(\alpha_k^2\|g_k - \mathbb{E}[g_k|\mathcal{H}_k]\|^2) \\ =& \sum_{k=K}^{\infty}\alpha_k^2\mathbb{E}(\|g_k\|^2) - \mathbb{E}_{\mathcal{H}_k}(\|\mathbb{E}[g_k|\mathcal{H}_k]\|^2) \\ \leq& \sum_{k=K}^{\infty}\alpha_k^2\mathbb{E}(\|g_k\|^2) \\ \overset{(b)}{\leq}& c_2\sum_{k=K}^{\infty}\alpha_k^2\gamma_k^2 \quad \Big(\leq c_2'\sum_{k=K}^{\infty}\alpha_k^2 \quad \text{in case of noise}\Big) \\ \overset{(c)}{<}& \infty,\end{aligned} \tag{34}$$

where (a) is due to the uncorrelatedness $\mathbb{E}[\langle e_k, e_{k'}\rangle] = 0$, (b) is by Lemma 7, and (c) is by Assumption 9. Therefore, both (i) and (ii) are satisfied and we can say that $\{\sum_{k=K}^{K'}\alpha_k e_k\}_{K'\geq K}$ is a martingale. This permits us to use Doob's martingale inequality (Doob, 1953):

For any constant $\nu > 0$,

$$\begin{aligned}\mathbb{P}(\sup_{K'\geq K}\|\sum_{k=K}^{K'}\alpha_k e_k\| \geq \nu) \leq& \frac{1}{\nu^2}\mathbb{E}(\|\sum_{k=K}^{K'}\alpha_k e_k\|^2) \\ \overset{(a)}{\leq}& \frac{c_2}{\nu^2}\sum_{k=K}^{\infty}\alpha_k^2\gamma_k^2 \quad \Big(\leq \frac{c_2'}{\nu^2}\sum_{k=K}^{\infty}\alpha_k^2 \quad \text{in case of noise}\Big),\end{aligned} \tag{35}$$

where (a) is following the exact same steps as (34).

Since c_2 (c_2' in case of noise) is a bounded constant and $\lim_{K\to\infty}\sum_{k=K}^{\infty}\alpha_k^2\gamma_k^2 = 0$ $\big(\lim_{K\to\infty}\sum_{k=K}^{\infty}\alpha_k^2 = 0$ with noise$\big)$ by Assumption 9, we get $\lim_{K\to\infty}\frac{c_2}{\nu^2}\sum_{k=K}^{\infty}\alpha_k^2\gamma_k^2 = 0$ $\big(\lim_{K\to\infty}\frac{c_2}{\nu^2}\sum_{k=K}^{\infty}\alpha_k^2 = 0$ with noise$\big)$ for any bounded constant ν. Hence, the probability that $\|\sum_{k=K}^{K'}\alpha_k e_k\| \geq \nu$ also vanishes as $K \to \infty$, which concludes the proof.

Appendix C. Convergence with Non-Convex Objective Function

By the L-smoothness assumption, we have

$$\begin{aligned}
F(\theta_{k+1}) \leq & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), g_k \rangle + \frac{\alpha_k^2 L}{2} \|g_k\|^2 \\
= & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), g_k - \mathbb{E}[g_k|\mathcal{H}_k] + \mathbb{E}[g_k|\mathcal{H}_k] \rangle + \frac{\alpha_k^2 L}{2} \|g_k\|^2 \\
= & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), e_k \rangle - c_1 \alpha_k \gamma_k \langle \nabla F(\theta_k), \nabla F(\theta_k) + b_k \rangle + \frac{\alpha_k^2 L}{2} \|g_k\|^2 \\
= & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), e_k \rangle - c_1 \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 - c_1 \alpha_k \gamma_k \langle \nabla F(\theta_k), b_k \rangle + \frac{\alpha_k^2 L}{2} \|g_k\|^2 \\
\overset{(a)}{\leq} & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), e_k \rangle - c_1 \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|b_k\|^2 \\
& + \frac{\alpha_k^2 L}{2} \|g_k\|^2 \\
= & F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), e_k \rangle - \frac{c_1 \alpha_k \gamma_k}{2} \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|b_k\|^2 + \frac{\alpha_k^2 L}{2} \|g_k\|^2
\end{aligned} \tag{36}$$

where (a) is by $-\langle a, b \rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. Note that b_k is replaced by b'_k for the case of noisy channels.

By taking the telescoping sum, we get

$$\begin{aligned}
F(\theta^*) \leq F(\theta_{K+1}) \leq & F(\theta_0) - \frac{c_1}{2} \sum_{k=0}^{K} \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 - \sum_{k=0}^{K} \alpha_k \langle \nabla F(\theta_k), e_k \rangle + \frac{c_1}{2} \sum_{k=0}^{K} \alpha_k \gamma_k \|b_k\|^2 \\
& + \frac{L}{2} \sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2
\end{aligned} \tag{37}$$

Hence,

$$\begin{aligned}
\sum_{k=0}^{K} \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 \leq & \frac{2}{c_1} (F(\theta_0) - F(\theta^*)) - \frac{2}{c_1} \sum_{k=0}^{K} \alpha_k \langle \nabla F(\theta_k), e_k \rangle + \sum_{k=0}^{K} \alpha_k \gamma_k \|b_k\|^2 \\
& + \frac{L}{c_1} \sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2
\end{aligned} \tag{38}$$

By Lemma 3, $\|\nabla F(\theta_k)\|$ is bounded for any $\theta_k \in \mathbb{R}^d$ by taking the summation in (35) between 0 and ∞, we have

$$\lim_{K \to \infty} \| \sum_{k=0}^{K} \alpha_k \langle \nabla F(\theta_k), e_k \rangle \| < \infty. \tag{39}$$

From Lemma 8, we know that $\|b_k\|^2 \sim \gamma_k^2$ (similarly, $\|b_k'\|^2 \sim \gamma_k^2$). Hence, by Assumption 11,

$$\lim_{K\to\infty} \sum_{k=0}^{K} \alpha_k \gamma_k^3 < \infty. \tag{40}$$

To prove the finiteness of $\sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2$, we let X_k be a centered Gaussian process. We know that for any σ^2-subgaussian random variables $X_1, \dots, X_K$, we have

$$\mathbb{P}\left[\sup_{1\le k\le K} X_k \ge \sqrt{2\sigma^2(\log K + t)}\right] \le e^{-t}$$

since let $u := \sqrt{2\sigma^2(\log K + t)}$,

$$\mathbb{P}\left[\sup_{1\le k\le K} X_k \ge u\right] = \mathbb{P}\left[\exists k, X_k \ge u\right] \le \sum_{k=1}^{K} \mathbb{P}\left[X_k \ge u\right] \le K e^{-\frac{u^2}{2\sigma^2}} = e^{-t}.$$

Then, for $t = c\log K$ with $c > 1$, $\mathbb{P}\left[\sup_{1\le k\le K} X_k \ge \sqrt{2\sigma^2(1+c)\log K}\right] \le \frac{1}{K^c}$ and

$$\sum_{K=1}^{\infty} \mathbb{P}\left[\sup_{1\le k\le K} X_k \ge \sqrt{2\sigma^2(1+c)\log K}\right] \le \sum_{K=1}^{\infty} \frac{1}{K^c} < \infty.$$

By Borel-Cantelli Lemma, we have

$$\mathbb{P}\left(\lim_{K\to\infty} \sup\{\sup_{1\le k\le K} X_k \ge \sqrt{2\sigma^2(1+c)\log K}\}\right) = 0.$$

Then, w.p.1 $\exists K' < \infty$ s.t. $\forall K \ge K'$, $\sup_{1\le k\le K} X_k \le \sqrt{2\sigma^2(1+c)\log K}$ with $c > 1$.

Thus, the supremum of the Gaussian elements of $\|g_k\|$, i.e., $h_{i,k+1}, h_{i,k}$, $n_{i,k}$, and $n_{i,k+1}$, at worst grow as $O(\sqrt{log(k+1)})$, and hence the upper bound on $\|g_k\|$ grows as $c_4\gamma_k \log(k+1)$ (and as $c_4' \log(k+1)$ with noise), where $c_4 = 4\beta_3^2 L_{\xi_{k+1}}(1+c)|a|\sigma_h^2$ and $c_4' = 4\gamma_k\beta_3^2 L_{\xi_{k+1}}(1+c)(\sigma_h^2|a| + \sigma_n\sigma_h) + \beta_3\sqrt{2\sigma_n^2(1+c)}$:

$$\begin{aligned}
\|g_k\| &= \Big\| \frac{\omega_k}{N} \Big(\sum_{i=1}^{N} h_{i,k+1} \big[f_i\big(\theta_k + \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} (h_{j,k}a + n_{j,k}), \xi_{k+1}\big) \\
&\qquad\qquad - f_i\big(\theta_k - \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} (h_{j,k}a + n_{j,k}), \xi_{k+1}\big)\big] + n_{i,k+1}\Big) \Big\| \\
&\leq \|\omega_k\| \frac{1}{N} \sum_{i=1}^{N} \Big\| h_{i,k+1} 2L_{\xi_{k+1}} \Big\| \gamma_k \frac{\omega_k}{N} \sum_{j=1}^{N} (h_{j,k}a + n_{j,k}) \Big\| \Big\| + \|n_{i,k+1}\| \\
&\leq \|\omega_k\| \Big[\|\omega_k\| \frac{2\gamma_k L_{\xi_{k+1}}}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |h_{i,k+1}| \cdot (|a h_{j,k}| + |n_{j,k}|) + \frac{1}{N} \sum_{i=1}^{N} |n_{i,k+1}| \Big] \\
&\leq \beta_3^2 \frac{2\gamma_k L_{\xi_{k+1}}}{N^2} N^2 \sqrt{2\sigma_h^2 (1+c) \log(k+1)} (|a| \sqrt{2\sigma_h^2 (1+c) \log(k)} + \sqrt{2\sigma_n^2 (1+c) \log(k)}) \\
&+ \frac{\beta_3}{N} N \sqrt{2\sigma_n^2 (1+c) \log(k+1)} \\
&\leq 4\gamma_k \beta_3^2 L_{\xi_{k+1}} (1+c) \sqrt{\sigma_h^2 \log(k+1)} (|a| \sqrt{\sigma_h^2 \log(k+1)} + \sqrt{\sigma_n^2 \log(k+1)}) \\
&+ \beta_3 \sqrt{2\sigma_n^2 (1+c) \log(k+1)} \\
&= 4\gamma_k \beta_3^2 L_{\xi_{k+1}} (1+c) (\sigma_h^2 |a| + \sigma_n \sigma_h) \log(k+1) + \beta_3 \sqrt{2\sigma_n^2 (1+c) \log(k+1)} \\
&\leq \Big(4\gamma_k \beta_3^2 L_{\xi_{k+1}} (1+c) (\sigma_h^2 |a| + \sigma_n \sigma_h) + \beta_3 \sqrt{2\sigma_n^2 (1+c)} \Big) \log(k+1),
\end{aligned}$$

where the last inequality holds for $k \geq e - 1$. Then, we write $\sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2 = \sum_{k=0}^{K'} \alpha_k^2 \|g_k\|^2 + \sum_{k=K'}^{K} \alpha_k^2 \|g_k\|^2$, where $\sum_{k=0}^{K'} \alpha_k^2 \|g_k\|^2 < \infty$ for $K' < \infty$.

We know that $\forall \epsilon > 0$, $\log(k+1) \leq \frac{1}{\epsilon}(k+1)^{\epsilon}$. Thus, by Assumption 9, for $\epsilon' > 0$ and $2(\upsilon_1 + \upsilon_2) = 1 + \epsilon'$ ($2\upsilon_1 = 1 + \epsilon'$ with noise),

$$\begin{aligned}
\lim_{K\to\infty} \sum_{k=K'}^{K} \alpha_k^2 \|g_k\|^2 &\leq \lim_{K\to\infty} c_4^2 \sum_{k=K'}^{K} \alpha_k^2 \gamma_k^2 \log^2(k+1) \\
&\leq \lim_{K\to\infty} c_4^2 \sum_{k=K'}^{K} \frac{1}{(k+1)^{1+\epsilon'}} \times \frac{1}{\epsilon^2} (k+1)^{2\epsilon} < \infty, \ \ \forall \epsilon' > 2\epsilon > 0. \\
\Bigg(\lim_{K\to\infty} \sum_{k=K'}^{K} \alpha_k^2 \|g_k\|^2 &\leq \lim_{K\to\infty} c_4'^2 \sum_{k=K'}^{K} \alpha_k^2 \log^2(k+1) \\
&\leq \lim_{K\to\infty} c_4'^2 \sum_{k=K'}^{K} \frac{1}{(k+1)^{1+\epsilon'}} \times \frac{1}{\epsilon^2} (k+1)^{2\epsilon} < \infty, \ \ \forall \epsilon' > 2\epsilon > 0 \text{ with noise} \Bigg).
\end{aligned} \tag{41}$$

We conclude that

$$\lim_{K\to\infty} \sum_{k=0}^{K} \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 < \infty. \tag{42}$$

Moreover, since the series $\sum_k \alpha_k \gamma_k$ diverges by Assumption 9, we have

$$\lim_{k\to\infty} \inf \|\nabla F(\theta_k)\| = 0. \tag{43}$$

To prove that $\lim_{k\to\infty} \|\nabla F(\theta_k)\| = 0$, we consider the hypothesis:

(H) $\lim_{k\to\infty} \sup \|\nabla F(\theta_k)\| \geq \rho$ for an arbitrary $\rho > 0$.

Assume (H) to be true. Then, we can always find an arbitrary subsequence $\left(\|\nabla F(\theta_{k_l})\|\right)_{l\in\mathbb{N}}$ of $\|\nabla F(\theta_k)\|$, such that $\|\nabla F(\theta_{k_l})\| \geq \rho - \varepsilon$, $\forall l$, for $\rho - \varepsilon > 0$ and $\varepsilon > 0$.

Then, by the L-smoothness property and applying the descent step of the algorithm,

$$\begin{aligned}
\|\nabla F(\theta_{k_l+1})\| \geq& \|\nabla F(\theta_{k_l})\| - \|\nabla F(\theta_{k_l+1}) - \nabla F(\theta_{k_l})\| \\
\geq& \rho - \varepsilon - L\|\theta_{k_l+1} - \theta_{k_l}\| \\
=& \rho - \varepsilon - L\alpha_{k_l}\|g_{k_l}\| \\
\geq& \rho - \varepsilon - L\sqrt{c}\alpha_{k_l}\gamma_{k_l} \quad \left(\geq \rho - \varepsilon - L\sqrt{c'}\alpha_{k_l} \text{ with noise}\right),
\end{aligned} \tag{44}$$

Since $k_l \to \infty$ as $l \to \infty$, we can always find a subsequence of $(k_{l_p})_{p\in\mathbb{N}}$ such that $k_{l_{p+1}} - k_{l_p} > 1$. As $\alpha_{k_l}\gamma_{k_l}$ is vanishing, we consider $(k_l)_{l\in\mathbb{N}}$ starting from $\alpha_{k_l}\gamma_{k_l} < \frac{\rho-\varepsilon}{L\sqrt{c}}$. Thus,

(without noise):

$$\begin{aligned}
&\sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\|\nabla F(\theta_{k+1})\|^2 \\
&\geq (\rho-\varepsilon)^2 \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1} - 2(\rho-\varepsilon)L\sqrt{c}\sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k\gamma_k + L^2 c \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k^2\gamma_k^2 \\
&\geq (\rho-\varepsilon)^2 \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1} - 2(\rho-\varepsilon)L\sqrt{c}\sum_{k=0}^{\infty} \alpha_k^2\gamma_k^2 + L^2 c \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k^2\gamma_k^2 \\
&= +\infty,
\end{aligned} \tag{45}$$

(with noise):

$$\begin{aligned}
&\sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\|\nabla F(\theta_{k+1})\|^2 \\
&\geq (\rho-\varepsilon)^2 \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1} - 2(\rho-\varepsilon)L\sqrt{c'}\sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k + L^2 c' \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k^2 \\
&\geq (\rho-\varepsilon)^2 \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1} - 2(\rho-\varepsilon)L\sqrt{c'}\sum_{k=0}^{\infty} \alpha_k^2 + L^2 c' \sum_{k=0}^{\infty} \alpha_{k+1}\gamma_{k+1}\alpha_k^2 \\
&= +\infty,
\end{aligned} \tag{46}$$

as the first series diverges, and the second and the third converge by Assumption 9. This implies that the series $\sum_k \alpha_k\gamma_k\|\nabla F(\theta_k)\|^2$ diverges. This is a contradiction as this series converges almost surely by (42). Therefore, hypothesis (H) cannot be true and $\|\nabla F(\theta_k)\|$ converges to zero almost surely.

C.1 Non-Convex Objective Function Convergence Rate

C.1.1 Without Noise

Considering again the L-smoothness inequality, we have

$$F(\theta_{k+1}) \leq F(\theta_k) - \alpha_k \langle \nabla F(\theta_k), g_k \rangle + \frac{\alpha_k^2 L}{2} \|g_k\|^2. \tag{47}$$

Taking the conditional expectation given $\mathcal{H}_k$,

$$\begin{aligned}
\mathbb{E}[F(\theta_{k+1})|\mathcal{H}_k] \leq & F(\theta_k) - c_1 \alpha_k \gamma_k \langle \nabla F(\theta_k), \nabla F(\theta_k) + b_k \rangle + \frac{c_2 L}{2} \alpha_k^2 \gamma_k^2 \\
= & F(\theta_k) - c_1 \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 - c_1 \alpha_k \gamma_k \langle \nabla F(\theta_k), b_k \rangle + \frac{c_2 L}{2} \alpha_k^2 \gamma_k^2 \\
\overset{(a)}{\leq} & F(\theta_k) - c_1 \alpha_k \gamma_k \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|b_k\|^2 + \frac{c_2 L}{2} \alpha_k^2 \gamma_k^2 \\
= & F(\theta_k) - \frac{c_1 \alpha_k \gamma_k}{2} \|\nabla F(\theta_k)\|^2 + \frac{c_1 \alpha_k \gamma_k}{2} \|b_k\|^2 + \frac{c_2 L}{2} \alpha_k^2 \gamma_k^2
\end{aligned} \tag{48}$$

where (a) is by $-\langle a, b \rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$

Taking the telescoping sum of (48),

$$\begin{aligned}
\mathbb{E}[F(\theta_{K+1})] &\leq F(\theta_0) - \frac{c_1}{2} \sum_k \alpha_k \gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2] + \frac{c_1}{2} \sum_k \alpha_k \gamma_k \mathbb{E}[\|b_k\|^2] + \frac{c_2 L}{2} \sum_k \alpha_k^2 \gamma_k^2 \\
0 \leq \mathbb{E}[\delta_{K+1}] &\leq \delta_0 - \frac{c_1}{2} \sum_k \alpha_k \gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2] + \frac{c_1}{2} \sum_k \alpha_k \gamma_k \mathbb{E}[\|b_k\|^2] + \frac{c_2 L}{2} \sum_k \alpha_k^2 \gamma_k^2.
\end{aligned} \tag{49}$$

Hence,

$$\begin{aligned}
\sum_k \alpha_k \gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2] \leq & \frac{2}{c_1} \delta_0 + \sum_k \alpha_k \gamma_k \mathbb{E}[\|b_k\|^2] + \frac{c_2 L}{c_1} \sum_k \alpha_k^2 \gamma_k^2 \\
\leq & \frac{2}{c_1} \delta_0 + c_3^2 \sum_k \alpha_k \gamma_k^3 + \frac{c_2 L}{c_1} \sum_k \alpha_k^2 \gamma_k^2
\end{aligned} \tag{50}$$

Let $\alpha_k = \alpha_0 (1+k)^{-\upsilon_1}$ and $\gamma_k = \gamma_0 (1+k)^{-\upsilon_2}$. Then, to satisfy Assumptions 9 and 11, it is sufficient to find υ_1 and υ_2 such that $0 < \upsilon_1 + \upsilon_2 \leq 1$, $\upsilon_1 + 3\upsilon_2 > 1$, and $\upsilon_1 + \upsilon_2 > 0.5$.

We know that, $\forall K > 0$,

$$\begin{aligned}\sum_{k=0}^{K} \alpha_k \gamma_k^3 &= \alpha_0\gamma_0^3 + \sum_{k=1}^{K} \alpha_k \gamma_k^3 \\ &\leq \alpha_0\gamma_0^3 \bigg(1 + \int_0^K (x+1)^{-\upsilon_1 - 3\upsilon_2} dx\bigg) \\ &= \alpha_0\gamma_0^3 \bigg(1 + \frac{1}{\upsilon_1 + 3\upsilon_2 - 1} - \frac{(K+1)^{-\upsilon_1 - 3\upsilon_2 + 1}}{\upsilon_1 + 3\upsilon_2 - 1}\bigg) \\ &\leq \alpha_0\gamma_0^3 \bigg(1 + \frac{1}{\upsilon_1 + 3\upsilon_2 - 1}\bigg) \\ &= \alpha_0\gamma_0^3 \bigg(\frac{\upsilon_1 + 3\upsilon_2}{\upsilon_1 + 3\upsilon_2 - 1}\bigg).\end{aligned} \tag{51}$$

Similarly,

$$\sum_{k=0}^{K} \alpha_k^2 \gamma_k^2 \leq \alpha_0^2 \gamma_0^2 \bigg(\frac{2\upsilon_1 + 2\upsilon_2}{2\upsilon_1 + 2\upsilon_2 - 1}\bigg) \tag{52}$$

- Next, when $0 < \upsilon_1 + \upsilon_2 < 1$,

$$\begin{aligned}\sum_{k=0}^{K} \alpha_k \gamma_k &\geq \alpha_0\gamma_0 \int_0^{K+1} (x+1)^{-\upsilon_1 - \upsilon_2} dx \\ &= \frac{\alpha_0\gamma_0}{(1 - \upsilon_1 - \upsilon_2)} \bigg((K+2)^{1-\upsilon_1-\upsilon_2} - 1\bigg).\end{aligned} \tag{53}$$

Thus, making use of inequality (50)

$$\begin{aligned}\frac{\sum_k \alpha_k \gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2]}{\sum_k \alpha_k \gamma_k} \leq & \frac{(1 - \upsilon_1 - \upsilon_2)}{(K+2)^{1-\upsilon_1-\upsilon_2} - 1} \times \\ & \bigg[\frac{2\delta_0}{c_1 \alpha_0 \gamma_0} + \frac{(\upsilon_1 + 3\upsilon_2)(c_3\gamma_0)^2}{\upsilon_1 + 3\upsilon_2 - 1} + \frac{2(\upsilon_1 + \upsilon_2) c_2 \alpha_0 \gamma_0 L}{c_1(2\upsilon_1 + 2\upsilon_2 - 1)}\bigg]\end{aligned} \tag{54}$$

In the pursuit of optimizing the time-varying component, which follows the scaling of $O\left(\frac{1}{K^{1-\upsilon_1-\upsilon_2}}\right)$, we find that the most suitable values for the exponents are $\upsilon_1 = \upsilon_2 = \frac{1}{4}$, resulting in a rate of $O\left(\frac{1}{\sqrt{K}}\right)$. However, it is worth noting that with this specific selection, the constant portion becomes excessively large, underscoring the need for a compromise.

- Otherwise, when $\upsilon_1 + \upsilon_2 = 1$,

$$\begin{aligned}\sum_{k=0}^{K} \alpha_k \gamma_k &\geq \alpha_0\gamma_0 \int_0^{K+1} \frac{1}{x+1} dx \\ &= \alpha_0\gamma_0 \ln(K+2).\end{aligned} \tag{55}$$

Thus, we get

$$\begin{aligned}\frac{\sum_k \alpha_k\gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2]}{\sum_k \alpha_k\gamma_k} \leq & \frac{1}{\ln(K+2)}\times \\ & \left[\frac{2\delta_0}{c_1\alpha_0\gamma_0} + \frac{(\upsilon_1+3\upsilon_2)(c_3\gamma_0)^2}{\upsilon_1+3\upsilon_2-1} + \frac{2(\upsilon_1+\upsilon_2)c_2\alpha_0\gamma_0 L}{c_1(2\upsilon_1+2\upsilon_2-1)}\right].\end{aligned} \tag{56}$$

C.1.2 With Noise

By the L-smoothness inequality,

$$F(\theta_{k+1}) \leq F(\theta_k) - \alpha_k\langle\nabla F(\theta_k), g_k\rangle + \frac{\alpha_k^2 L}{2}\|g_k\|^2. \tag{57}$$

Taking the conditional expectation given $\mathcal{H}_k$,

$$\begin{aligned}\mathbb{E}[F(\theta_{k+1})|\mathcal{H}_k] \leq & F(\theta_k) - c_1\alpha_k\gamma_k\langle\nabla F(\theta_k), \nabla F(\theta_k)+b_k'\rangle + \frac{c_2'L}{2}\alpha_k^2 \\ = & F(\theta_k) - c_1\alpha_k\gamma_k\|\nabla F(\theta_k)\|^2 - c_1\alpha_k\gamma_k\langle\nabla F(\theta_k), b_k'\rangle + \frac{c_2'L}{2}\alpha_k^2 \\ \overset{(a)}{\leq} & F(\theta_k) - c_1\alpha_k\gamma_k\|\nabla F(\theta_k)\|^2 + \frac{c_1\alpha_k\gamma_k}{2}\|\nabla F(\theta_k)\|^2 + \frac{c_1\alpha_k\gamma_k}{2}\|b_k'\|^2 + \frac{c_2'L}{2}\alpha_k^2 \\ = & F(\theta_k) - \frac{c_1\alpha_k\gamma_k}{2}\|\nabla F(\theta_k)\|^2 + \frac{c_1\alpha_k\gamma_k}{2}\|b_k'\|^2 + \frac{c_2'L}{2}\alpha_k^2\end{aligned} \tag{58}$$

where (a) is by $-\langle a,b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$

Taking the telescoping sum of (58),

$$\begin{aligned}\mathbb{E}[F(\theta_{K+1})] \leq F(\theta_0) - \frac{c_1}{2}\sum_k \alpha_k\gamma_k\mathbb{E}[\|\nabla F(\theta_k)\|^2] + \frac{c_1}{2}\sum_k \alpha_k\gamma_k\mathbb{E}[\|b_k'\|^2] + \frac{c_2'L}{2}\sum_k \alpha_k^2 \\ 0 \leq \mathbb{E}[\delta_{K+1}] \leq \delta_0 - \frac{c_1}{2}\sum_k \alpha_k\gamma_k\mathbb{E}[\|\nabla F(\theta_k)\|^2] + \frac{c_1}{2}\sum_k \alpha_k\gamma_k\mathbb{E}[\|b_k'\|^2] + \frac{c_2'L}{2}\sum_k \alpha_k^2.\end{aligned} \tag{59}$$

Hence,

$$\begin{aligned}\sum_k \alpha_k\gamma_k\mathbb{E}[\|\nabla F(\theta_k)\|^2] \leq & \frac{2}{c_1}\delta_0 + \sum_k \alpha_k\gamma_k\mathbb{E}[\|b_k'\|^2] + \frac{c_2'L}{c_1}\sum_k \alpha_k^2 \\ \leq & \frac{2}{c_1}\delta_0 + (c_3')^2\sum_k \alpha_k\gamma_k^3 + \frac{c_2'L}{c_1}\sum_k \alpha_k^2.\end{aligned} \tag{60}$$

Let $\alpha_k = \alpha_0(1+k)^{-\upsilon_1}$ and $\gamma_k = \gamma_0(1+k)^{-\upsilon_2}$. Then, to satisfy Assumptions 9 and 11, it is sufficient to find υ_1 and υ_2 such that $0 < \upsilon_1+\upsilon_2 \leq 1$, $\upsilon_1+3\upsilon_2 > 1$, and $\upsilon_1 > 0.5$.

Similarly to (51), $\forall K > 0$, $\sum_{k=0}^K \alpha_k\gamma_k^3 \leq \alpha_0\gamma_0^3\Big(\frac{\upsilon_1+3\upsilon_2}{\upsilon_1+3\upsilon_2-1}\Big)$ and $\sum_{k=0}^K \alpha_k^2 \leq \alpha_0^2\Big(\frac{2\upsilon_1}{2\upsilon_1-1}\Big)$.

- When $0 < \upsilon_1 + \upsilon_2 < 1$, $\sum_{k=0}^{K} \alpha_k \gamma_k \geq \frac{\alpha_0\gamma_0}{(1-\upsilon_1-\upsilon_2)}\Big((K+2)^{1-\upsilon_1-\upsilon_2} - 1\Big)$.

  Thus,

$$\frac{\sum_k \alpha_k\gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2]}{\sum_k \alpha_k\gamma_k} \leq \frac{(1-\upsilon_1-\upsilon_2)}{(K+2)^{1-\upsilon_1-\upsilon_2}-1} \times \left[\frac{2\delta_0}{c_1\alpha_0\gamma_0} + \frac{(\upsilon_1+3\upsilon_2)(c_3'\gamma_0)^2}{\upsilon_1+3\upsilon_2-1} + \frac{2\upsilon_1 c_2'\alpha_0 L}{c_1\gamma_0(2\upsilon_1-1)}\right].$$

  When optimizing for the time-varying component, which scales as $O\left(\frac{1}{K^{1-\upsilon_1-\upsilon_2}}\right)$, we discover that the most favorable values for the exponents are $\upsilon_1 = \frac{1}{2}$ and $\upsilon_2 = \frac{1}{6}$, resulting in a rate of $O\left(\frac{1}{\sqrt[3]{K}}\right)$. However, it is important to note that with this particular choice, the constant part becomes excessively large, indicating the necessity for a trade-off or compromise.

- Otherwise, when $\upsilon_1 + \upsilon_2 = 1$, $\sum_{k=0}^{K} \alpha_k\gamma_k \geq \alpha_0\gamma_0 \ln(K+2)$.

  Thus, we get

$$\frac{\sum_k \alpha_k\gamma_k \mathbb{E}[\|\nabla F(\theta_k)\|^2]}{\sum_k \alpha_k\gamma_k} \leq \frac{1}{\ln(K+2)}\left[\frac{2\delta_0}{c_1\alpha_0\gamma_0} + \frac{(\upsilon_1+3\upsilon_2)(c_3'\gamma_0)^2}{\upsilon_1+3\upsilon_2-1} + \frac{2\upsilon_1 c_2'\alpha_0 L}{c_1\gamma_0(2\upsilon_1-1)}\right]. \tag{61}$$

C.2 κ-Gradient Dominated Non-Convex Objective Function Convergence Rate

C.2.1 Without Noise

Making use of inequalities (48) and (13),

$$\begin{aligned}\mathbb{E}[\delta_{k+1}|\mathcal{H}_k] &\leq \delta_k - \frac{c_1\alpha_k\gamma_k}{2}2\kappa\delta_k + \frac{c_1\alpha_k\gamma_k}{2}\|b_k\|^2 + \frac{\alpha_k^2 L c_2\gamma_k^2}{2} \\ &\leq (1 - c_1\alpha_k\gamma_k\kappa)\delta_k + \frac{c_1c_3^2\alpha_k\gamma_k^3}{2} + \frac{c_2L}{2}\alpha_k^2\gamma_k^2.\end{aligned} \tag{62}$$

By recursion,

$$\begin{aligned}\mathbb{E}[\delta_K] \leq& \prod_{k=0}^{K-1}(1 - c_1\alpha_k\gamma_k\kappa)\delta_0 + \frac{c_1c_3^2}{2}\sum_{k=0}^{K-1}\alpha_k\gamma_k^3\prod_{j=k+1}^{K-1}(1 - c_1\alpha_j\gamma_j\kappa) \\ &+ \frac{c_2L}{2}\sum_{k=0}^{K-1}\alpha_k^2\gamma_k^2\prod_{j=k+1}^{K-1}(1 - c_1\alpha_j\gamma_j\kappa).\end{aligned} \tag{63}$$

Substituting by $\alpha_k = \alpha_0(k+2)^{-\upsilon_1}$ and $\gamma_k = \gamma_0(k+2)^{-\upsilon_2}$ and letting $c_1\alpha_0\gamma_0\kappa < 2$,

$$\begin{aligned}\mathbb{E}[\delta_K] \leq \delta_0\prod_{k=0}^{K-1}\left(1 - \frac{c_1\alpha_0\gamma_0\kappa}{(k+2)^{\upsilon_1+\upsilon_2}}\right) + \frac{c_1c_3^2\alpha_0\gamma_0^3}{2}\sum_{k=0}^{K-1}\frac{1}{(k+2)^{\upsilon_1+3\upsilon_2}}\prod_{j=k+1}^{K-1}\left[1 - \frac{c_1\alpha_0\gamma_0\kappa}{(j+2)^{\upsilon_1+\upsilon_2}}\right] \\ + \frac{c_2\alpha_0^2\gamma_0^2L}{2}\sum_{k=0}^{K-1}\frac{1}{(k+2)^{2\upsilon_1+2\upsilon_2}}\prod_{j=k+1}^{K-1}\left[1 - \frac{c_1\alpha_0\gamma_0\kappa}{(j+2)^{\upsilon_1+\upsilon_2}}\right].\end{aligned} \tag{64}$$

As shown by Pu et al. (2022, Lemma 11), we know that

$$\prod_{k=0}^{K-1}\left(1-\frac{c_1\alpha_0\gamma_0\kappa}{k+2}\right) \leq \frac{2^{c_1\alpha_0\gamma_0\kappa}}{(K+2)^{c_1\alpha_0\gamma_0\kappa}} \tag{65}$$

and

$$\prod_{j=k+1}^{K-1}\left(1-\frac{c_1\alpha_0\gamma_0\kappa}{j+2}\right) \leq \frac{(k+3)^{c_1\alpha_0\gamma_0\kappa}}{(K+2)^{c_1\alpha_0\gamma_0\kappa}}. \tag{66}$$

By allowing $\upsilon_1+\upsilon_2=1$, we get

$$\begin{aligned}\mathbb{E}[\delta_K] \leq &\frac{1}{(K+2)^{c_1\alpha_0\gamma_0\kappa}}\times \\ &\left(2^{c_1\alpha_0\gamma_0\kappa}\delta_0+\frac{c_1c_3^2\alpha_0\gamma_0^3}{2}\sum_{k=0}^{K-1}\frac{(k+3)^{c_1\alpha_0\gamma_0\kappa}}{(k+2)^{\upsilon_1+3\upsilon_2}}+\frac{c_2\alpha_0^2\gamma_0^2L}{2}\sum_{k=0}^{K-1}\frac{(k+3)^{c_1\alpha_0\gamma_0\kappa}}{(k+2)^{2\upsilon_1+2\upsilon_2}}\right).\end{aligned} \tag{67}$$

For δ_K to converge, we need $\upsilon_1+3\upsilon_2-c_1\alpha_0\gamma_0\kappa>1$ and $2\upsilon_1+2\upsilon_2-c_1\alpha_0\gamma_0\kappa>1$. Then, optimizing for the rate, we obtain a rate a little short of $O(\frac{1}{K})$ by considering $c_1\alpha_0\gamma_0\kappa$ as near as possible to 1 and $\upsilon_1=\upsilon_2=\frac{1}{2}$.

C.2.2 With Noise

Following similar steps as in subsection C.2.1, we obtain for $\upsilon_1+\upsilon_2=1$

$$\begin{aligned}\mathbb{E}[\delta_K] \leq &\frac{1}{(K+2)^{c_1\alpha_0\gamma_0\kappa}}\times \\ &\left(2^{c_1\alpha_0\gamma_0\kappa}\delta_0+\frac{c_1c_3'^2\alpha_0\gamma_0^3}{2}\sum_{k=0}^{K-1}\frac{(k+3)^{c_1\alpha_0\gamma_0\kappa}}{(k+2)^{\upsilon_1+3\upsilon_2}}+\frac{c_2'\alpha_0^2L}{2}\sum_{k=0}^{K-1}\frac{(k+3)^{c_1\alpha_0\gamma_0\kappa}}{(k+2)^{2\upsilon_1}}\right).\end{aligned} \tag{68}$$

For δ_K to converge, we should have $\upsilon_1+3\upsilon_2-c_1\alpha_0\gamma_0\kappa>1$ and $2\upsilon_1-c_1\alpha_0\gamma_0\kappa>1$. Then, we optimize for the rate, we find a rate a little short of $O(\frac{1}{\sqrt{K}})$ by considering $c_1\alpha_0\gamma_0\kappa$ as near as possible to $\frac{1}{2}$, $\upsilon_1=\frac{3}{4}$, and $\upsilon_2=\frac{1}{4}$.

C.3 κ-Gradient Dominated Non-Convex Objective Function with Fixed Step Sizes Convergence Rate

C.3.1 Without Noise

Following up from (62), and letting $\varsigma=1-c_1\alpha\gamma\kappa$, $\alpha_k=\alpha$, and $\gamma_k=\gamma$,

$$\begin{aligned}\mathbb{E}[\delta_{k+1}|\mathcal{H}_k] &\leq (1-c_1\alpha\gamma\kappa)\delta_k+\frac{c_1c_3^2}{2}\alpha\gamma^3+\frac{c_2L}{2}\alpha^2\gamma^2 \\ &= \varsigma\delta_k+\frac{c_1c_3^2}{2}\alpha\gamma^3+\frac{c_2L}{2}\alpha^2\gamma^2.\end{aligned} \tag{69}$$

Then, for $\alpha\gamma < \frac{1}{c_1\kappa}$, we take the telescoping sum,

$$\begin{aligned}\mathbb{E}[\delta_{k+1}] &\leq \varsigma^{k+1}\delta_0 + \alpha\gamma\bigg(\frac{c_1c_3^2}{2}\gamma^2 + \frac{c_2L}{2}\alpha\gamma\bigg)\sum_{j=0}^{k}\varsigma^j \\ &= \varsigma^{k+1}\delta_0 + \alpha\gamma\bigg(\frac{c_1c_3^2}{2}\gamma^2 + \frac{c_2L}{2}\alpha\gamma\bigg)\frac{1-\varsigma^{k+1}}{1-\varsigma}.\end{aligned} \tag{70}$$

C.3.2 With Noise

Similarly, following up from (58) and (13), and setting again $\varsigma = 1 - c_1\alpha\gamma\kappa$, $\alpha_k = \alpha$, and $\gamma_k = \gamma$,

$$\begin{aligned}\mathbb{E}[\delta_{k+1}|\mathcal{H}_k] \leq& (1 - c_1\alpha\gamma\kappa)\delta_k + \frac{c_1c_3'^2}{2}\alpha\gamma^3 + \frac{c_2'L}{2}\alpha^2 \\ =&\varsigma\delta_k + \frac{c_1c_3'^2}{2}\alpha\gamma^3 + \frac{c_2'L}{2}\alpha^2.\end{aligned} \tag{71}$$

For $\alpha\gamma < \frac{1}{c_1\kappa}$,

$$\mathbb{E}[\delta_{k+1}] \leq \varsigma^{k+1}\delta_0 + \alpha\bigg(\frac{c_1c_3'^2}{2}\gamma^3 + \frac{c_2'L}{2}\alpha\bigg)\frac{1-\varsigma^{k+1}}{1-\varsigma}. \tag{72}$$

Appendix D. Convergence with a Strictly Convex Objective Function

The goal is to write the divergence in terms of its previous term and to prove that it is finally vanishing. We know that $\theta_{k+1} = \theta_k - \alpha_k g_k$. With this equation, the divergence at time $k+1$ can be written as

$$\begin{aligned}d_{k+1} =& \|\theta_{k+1} - \theta^*\|^2 \\ =& \|\Pi_{\mathcal{K}}(\theta_k - \alpha_k g_k) - \theta^*\|^2 \\ \overset{(a)}{\leq}& \|\theta_k - \alpha_k g_k - \theta^*\|^2 \\ =& \|\theta_k - \theta^*\|^2 - 2\alpha_k\langle\theta_k - \theta^*, g_k - \mathbb{E}[g_k|\mathcal{H}_k] + \mathbb{E}[g_k|\mathcal{H}_k]\rangle + \alpha_k^2\|g_k\|^2 \\ =& \|\theta_k - \theta^*\|^2 - 2\alpha_k\langle\theta_k - \theta^*, \mathbb{E}[g_k|\mathcal{H}_k]\rangle - 2\alpha_k\langle\theta_k - \theta^*, e_k\rangle + \alpha_k^2\|g_k\|^2 \\ \overset{(b)}{=}& d_k - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, \nabla F(\theta_k) + b_k\rangle - 2\alpha_k\langle\theta_k - \theta^*, e_k\rangle + \alpha_k^2\|g_k\|^2 \\ =& d_k - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, \nabla F(\theta_k)\rangle - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, b_k\rangle - 2\alpha_k\langle\theta_k - \theta^*, e_k\rangle + \alpha_k^2\|g_k\|^2 \\ \overset{(c)}{\leq}& d_k - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, \nabla F(\theta_k)\rangle + 2c_1c_3\alpha_k\gamma_k^2\|\theta_k - \theta^*\| - 2\alpha_k\langle\theta_k - \theta^*, e_k\rangle + \alpha_k^2\|g_k\|^2\end{aligned}$$

where (a) is by (19) and the fact that $\theta^* \in \mathcal{K}$ in Assumption 19, (b) is due to Lemma 6, and (c) is by Lemma 8. Note that b_k is replaced by b_k' and c_3 by c_3' for the case of noisy channels.

By recursion, we have

$$\begin{aligned} d_{K+1} \leq & d_0 - 2c_1 \sum_{k=0}^{K} \alpha_k \gamma_k \langle \theta_k - \theta^*, \nabla F(\theta_k) \rangle + 2c_1 c_3 \sum_{k=0}^{K} \alpha_k \gamma_k^2 \|\theta_k - \theta^*\| - 2 \sum_{k=0}^{K} \alpha_k \langle \theta_k - \theta^*, e_k \rangle \\ & + \sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2. \end{aligned} \tag{73}$$

By Lemma 10, we have $\lim_{K\to\infty} \| \sum_{k=0}^{K} \alpha_k e_k \| < \infty$ almost surely. Since $\|\theta_k - \theta^*\| < \infty$ by the compactness of $\mathcal{K}$,

$$\lim_{K\to\infty} \| \sum_{k=0}^{K} \alpha_k \langle \theta_k - \theta^*, e_k \rangle \| < \infty. \tag{74}$$

Following the exact same steps in the analysis leading to (41),

$$\lim_{K\to\infty} \sum_{k=0}^{K} \alpha_k^2 \|g_k\|^2 < \infty. \tag{75}$$

By Assumption 17 and the compactness of $\mathcal{K}$, we have

$$\lim_{K\to\infty} \sum_{k=0}^{K} \alpha_k \gamma_k^2 \|\theta_k - \theta^*\| < \infty. \tag{76}$$

From the inequalities (73)-(76), we conclude that there exists D such that $d_{K+1} \leq D + z_K$, with

$$z_K = -2c_1 \sum_{k=0}^{K} \gamma_k \alpha_k \langle \theta_k - \theta^*, \nabla F(\theta_k) \rangle. \tag{77}$$

We know that $z_K < 0$ due to the strict convexity of F in (18).

Consequently, for any big K, $0 \leq d_{K+1} < \infty$ and the limit $\lim_{K\to\infty} d_{K+1} = \bar{d}$ exists.

Thus, there are 2 cases: $\bar{d} > 0$ or $\bar{d} = 0$. Assume hypothesis *H1)* $\bar{d} > 0$ to be valid, i.e., θ_k does not converge to θ^*, then $\forall \epsilon_h > 0$, $\exists K_h$ such that

$$-\langle \theta_k - \theta^*, \nabla F(\theta_k) \rangle < -\epsilon_h, \ \forall k \geq K_h,$$

implying

$$\lim_{K\to\infty} - \sum_{k=K_h}^{K} \gamma_k \alpha_k \langle \theta_k - \theta^*, \nabla F(\theta_k) \rangle < -\epsilon_h \lim_{K\to\infty} \sum_{k=K_m}^{K} \gamma_k \alpha_k < -\infty$$

since $\sum \gamma_k \alpha_k$ diverges by Assumption 9. As a result, we get $\lim_{K\to\infty} z_K < -\infty$ and $\lim_{K\to\infty} d_{K+1} < -\infty$. However, by definition in (20), $d_{K+1} > 0$. Accordingly, the hypothesis *H1* cannot be true and the case $\bar{d} = 0$ is the valid one. We conclude that $\lim_{k\to\infty} d_k = 0$, $\lim_{k\to\infty} \nabla \mathcal{F}(\theta_k) = 0$, and $\lim_{k\to\infty} \theta_k = \theta^*$ almost surely.

D.1 Strongly Convex Objective Function Convergence Rate

D.1.1 Without Noise

In this subsection, we study the convergence rate. Let the objective function F be μ-strongly convex and $\lambda_k = 1 - c_1\alpha_k\gamma_k\mu$, then assuming $c_1\alpha_k\gamma_k \leq \frac{2}{\mu+L}$, for all $k \geq 0$,

$$
\begin{aligned}
&\mathbb{E}\big[\|\theta_{k+1} - \theta^*\|^2\big|\mathcal{H}_k\big]\\
=&\mathbb{E}\big[\|\Pi_{\mathcal{K}}(\theta_k - \alpha_k g_k) - \theta^*\|^2\big|\mathcal{H}_k\big]\\
\overset{(a)}{\leq}&\mathbb{E}\big[\|\theta_k - \alpha_k g_k - \theta^*\|^2\big|\mathcal{H}_k\big]\\
=&\|\theta_k - \theta^*\|^2 - 2\alpha_k\mathbb{E}\big[\langle\theta_k - \theta^*, g_k\rangle\big|\mathcal{H}_k\big] + \alpha_k^2\mathbb{E}\big[\|g_k\|^2\big|\mathcal{H}_k\big]\\
\overset{(b)}{\leq}&\|\theta_k - \theta^*\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, \nabla F(\theta_k) + b_k\rangle + c_2\alpha_k^2\gamma_k^2\\
=&\|\theta_k - \theta^*\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, \nabla F(\theta_k)\rangle + (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2\\
&- 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, b_k\rangle + c_2\alpha_k^2\gamma_k^2\\
=&\|\theta_k - c_1\alpha_k\gamma_k\nabla F(\theta_k) - \theta^*\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k - \theta^*, b_k\rangle + c_2\alpha_k^2\gamma_k^2\\
\overset{(c)}{\leq}&\lambda_k^2\|\theta_k - \theta^*\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k - c_1\alpha_k\gamma_k\nabla F(\theta_k) - \theta^*, b_k\rangle\\
&- 2(c_1\alpha_k\gamma_k)^2\langle\nabla F(\theta_k), b_k\rangle + c_2\alpha_k^2\gamma_k^2\\
\overset{(d)}{\leq}&\lambda_k^2\|\theta_k - \theta^*\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 + c_1\alpha_k\gamma_k\mu\lambda_k^2\|\theta_k - \theta^*\|^2 + \frac{c_1\alpha_k\gamma_k}{\mu}\|b_k\|^2\\
&+ (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 + (c_1\alpha_k\gamma_k)^2\|b_k\|^2 + c_2\alpha_k^2\gamma_k^2
\end{aligned}
$$

$$
\begin{aligned}
=&\lambda_k^2(1 + c_1\alpha_k\gamma_k\mu)\|\theta_k - \theta^*\|^2 + c_1\alpha_k\gamma_k\left(\frac{1}{\mu} + c_1\alpha_k\gamma_k\right)\|b_k\|^2 + c_2\alpha_k^2\gamma_k^2\\
\leq&\lambda_k\|\theta_k - \theta^*\|^2 + c_1\alpha_k\gamma_k\left(\frac{1}{\mu} + c_1\alpha_k\gamma_k\right)\|b_k\|^2 + c_2\alpha_k^2\gamma_k^2,
\end{aligned}
\tag{78}
$$

where (a) by (19) and the fact that $\theta^* \in \mathcal{K}$ in Assumption 19, (b) is due to the inequalities (28) and (32), (c) is due to (24), and (d) is since $-2\epsilon\times\frac{1}{\epsilon}\langle a, b\rangle = -2\langle\epsilon a, \frac{1}{\epsilon}b\rangle \leq \epsilon^2\|a\|^2 + \frac{1}{\epsilon^2}\|b\|^2$.

Taking the full expectation on both sides of (78), and letting $D_k = \mathbb{E}\big[\|\theta_k - \theta^*\|^2\big]$, we get

$$
\begin{aligned}
D_{k+1} \leq&\lambda_k D_k + c_1c_3^2\left(\frac{1}{\mu} + c_1\alpha_k\gamma_k\right)\alpha_k\gamma_k^3 + c_2\alpha_k^2\gamma_k^2.\\
\leq&\lambda_k D_k + c_1c_3^2\left(\frac{1}{\mu} + c_1\alpha_0\gamma_0\right)\alpha_k\gamma_k^3 + c_2\alpha_k^2\gamma_k^2.
\end{aligned}
\tag{79}
$$

For the following part, we let $c_1\alpha_k\gamma_k = \frac{\nu}{l+k}$, with $l > \frac{\nu}{2}(\mu + L)$ and $\nu\mu > 1$ two constant values. A practical example is to take $\alpha_k = \gamma_k = \sqrt{\frac{\nu/c_1}{l+k}}$, $\forall k \geq 0$.

Suppose that $D_k \leq \frac{D}{l+k}$ for some $k \geq 0$, we want to prove that $D_{k+1} \leq \frac{D}{l+k+1}$. Thus, making use of (79) and substituting by the step-sizes' chosen form, we have to solve for D the following inequality,

$$\begin{aligned} D_{k+1} &\leq \frac{(1-\frac{\nu}{l+k}\mu)D}{l+k} + \left(\frac{c_3^2}{c_1}\big(\frac{1}{\mu} + c_1\alpha_0\gamma_0\big) + \frac{c_2}{c_1^2}\right)\frac{\nu^2}{(l+k)^2} \\ &\leq \frac{D}{l+k+1}. \end{aligned} \tag{80}$$

Let $c = \frac{c_3^2}{c_1}\big(\frac{1}{\mu} + c_1\alpha_0\gamma_0\big) + \frac{c_2}{c_1^2}$ for simplification,

$$\begin{aligned} \frac{(l+k-\nu\mu)D}{(l+k)^2} + c\frac{\nu^2}{(l+k)^2} &\leq \frac{D}{l+k+1} \\ D(l+k-\nu\mu)(l+k+1) + c\nu^2(l+k+1) &\leq D(l+k)^2. \end{aligned} \tag{81}$$

Then,

$$\frac{c\nu^2(l+k+1)}{(\nu\mu-1)(l+k)+\nu\mu} \leq D. \tag{82}$$

Next, we define a function $q(x) = \frac{c\nu^2(l+x+1)}{(\nu\mu-1)(l+x)+\nu\mu}$ for $x \geq 0$. We know that

$$q'(x) = \frac{c\nu^2}{\big((\nu\mu-1)(l+x)+\nu\mu\big)^2} > 0, \tag{83}$$

meaning q is strictly increasing for $x \geq 0$, and $\lim_{x\to\infty} q(x) = \frac{c\nu^2}{\nu\mu-1}$. Therefore, we must have $D \geq \frac{c\nu^2}{\nu\mu-1}$. Since we can find such a constant D, we conclude that D_k is indeed bounded from above by $\frac{D}{l+k}$ for $k \geq 0$.

D.1.2 With Noise

Let the objective function F be μ-strongly convex and let $\lambda_k = 1 - c_1\alpha_k\gamma_k\mu$, then assuming $c_1\alpha_k\gamma_k \leq \frac{2}{\mu+L}$, for all $k \geq K_0$,

$$\begin{aligned} &\mathbb{E}\big[\|\theta_{k+1}-\theta^*\|^2\big|\mathcal{H}_k\big] \\ =&\mathbb{E}\big[\|\Pi_{\mathcal{K}}(\theta_k-\alpha_k g_k)-\theta^*\|^2\big|\mathcal{H}_k\big] \\ \overset{(a)}{\leq}&\mathbb{E}\big[\|\theta_k-\alpha_k g_k-\theta^*\|^2\big|\mathcal{H}_k\big] \\ =&\|\theta_k-\theta^*\|^2 - 2\alpha_k\mathbb{E}\big[\langle\theta_k-\theta^*, g_k\rangle\big|\mathcal{H}_k\big] + \alpha_k^2\mathbb{E}\big[\|g_k\|^2\big|\mathcal{H}_k\big] \\ \overset{(b)}{\leq}&\|\theta_k-\theta^*\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k-\theta^*, \nabla F(\theta_k)+b_k'\rangle + c_2'\alpha_k^2 \\ =&\|\theta_k-\theta^*\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k-\theta^*, \nabla F(\theta_k)\rangle + (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 \\ &- 2c_1\alpha_k\gamma_k\langle\theta_k-\theta^*, b_k'\rangle + c_2'\alpha_k^2 \\ =&\|\theta_k - c_1\alpha_k\gamma_k\nabla F(\theta_k)-\theta^*\|^2 - (c_1\alpha_k\gamma_k)^2\|\nabla F(\theta_k)\|^2 - 2c_1\alpha_k\gamma_k\langle\theta_k-\theta^*, b_k'\rangle + c_2'\alpha_k^2 \\ \overset{(c)}{\leq}&\lambda_k^2\|\theta_k-\theta^*\|^2 - (c_1\alpha_k\gamma_k\mu)^2\|\theta_k-\theta^*\|^2 + 2c_1c_3'\alpha_k\gamma_k^2\|\theta_k-\theta^*\| + c_2'\alpha_k^2 \\ =&(1-2c_1\alpha_k\gamma_k\mu)\|\theta_k-\theta^*\|^2 + 2c_1c_3'\alpha_k\gamma_k^2\|\theta_k-\theta^*\| + c_2'\alpha_k^2. \end{aligned} \tag{84}$$

where (a) by (19) and the fact that $\theta^* \in \mathcal{K}$ in Assumption 19 and (b) is due to (29) and (33). (c) is due to (24), the strong convexity inequality, and (31) respectively.

Thus, further assuming $2c_1\alpha_k\gamma_k\mu < 1$ for $k \geq K_0$, and taking the expectation over $\mathcal{H}_k$ on both sides of the previous inequality,

$$D_{k+1} \leq (1 - 2c_1\alpha_k\gamma_k\mu)D_k + 2c_1c_3'\alpha_k\gamma_k^2\sqrt{D_k} + c_2'\alpha_k^2. \tag{85}$$

To study the evolution of D_k, we let $U_k \leq U_{k+1}$ be a decreasing sequence, and we assume that $D_k \leq U_k$, $\forall k \geq 0$. Then, to find U_k and verify its existence, we substitute in (85),

$$\begin{aligned} D_{k+1} &\leq (1 - 2c_1\alpha_k\gamma_k\mu)U_k + 2c_1c_3'\alpha_k\gamma_k^2\sqrt{U_k} + c_2'\alpha_k^2 \\ &\leq U_{k+1} \leq U_k. \end{aligned} \tag{86}$$

Therefore, we must have

$$U_k \geq \left(\frac{c_3'}{2\mu}\gamma_k + \frac{1}{2}\sqrt{\frac{c_3'^2}{\mu^2}\gamma_k^2 + \frac{2c_2'}{c_1\mu}\frac{\alpha_k}{\gamma_k}}\right)^2, \tag{87}$$

meaning we cannot get a rate better than that of γ_k^2 or $\frac{\alpha_k}{\gamma_k}$. Thus, we study both possibilities:

1. We assume $D_k \leq \zeta_1^2\gamma_k^2$ with ζ_1 some constant.

   We want to make sure that $D_{k+1} \leq U_{k+1}$ can be obtained from $D_k \leq U_k$, $\forall k \geq K_0$. Take $U_k = \zeta_1^2\gamma_k^2$, let $D_k \leq U_k$ hold, and substitute in (85),

   $$D_{k+1} \leq (1 - 2c_1\alpha_k\gamma_k\mu)\zeta_1^2\gamma_k^2 + 2c_1c_3'\zeta_1\alpha_k\gamma_k^3 + c_2'\alpha_k^2. \tag{88}$$

   We solve $D_{k+1} \leq U_{k+1}$ for $\zeta_1 \in \mathbb{R}^+$,

   $$\begin{aligned} (1 - 2c_1\alpha_k\gamma_k\mu)\zeta_1^2\gamma_k^2 + 2c_1c_3'\zeta_1\alpha_k\gamma_k^3 + c_2'\alpha_k^2 &\leq U_{k+1} \\ &= \zeta_1^2\gamma_{k+1}^2. \end{aligned} \tag{89}$$

   Then, by considering $\varpi_k = \frac{1-(\frac{\gamma_{k+1}}{\gamma_k})^2}{\alpha_k\gamma_k} > 0$

   $$(\varpi_k - 2c_1\mu)\zeta_1^2 + 2c_1c_3'\zeta_1 + c_2'\alpha_k\gamma_k^{-3} \leq 0, \tag{90}$$

   and assuming $\varpi_k - 2c_1\mu < 0$, we find a constant $\bar{\zeta_1}$ such that $\zeta_1 \geq \bar{\zeta_1}$ and

   $$\bar{\zeta_1} = \frac{c_1c_3'}{2c_1\mu - \varpi_k} + \sqrt{\left(\frac{c_1c_3'}{2c_1\mu - \varpi_k}\right)^2 + \frac{c_2'\alpha_k\gamma_k^{-3}}{2c_1\mu - \varpi_k}}, \tag{91}$$

   keeping in mind that $2c_1c_3'$ and $c_2'\alpha_k\gamma_k^{-3}$ are positive by definition. Then, for $\sigma_1 = \max_k \varpi_k$ and $\sigma_2 = \max_k \alpha_k\gamma_k^{-3}$,

   $$\bar{\zeta_1} \leq \frac{c_1c_3'}{2c_1\mu - \sigma_1} + \sqrt{\left(\frac{c_1c_3'}{2c_1\mu - \sigma_1}\right)^2 + \frac{c_2'\sigma_2}{2c_1\mu - \sigma_1}}. \tag{92}$$

   We conclude that $D_k \leq \zeta_1^2\gamma_k^2$ where ζ_1 satisfies

   $$\zeta_1 \geq \max\left\{\frac{\sqrt{D_{K_0}}}{\gamma_{K_0}}, \bar{\zeta_1}\right\}. \tag{93}$$

2. We assume $D_k \le \zeta_2^2 \frac{\gamma_k}{\alpha_k}$ with ζ_2 some constant.

$$D_{k+1} \le (1 - 2c_1\mu\alpha_k\gamma_k)\zeta_2^2 \frac{\alpha_k}{\gamma_k} + 2c_1c_3'\alpha_k\gamma_k^2\zeta_2\sqrt{\frac{\alpha_k}{\gamma_k}} + \alpha_k^2 c_2'. \tag{94}$$

Solving $D_{k+1} \le \zeta_2^2 \frac{\alpha_{k+1}}{\gamma_{k+1}}$ for $\zeta_2 \in \mathbb{R}^+$,

$$(1 - 2c_1\mu\alpha_k\gamma_k)\zeta_2^2 \frac{\alpha_k}{\gamma_k} + 2c_1c_3'\alpha_k\gamma_k^2\zeta_2\sqrt{\frac{\alpha_k}{\gamma_k}} + \alpha_k^2 c_2' \le \zeta_2^2 \frac{\alpha_{k+1}}{\gamma_{k+1}}. \tag{95}$$

Take $\tau_k = \frac{\frac{\alpha_k}{\gamma_k} - \frac{\alpha_{k+1}}{\gamma_{k+1}}}{\alpha_k^2} > 0$, then

$$(\tau_k - 2c_1\mu)\zeta_2^2 + 2c_1c_3'\gamma_k^{\frac{3}{2}}\alpha_k^{-\frac{1}{2}}\zeta_2 + c_2' \le 0. \tag{96}$$

If $\frac{\alpha_k}{\gamma_k} - \frac{\alpha_{k+1}}{\gamma_{k+1}} < 2c_1\mu\alpha_k^2$, then $\exists\ \bar{\zeta_2}$ such that $\zeta_2 \ge \bar{\zeta_2}$ and

$$\bar{\zeta_2} = \frac{c_1c_3'\gamma_k^{\frac{3}{2}}\alpha_k^{-\frac{1}{2}}}{2c_1\mu - \tau_k} + \sqrt{\left(\frac{c_1c_3'\gamma_k^{\frac{3}{2}}\alpha_k^{-\frac{1}{2}}}{2c_1\mu - \tau_k}\right)^2 + \frac{c_2'}{2c_1\mu - \tau_k}} \tag{97}$$

Let $\sigma_3 = \max_k \tau_k$ and $\sigma_4 = \max_k \gamma_k^{\frac{3}{2}}\alpha_k^{-\frac{1}{2}}$, we can say

$$\bar{\zeta_2} \le \frac{c_1c_3'\sigma_4}{2c_1\mu - \sigma_3} + \sqrt{\left(\frac{c_1c_3'\sigma_4}{2c_1\mu - \sigma_3}\right)^2 + \frac{c_2'}{c_1\mu - \sigma_3}}. \tag{98}$$

We conclude that $D_k \le \zeta_2^2 \frac{\alpha_k}{\gamma_k}$ with ζ_2 satisfying

$$\zeta_2 \ge \max\left\{\sqrt{D_{K_0}\frac{\gamma_{K_0}}{\alpha_{K_0}}}, \bar{\zeta_2}\right\}. \tag{99}$$

The previous analysis indicates that the convergence rate is a function of υ_1 and υ_2, as $\gamma_k^2 \propto (k+1)^{-2\upsilon_2}$ and $\frac{\alpha_k}{\gamma_k} \propto (k+1)^{-(\upsilon_1 - \upsilon_2)}$. -theless, we must still verify the validity of the assumptions we utilized for the analysis, meaning:

- Are $\sigma_1 < 2c_1\mu$ and $\sigma_3 < 2c_1\mu$ fulfilled?
- Are ζ_1 and ζ_2 bounded?

Let α_k and γ_k have the forms given in Example 2 with $l = 1$, i.e., $\alpha_k = \alpha_0(1+k)^{-\upsilon_1}$ and $\gamma_k = \gamma_0(1+k)^{-\upsilon_2}$.

1. **Verifying that $\sigma_1 < 2c_1\mu$ and $\sigma_3 < 2c_1\mu$**

   In this part, we find conditions on α_0 and γ_0 that allow $\sigma_1 < 2c_1\mu$ and $\sigma_3 < 2c_1\mu$ to be satisfied.

$$\sigma_1 = \max_{k \ge K_0} \frac{1 - (\frac{\gamma_{k+1}}{\gamma_k})^2}{\alpha_k\gamma_k} = \max_{k \ge K_0} \frac{1 - (1 + \frac{1}{k+1})^{-2\upsilon_2}}{\alpha_0\gamma_0(k+1)^{-\upsilon_1 - \upsilon_2}}$$

and

$$\sigma_3 = \max_{k \geq K_0} \frac{1 - \frac{\alpha_{k+1}\gamma_{k+1}^{-1}}{\alpha_k \gamma_k^{-1}}}{\alpha_k \gamma_k} = \max_{k \geq K_0} \frac{1 - (1 + \frac{1}{k+1})^{-(\upsilon_1 - \upsilon_2)}}{\alpha_0 \gamma_0 (k+1)^{-\upsilon_1 - \upsilon_2}}.$$

To find an upper bound on σ_3 and σ_1, we define a function $q(x) = x^{-a}(1 - (1+x)^{-b})$ with $a, b, x \in (0, 1]$. Since $x^{-a} \leq x^{-1}$, we have $q(x) \leq x^{-1}(1 - (1+x)^{-b}) = r(x)$. We then analyze the derivative of $r(x)$ to study the variation of $q(x)$,

$$r'(x) = x^{-2}\bigg(((b+1)x + 1)(1+x)^{-b-1} - 1 \bigg) = x^{-2} s(x).$$

We can see that the sign of $r'(x)$ is that of $s(x)$. We again write the derivative of $s(x)$ to find its sign,

$$s'(x) = -b(b+1)x(1+x)^{-b-2} \leq 0$$

since $b > 0$ and $x > 0$. Then, $s(x)$ is a decreasing function of x over $(0, 1]$. We remark that $\lim_{x \to 0} s(x) = 0$, meaning $s(x) < 0$ and $r'(x) < 0$, $\forall x \in (0, 1]$. Finally,

$$r(x) < \lim_{x \to 0} r(x) = \frac{1 - (1+x)^{-b}}{x} = b,$$

and $q(x) \leq r(x) < b$, noting that $\lim_{x \to 0} q(x) = b$ for $a = 1$. We conclude that $\sigma_1 < \frac{2\upsilon_2}{\alpha_0 \gamma_0}$ and $\sigma_1 < \frac{\upsilon_1 - \upsilon_2}{\alpha_0 \gamma_0}$. For $\sigma_1 < 2c_1\mu$ and $\sigma_3 < 2c_1\mu$ to be valid, we must have

$$\alpha_0 \gamma_0 \geq \max\{2\upsilon_2, \upsilon_1 - \upsilon_2\}/2c_1\mu. \tag{100}$$

2. **Verifying that ζ_1 and ζ_2 are bounded**

As $D_k \leq \zeta_1^2 \gamma_k^2$ and $D_k \leq \zeta_2^2 \frac{\alpha_k}{\gamma_k}$, the goal here to verify that the constant part is bounded. To do so, and referring to (92) and (98), we only need to analyze $\sigma_2 = \max_k \alpha_k \gamma_k^{-3}$ and $\sigma_4 = \max_k \gamma_k^{\frac{3}{2}} \alpha_k^{-\frac{1}{2}}$.

$$\sigma_2 = \alpha_0 \gamma_0^{-3} \max_{k \geq K_0} (1+k)^{-(\upsilon_1 - 3\upsilon_2)} = \begin{cases} \alpha_0 \gamma_0^{-3} (1 + K_0)^{-(\upsilon_1 - 3\upsilon_2)}, & \text{if } \upsilon_1 \geq 3\upsilon_2, \\ \infty, & \text{if } \upsilon_1 < 3\upsilon_2, \end{cases}$$

and

$$\sigma_4 = \alpha_0^{-\frac{1}{2}} \gamma_0^{\frac{3}{2}} \max_{k \geq K_0} (1+k)^{-\frac{1}{2}(3\upsilon_2 - \upsilon_1)} = \begin{cases} \alpha_0^{-\frac{1}{2}} \gamma_0^{\frac{3}{2}} (1 + K_0)^{-\frac{1}{2}(3\upsilon_2 - \upsilon_1)}, & \text{if } \upsilon_1 \leq 3\upsilon_2, \\ \infty, & \text{if } \upsilon_1 > 3\upsilon_2. \end{cases}$$

We deduce the following 3 cases:

- When $\upsilon_1 > 3\upsilon_2$:
  σ_2 is bounded. ζ_1 (by definition in (92)) is also bounded provided that $\alpha_0 \gamma_0 \geq \frac{\upsilon_2}{c_1 \mu}$ in (100).
  However, $\sigma_4 \to \infty$ causing $\zeta_2 \to \infty$ (in (98)) and thus a loose upper bound in $D_k \leq \zeta_2^2 \frac{\alpha_k}{\gamma_k}$.
  To that end, we can write $D_k \leq D_1 (1+k)^{-2\upsilon_2}$ with D_1 a bounded constant.

- When $\upsilon_1 < 3\upsilon_2$:
  Similarly, σ_4 is bounded while $\sigma_2 \to \infty$. Then, $\exists\ D_2 < \infty$, where $D_k \leq D_2(1+k)^{-(\upsilon_1-\upsilon_2)}$ provided that $\alpha_0\gamma_0 \geq \frac{\upsilon_1-\upsilon_2}{2c_1\mu}$.
- When $\upsilon_1 = 3\upsilon_2$:
  Both σ_2 and σ_4 are bounded allowing both previous inequalities corresponding to D_k to be valid.

Optimizing over υ_1 and υ_2 under the constraints $0 < \upsilon_1 + \upsilon_2 \leq 1$, $1 < \upsilon_1 + 2\upsilon_2$, and $\upsilon_1 > 0.5$ given in Assumptions 9 and 17 for noisy environments, we obtain the optimal rate of $D_k \leq D'(1+k)^{-\frac{1}{2}}$ for $\upsilon_1 = \frac{3}{4}$, $\upsilon_2 = \frac{1}{4}$, and D' a bounded constant.

D.2 Strongly Convex Objective Function with Fixed Step Sizes Convergence Rate

D.2.1 Without Noise

Following up from inequality (78) and setting $\alpha_k = \alpha$ and $\gamma_k = \gamma$,

$$D_{k+1} \leq \lambda D_k + c_1 c_3^2 \alpha\gamma^3 \left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c_2\alpha^2\gamma^2. \tag{101}$$

By taking the telescoping sum,

$$\begin{aligned} D_{K+1} \leq& \lambda^{K+1} D_0 + \alpha\gamma\bigg(c_1 c_3^2\gamma^2\left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c_2\alpha\gamma\bigg)\sum_{j=0}^{K}\lambda^j \\ =& \lambda^{K+1} D_0 + \alpha\gamma\bigg(c_1 c_3^2\gamma^2\left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c_2\alpha\gamma\bigg)\frac{1-\lambda^{K+1}}{1-\lambda}. \end{aligned} \tag{102}$$

D.2.2 With Noise

Note that inequality (78) is valid for the case of noisy environments when b_k is replaced by b'_k, $c_2\gamma_k^2$ by c'_2, and c_3 by c'_3. Thus, setting $\alpha_k = \alpha$ and $\gamma_k = \gamma$,

$$D_{k+1} \leq \lambda D_k + c_1 c_3'^2 \alpha\gamma^3 \left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c'_2\alpha^2. \tag{103}$$

Similarly, taking the telescoping sum,

$$\begin{aligned} D_{K+1} \leq& \lambda^{K+1} D_0 + \alpha\bigg(c_1 c_3'^2\gamma^3\left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c'_2\alpha\bigg)\sum_{j=0}^{K}\lambda^j \\ =& \lambda^{K+1} D_0 + \alpha\bigg(c_1 c_3'^2\gamma^3\left(\frac{1}{\mu} + c_1\alpha\gamma\right) + c'_2\alpha\bigg)\frac{1-\lambda^{K+1}}{1-\lambda}. \end{aligned} \tag{104}$$

Appendix E. Experimental Details

- Figure 2: For the training model, for 2P-ZOFL without noise, we set $\alpha_k = 0.4(1+k)^{-0.26}$ and $\gamma_k = 0.7(1+k)^{-0.26}$ and with noise, we set $\alpha_k = 0.65(1+k)^{-0.51}$ and $\gamma_k = 0.75(1+k)^{-0.18}$. For FedAvg, $\eta = 0.01$. For FedZO, $\eta = 0.0001$ and $\mu = 0.001$.

- Figure 3: For 2P-ZOFL with the non-convex MNIST classification example and different noise variances, $\alpha_k = \alpha_0(1+k)^{-0.51}$ and $\gamma_k = \gamma_0(1+k)^{-0.18}$ and for $\sigma_n^2 = \{0.25, 1, 2.25, 3.1684, 10.0489\}$, we set $\alpha_0 = \{1.5, 1.1, 1.1, 0.5, 0.3\}$ and $\gamma_0 = \{3.5, 3.1, 3.1, 1.5, 1.3\}$, respectively.

- Figure 5: For 2P-ZOFL without noise, we set $\alpha_k = 0.7(1+k)^{-\frac{1}{2}}$ $\gamma_k = 1.5(1+k)^{-\frac{1}{2}}$ and $\alpha = 0.5$ and $\gamma = 0.95$ when the step sizes are constant. When there is noise, $\alpha_k = (1+k)^{-\frac{3}{4}}$ and $\gamma_k = 2(1+k)^{-\frac{1}{4}}$ and and $\alpha = 0.5$ and $\gamma = 0.95$ for constant step sizes.

- Figure 6: For the mushroom classification example, we consider the following step sizes for every algorithm. For 2P-ZOFL without noise, we set $\alpha_k = \gamma_k = 0.3(1+k)^{-\frac{1}{2}}$ and $\alpha = \gamma = 0.1$ when the step sizes are constant. When there is noise, $\alpha_k = 0.75(1+k)^{-\frac{3}{4}}$ and $\gamma_k = 0.75(1+k)^{-\frac{1}{4}}$ and $\alpha = \gamma = 0.5$ for constant step sizes. For FedAvg, $\eta = 0.01$.

- Figure 7: For the MNIST classification example, we consider the following step sizes for every algorithm. For 2P-ZOFL without noise, we set $\alpha_k = \gamma_k = 0.3(1+k)^{-\frac{1}{2}}$ and $\alpha = \gamma = 0.3$ when the step sizes are constant. When there is noise, $\alpha_k = 0.3(1+k)^{-\frac{3}{4}}$ and $\gamma_k = 0.3(1+k)^{-\frac{1}{4}}$ and $\alpha = \gamma = 0.3$ for constant step sizes. For FedAvg, $\eta = 0.01$.

- Figure 9: For 2P-ZOFL with the MNIST classification example and different noise variances, for $\sigma_n^2 = \{1, 2.25, 3.1684, 10.0489\}$, we set $\alpha = \gamma = \{0.2, 0.1, 0.07, 0.05\}$, respectively.

E.1 Convergence Time vs Rate

Table 2: Upload communication efficiency of ZOFL vs FedAvg (McMahan et al., 2017) vs FedZO (Fang et al., 2022) till convergence and per iteration.

Algorithm	Total Symbols Until Convergence for 1 Device	Total Symbols Until Convergence for 100 Devices	Number of Symbols per Round for 1 Device	Number of Symbols per Round for 100 Devices
ZOFL	$4,000$	$400,000$	2	200
FedAvg	$59,280,600$	$5,928,060,000$	$197,602$	$19,760,200$
FedZO	$148,201,500$	$148,201,500 \times 10$ $= 1,482,015,000$	$197,602$	$197,602 \times 10$ $= 1,976,020$

We include this quantitative study to compare with other communication-efficient strategies, like local SGD (multiple local gradient descent steps before upload) and partial device participation at every iteration. We compare with FedZO (Fang et al., 2022), which incorporates both strategies and communicates over wireless channels via analog modulation and is also a ZO method. While both these strategies help to save resources as compared with the standard FL method, they are still much less efficient than our method.

References

Mushroom. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5959T.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010.

Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 9017–9027, 2020.

Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Distributed zero-order optimization under adversarial noise. In *Advances in Neural Information Processing Systems*, volume 34, pages 10209–10220, 2021.

Mohammad Mohammadi Amiri and Deniz Gündüz. Federated learning over wireless fading channels. *IEEE Transactions on Wireless Communications*, 19(5):3546–3557, 2020. doi: 10.1109/TWC.2020.2974748.

Mohammad Mohammadi Amiri, Deniz Gündüz, Sanjeev R. Kulkarni, and H. Vincent Poor. Convergence of update aware device scheduling for federated learning at the wireless edge. *IEEE Transactions on Wireless Communications*, 20(6):3643–3658, 2021. doi: 10.1109/TWC.2021.3052681.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points, 2018. URL https://arxiv.org/abs/1809.06474.

Emil Björnson and Luca Sanguinetti. Making cell-free massive mimo competitive with mmse processing and centralized implementation. *IEEE Transactions on Wireless Communications*, 19(1):77–90, 2020. doi: 10.1109/TWC.2019.2941478.

Keith Bonawitz, Hubert Eichner, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019. URL https://proceedings.mlsys.org/paper_files/paper/2019/file/7b770da633baf74895be22a8807f1a8f-Paper.pdf.

Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin. Lag: Lazily aggregated gradient for communication-efficient distributed learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/feecee9f1643651799ede2740927317a-Paper.pdf.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/576d026223582a390cd323bef4bad026-Paper.pdf.

Yicheng Chen, Rick S. Blum, Martin Takáč, and Brian M. Sadler. Distributed learning with sparsified gradient differences. *IEEE Journal of Selected Topics in Signal Processing*, 16 (3):585–600, 2022. doi: 10.1109/JSTSP.2022.3162989.

Zhongxiang Dai, Bryan Kian Hsiang Low, and Patrick Jaillet. Federated bayesian optimization via thompson sampling. In *Advances in Neural Information Processing Systems*, volume 33, pages 9687–9699. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/6dfe08eda761bd321f8a9b239f6f4ec3-Paper.pdf`.

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. Model agnostic contrastive explanations for structured data, 2019.

Joseph L. Doob. Stochastic processes. 1953.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Anis Elgabli, Jihong Park, Amrit S. Bedi, Mehdi Bennis, and Vaneet Aggarwal. Q-gadmm: Quantized group admm for communication efficient decentralized machine learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8876–8880, 2020. doi: 10.1109/ICASSP40776.2020.9054491.

Anis Elgabli, Chaouki Ben Issaid, Amrit Singh Bedi, Ketan Rajawat, Mehdi Bennis, and Vaneet Aggarwal. FedNew: A communication-efficient and privacy-preserving Newton-type method for federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5861–5877. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/elgabli22a.html`.

Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N. Jones, and Yong Zhou. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022. doi: 10.1109/TSP.2022.3214122.

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8759–8770, 2018.

Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksandr Beznosikov, and Alexander Lobanov. *Randomized Gradient-Free Methods in Convex Optimization*, page 1–15. Springer International Publishing, September 2023. ISBN 9783030546212. doi: 10.1007/978-3-030-54621-2_859-1. URL `http://dx.doi.org/10.1007/978-3-030-54621-2_859-1`.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. In *Advances in Neural Information Processing Systems*, volume 34, pages 12052–12064. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64be20f6dd1dd46adf110cf871e3ed35-Paper.pdf.

Huayan Guo, An Liu, and Vincent K. N. Lau. Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis. *IEEE Internet of Things Journal*, 8(1):197–210, 2021. doi: 10.1109/JIOT.2020.3002925.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilyas18a.html.

Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *NIPS*, 2012.

Peter Kairouz, H. Brendan McMahan, et al. *Advances and Open Problems in Federated Learning.* 2021.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4519–4529. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/bayoumi20a.html.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients, 2018. URL https://arxiv.org/abs/1806.06573.

David Kinderlehrer and Guido Stampacchia. An introduction to variational inequalities and their application. 31, 01 2000. doi: 10.1137/1.9780898719451.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency, 2016. URL https://arxiv.org/abs/1610.05492.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Feddane: A federated newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231, 2019. doi: 10.1109/IEEECONF44664.2019.9049023.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020. doi: 10.1109/MSP.2020.2975749.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2916–2925. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/malik19a.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Elissa Mhanna and Mohamad Assaad. Zero-order one-point estimate with distributed stochastic gradient-tracking technique, 2022.

Elissa Mhanna and Mohamad Assaad. Single point-based distributed zeroth-order optimization with a non-convex stochastic objective function. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24701–24719. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mhanna23a.html.

Elissa Mhanna and Mohamad Assaad. Rendering wireless environments useful for gradient estimators: A zero-order stochastic federated learning method. In *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, pages 1–8, 2024. doi: 10.1109/Allerton63246.2024.10735291.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences, 2019. URL https://arxiv.org/abs/1901.09269.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Yurii Nesterov and Vladimir G. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527 – 566, 2015. URL https://api.semanticscholar.org/CorpusID:2147817.

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

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods, 2018. URL https://arxiv.org/abs/1805.11454.

Shi Pu, Alex Olshevsky, and Ioannis Ch. Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67 (11):5900–5915, 2022. doi: 10.1109/TAC.2021.3126253.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018. doi: 10.1109/TCNS.2017.2698261.

Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. *IEEE Journal on Selected Areas in Information Theory*, 3(2):197–205, 2022. doi: 10.1109/JSAIT.2022.3205475.

Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23(64):1–47, 2022. URL `http://jmlr.org/papers/v23/19-750.html`.

Tomer Sery and Kobi Cohen. On analog gradient descent learning over multiple access fading channels. *IEEE Transactions on Signal Processing*, 68:2897–2911, 2020. doi: 10.1109/TSP.2020.2989580.

Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C. Eldar. Over-the-air federated learning from heterogeneous data. *IEEE Transactions on Signal Processing*, 69:3796–3811, 2021. doi: 10.1109/TSP.2021.3090323.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 3–24, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v30/Shamir13.html`.

Yuxuan Sun, Sheng Zhou, Zhisheng Niu, and Deniz Gündüz. Dynamic scheduling for over-the-air federated edge learning with energy constraints. *IEEE Journal on Selected Areas in Communications*, 40(1):227–242, 2022. doi: 10.1109/JSAC.2021.3126078.

David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

Anirudh Vemula, Wen Sun, and J. Bagnell. Contrasting exploration in parameter and action space: A zeroth-order optimization perspective. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2926–2935. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/vemula19a.html`.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021. doi: 10.1109/TSP.2021.3106104.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated learning via over-the-air computation. *IEEE Transactions on Wireless Communications*, 19(3):2022–2035, 2020. doi: 10.1109/TWC.2019.2961673.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021. doi: 10.1109/TSP.2021.3115952.