DIANA WITH COMPRESSION FOR DISTRIBUTED VARIATIONAL INEQUALITIES: ELIMINATING THE NEED TO TRANSMIT FULL GRADIENTS

Anonymous authors

Paper under double-blind review

ABSTRACT

Variational inequalities (VIs) are attracting increasing interest among machine learning (ML) researchers due to their applicability in numerous areas, such as empirical risk minimization (ERM) problems, adversarial learning, generative adversarial networks (GANs), and robust optimization. The growing volume of training data necessitates the use of advanced architectures beyond single-node computations. Distributed optimization has emerged as the most natural and efficient paradigm, enabling multiple devices to perform training simultaneously. However, this setup introduces a significant challenge: devices must exchange information with each other, which can substantially reduce the speed of learning. A standard approach to mitigating this issue involves the use of heuristics that allow only partial information transmission. State-of-the-art methods with compression for distributed VIs rely on variance reduction techniques, which makes them inapplicable to practical tasks due to full gradient computation and transmission. In this paper, we obviate the need to consider full gradient computations and introduce a novel algorithm for solving distributed variational inequalities. It combines the classical DIANA algorithm with the EXTRAGRADIENT technique. Additionally, we incorporate an error compensation mechanism, enabling our algorithm to handle the class of contractive compression operators, which are more practical for real-world applications. We provide a comprehensive theoretical analysis with near-optimal convergence guarantees and additionally outperform competitors in CNN and GAN training experiments.

1 INTRODUCTION

The variational inequalities (VIs) problem first attracted the attention of researchers in economics and game theory (Von Neumann & Morgenstern, 1953). Nowadays such a problem, due to its flexible nature, is used in a variety of other fields. In this paper, we consider it in the following form:

$$\text{find } z^* \in Z \text{ such that } \langle F(z^*), z - z^* \rangle \geqslant 0 \text{ for all } z \in Z, \tag{1}$$

where $F\colon \mathbb{R}^d \to \mathbb{R}$ is a convex and monotone operator. The general setting (1) can be specified to address particular types of problems, such as classical minimization, saddle point, and fixed point problems (Kinderlehrer & Stampacchia, 2000; Facchinei & Pang, 2003). We provide several examples to show aforementioned relations.

Example 1.1. Consider the following convex unconstrained minimization problem:

$$\min_{z\in\mathbb{R}^d} \left[f(z)\right], \tag{2}$$

where f represents the smooth convex objective function. We define $F(z) = \nabla f(z)$. Then $z^* \in \mathbb{R}^d$ is the solution of equation 1 if and only if $z^* \in \mathbb{R}^d$ is the solution of equation 2. In this way, the problem (2) can be considered as a variational inequality.

Example 1.2. Now consider the convex-concave saddle point problem:

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left[f(x,y)\right], \tag{3}$$

where f represents the convex-concave smooth objective function. In this setting, we define $F(z) = F(x,y) = [\nabla_x f(x,y), -\nabla_y f(x,y)]$. Then $z^* \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ is the solution of equation 1 if and only if $z^* \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ is the solution of equation 3. Consequently, the problem (3) admits a reformulation as a variational inequality.

Highlighting the numerous applications of variational inequalities, their expressive power is evident in areas such as discriminative clustering (Xu et al., 2004), matrix factorization (Bach et al., 2008), image denoising (Esser et al., 2010; Chambolle & Pock, 2011), and robust optimization (Ben-Tal et al., 2009). Furthermore, the problem (1) has found applications in cutting-edge machine learning approaches, including adversarial training (Madry et al., 2017), reinforcement learning (Omidshafiei et al., 2017; Jin & Sidford, 2020), and generative adversarial networks (GANs) (Goodfellow et al., 2014). The study of classical techniques for variational inequalities, such as averaging and extrapolation, provides valuable tools for a deeper understanding of DL training (Daskalakis et al., 2017; Gidel et al., 2018; Liang & Stokes, 2019; Peng et al., 2020). In summary, variational inequalities have recently become a pivotal tool for the machine learning community. The development of methods for solving these problems is essential for achieving advanced results in various areas of ML in the future.

On the other hand, the increasing volume of data in training samples presents new challenges for the ML community. Indeed, relying on a single node for computation makes it impossible to train state-of-the-art models with the desired quality within a reasonable time frame. The distributed learning paradigm has emerged as the natural solution to this problem (Verbraeken et al., 2020; Konečný et al., 2016; McMahan et al., 2017). Formally, we can describe this setting by choosing F in equation 1 of the following form:

$$F(z) = \frac{1}{M}\sum_{i=1}^{M} F_i(z), \tag{4}$$

where M is the number of devices in the network, and $F_i(\cdot)$ is a local operator based on the data stored at the i-th device. Consider an empirical risk minimization problem

$$\min_{w\in\mathbb{R}^d} \frac{1}{M}\sum_{i=1}^{M} \ell_i(g(w,x_i),y_i) + \frac{\lambda}{2}\|w\|^2 \tag{5}$$

(with $g(w,x_i)$ as the output of the machine learning model, $\ell(g(w,x_i),y_i)$ — loss function, $w \in \mathbb{R}^d$ — the model parameters, x_i and y_i — input data and label, respectively). In Madry et al. (2017) it was reformulated as a more broader $\min\max$ problem

$$\min_{w}\max_{\sigma} \frac{1}{M}\sum_{i=1}^{M} \ell_i(g(w,x_i+\sigma),y_i) + \frac{\lambda_1}{2}\|w\|^2 - \frac{\lambda_2}{2}\|\sigma\|^2, \tag{6}$$

where σ represents adversarial noise introduced to model data perturbations, and λ_1, λ_2 are regularization parameters. This formulation can be expressed as a variational inequality (see Example 1.2) with:

$$z = \begin{pmatrix} w \\ \sigma \end{pmatrix}, \quad F_i(z) = \begin{pmatrix} \nabla_w \ell_i(w,x_i+\sigma,y_i) + \lambda_1 w \\ -\nabla_\sigma \ell_i(w,x_i+\sigma,y_i) + \lambda_2\sigma \end{pmatrix}.$$

However, despite this setting paralleling computations across M nodes, there is no M-fold acceleration. This bottleneck arises from the need for devices to exchange information to solve a common problem. One successful approach to mitigate communication costs is to compress the transmitted data. The core idea behind this approach is that by utilizing a sufficiently large number of workers in the network, the error caused by partial information exchange decreases as the number of transmitted bits is reduced, resulting in faster training (Gorbunov et al., 2021a; Mishchenko et al., 2024).

Proceeding with the formal description of this approach, we introduce two general classes of compression operators.

Definition 1.3 (Unbiased compression operator)**.** We say that a (possibly) stochastic mapping $Q : \mathbb{R}^d \to \mathbb{R}^d$ is an unbiased compression operator if there exists a constant $\omega \geq 1$ such that for all $z \in \mathbb{R}^d$ it implies

$$\mathbb{E}Q(z) = z,\ \mathbb{E}\|Q(z)\|^2 \leq \omega\|z\|^2.$$

Example 1.4. A canonical example of a *unbiased* compression operator is the $\textsc{Rand}n$ mapping Karimireddy et al. (2019); Horváth et al. (2023), defined as follows: given a vector $x \in \mathbb{R}^d$, $\textsc{Rand}n$ randomly selects n coordinates (uniformly without replacement) and returns their scaled identity projection:

$$\textsc{Rand}n(x) = \frac{d}{n}\cdot \mathrm{Proj}_S(x),$$

where $S \subseteq [d]$ denotes the randomly selected subset of size n, and $\mathrm{Proj}_S(x)$ retains only the coordinates in S, setting the rest to zero.

Definition 1.5 (Contractive compression operator)**.** We say that a (possibly) stochastic mapping $Q : \mathbb{R}^d \to \mathbb{R}^d$ is a contractive compression operator if there exists a constant $0 < \alpha \leq 1$ such that for all $z \in \mathbb{R}^d$ it implies

$$\mathbb{E}\left\|Q(z) - z\right\|^2 \leq (1 - \alpha)\left\|z\right\|^2.$$

Example 1.6. An important example of a *biased* compression operator is the Top$n$ compressor Stich et al. (2018); Karimireddy et al. (2019), which selects the n coordinates of largest magnitude while discarding the remaining entries. Formally, it is defined as:

$$\textsc{Top}n(x) = \mathrm{Proj}_T(x),$$

where $T \subseteq [d]$ contains the indices corresponding to the n largest absolute values in x. Unlike Rand$n$, this compressor introduces bias due to the deterministic selection of the most significant components.

Given the Examples 1.4, 1.6 of the compression operators, one can note that the Rand$n$ compressor satisfies Definition 1.3 with $\omega = \frac{d}{n}$, and Top$n$ satisfies Definition 1.5 with $\alpha = \frac{n}{d}$ (Alistarh et al., 2018; Beznosikov et al., 2023).

In this work, we present a novel approach to solving distributed variational inequalities with compression, encompassing both unbiased and contractive classes of compressors.

1.1 Brief literature review

Starting with methods for solving variational inequalities, the standard choice is the SGDA (Fallah et al., 2020) and Extragradient (Korpelevich, 1976) methods, however the last one has more accurate theoretical confirmation. This approach has been extended into various stochastic (Juditsky et al., 2011; Mishchenko et al., 2020; Gorbunov et al., 2022; Medyakov et al., 2024) and distributed versions (Srivastava et al., 2011; Liu et al., 2020; Mukherjee & Chakraborty, 2020; Rogozin et al., 2021). In addition to Extragradient, the Mirror-prox method (Nemirovski, 2004) allows for the incorporation of non-Euclidean geometry. Other notable methods include forward-backward-forward (Tseng, 2000), Dual extrapolation (Nesterov, 2007), reflected gradient (Malitsky, 2015), and forward-reflected-backward (FoRB) (Malitsky & Tam, 2020), each offering specific advancements.

Next, we review communication-efficient approaches that utilize unbiased compression operators (see Definition 1.3) for classical minimization problems (Example 1.1). A prominent example is QSGD (Alistarh et al., 2017), which introduced a variant of SGD (Robbins & Monro, 1951; Lan, 2020) with compressed updates. Other methods built upon SGD include (Bernstein et al., 2018; Khirirat et al., 2018; Konečný & Richtárik, 2018). More advanced algorithms, such as DIANA (Mishchenko et al., 2024; Li & Richtárik, 2020), VR-DIANA (Horváth et al., 2023), FedCOMGATE (Haddadpour et al., 2021), and MARINA (Gorbunov et al., 2021a), provide estimates under stochastic distributed settings. DIANA is based on compressing the difference between two gradient approximations, while MARINA employs the variance reduction technique SARAH (Nguyen et al., 2017). The MARINA algorithm achieves optimal communication and oracle complexity for non-convex smooth objectives. However, it requires periodic exchanges of full gradients, which limits its practical application in federated settings. The DASHA method (Tyurin & Richtárik, 2022) addresses this limitation while maintaining optimal communication complexity.

Table 1: Comparison of the convergence results.

	Algorithm	Setup	Convergence to optimum	No Full Grad.?
Unbiased Compressors	QSGD (Alistarh et al., 2017)	min	✗	✓
	DIANA (Mishchenko et al., 2020)	min	✓	✓
	MARINA (Gorbunov et al., 2021a)	min	✓	✗
	DASHA (Tyurin & Richtárik, 2022)	min	✓	✓
	MASHA1 (Beznosikov et al., 2022)	min max	✓	✗
	EG DIANA (Algorithm 1)	min max	✓	✓
Contractive Compressors	EF (Karimireddy et al., 2019)	min	✗	✓
	EF21 (Richtárik et al., 2021)	min	✓	✓
	MASHA2 (Beznosikov et al., 2022)	min max	✓	✗
	EG DIANA EF (Algorithm 2)	min max	✓	✓

Columns: Setup = min if considered classical minimization problem, min max if considered VIs problem, Compression Boost = communication complexity of the algorithm and uncompressed training ratio, Contr. Comp. = whether the method works with contractive compressors, No Full Grad.? = whether the method requires full gradients transmission.

Approaches that enable convergence with contractive compressors (see Definition 1.5) deserve special attention. The ERROR FEEDBACK technique was proposed in (Seide et al., 2014; Ström, 2015), with theoretical analysis provided in (Stich et al., 2018; Karimireddy et al., 2019). Later, (Richtárik et al., 2021) introduced an advanced variant of this technique, EF21. A comparison is presented in Table 1. Although compression techniques for minimization problems have been well-studied, their application to variational inequalities remains unexplored. Several works address the communication bottleneck using various approaches, such as local steps (Yuan et al., 2014; Hou et al., 2021; Deng & Mahdavi, 2021) and data similarity (Beznosikov et al., 2020; 2021). However, to the best of our knowledge, there are almost no works that utilize compression operators, with the exception of two papers.
In (Yuan et al., 2014), rounding to the nearest multiple of a certain quantity was proposed, but this compressor did not achieve convergence to the solution. The work (Beznosikov et al., 2022) introduced the MASHA1 algorithm, which achieves optimal communication complexity for unbiased compressors and is $(M\omega)^{1/2}$ times worse for contractive compressors. To handle contractive compressors, the authors of (Beznosikov et al., 2022) incorporated the ERROR FEEDBACK technique. However, MASHA2 is based on the variance reduction approach SVRG (Johnson & Zhang, 2013; Alacaoglu & Malitsky, 2022), which requires regular transmission of the full operator value to all devices. This is a significant limitation, as broadcasting such data can be extremely time-consuming and impractical in federated learning scenarios. Later, the authors of (Beznosikov & Gasnikov, 2022) extended MASHA1 to a similarity setup (Shamir et al., 2014; Arjevani & Shamir, 2015; Kovalev et al., 2022), and in (Beznosikov et al., 2024), they incorporated local steps (Khaled et al., 2020; Gorbunov et al., 2021b; Luo et al., 2025).

1.2 OUR CONTRIBUTIONS

• **Novel Approach.** In contrast to the MASHA algorithms, which require the full transmission of operators at each iteration, we propose a novel and more efficient approach that eliminates this necessity. Rather than employing the variance reduction mechanism of SVRG, we draw upon the principles of DIANA, a distributed optimization method designed to accommodate compressed communication. We introduce the EXTRAGRADIENT DIANA algorithm, specifically tailored to operate with unbiased compressors. Additionally, we incorporate the ERROR FEEDBACK mechanism, which ensures convergence even when contractive compressors are utilized. To the best of our knowledge, this work is the first to combine DIANA with ERROR FEEDBACK, not only in the context of extragradient methods but also in classical minimization settings, thereby opening new avenues for research in compressed gradient-based optimization.
• **Theoretical Analysis.** We provide a comprehensive theoretical analysis, considering a practically relevant setup with M nodes in a distributed network. For the algorithm with ERROR FEEDBACK, we assume strongly monotone operators in (1) and L_i-Lipschitz operators in (4). For the algorithm without error compensation, we additionally assume L-Lipschitz continuity for whole operator in (4).
• **Experiments.** Our comprehensive experiments demonstrate great advantages of our proposed algorithm: it shows not only superior performance to MASHA2 for $\min\max$ problems, but also an improved convergence compared to EF21 and DIANA originally designed for standard minimization problems. We validate these claims through two benchmark studies. First, we reformulate a standard minimization task (image classification on CIFAR-10 using RESNET18) as a saddle-point problem, where our method achieves faster convergence than MASHA2 and higher accuracy. Moreover, it shows competitive results with DIANA. Second, we demonstrate practical scalability through STYLEGAN training on the I'M SOMETHING OF A PAINTER MYSELF dataset, where our algorithm significantly outperforms MASHA2 in terms of convergence and final metrics.

2 ALGORITHMS AND ANALYSIS

We begin with the list of assumptions regarding the operator in equation 1 and equation 4, under which we develop our theoretical analysis.

Assumption 2.1. Each operator F_i is L_i-Lipschitz, i.e., it satisfies $\|F_i(z_1) - F_i(z_2)\| \leq L_i\|z_1 - z_2\|$ for any $z_1, z_2 \in Z$. We denote $\hat{L} = \sqrt{1/M \sum_{i=1}^{M} L_i^2}$.

Assumption 2.2. Full operator F is L-Lipschitz, i.e., it satisfies $\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|$ for any $z_1, z_2 \in Z$.

As a practical example of a Lipschitz operator F, consider the gradient of an objective function. In the distributed setting, each agent may be associated with its own local gradient $F_i = \nabla f_i$, which is L_i-Lipschitz continuous (Richtárik & Takáč, 2016; Beznosikov et al., 2020; 2022).

Assumption 2.3. Full operator F is μ-strongly monotone, i.e., it satisfies $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geqslant \mu \|z_1 - z_2\|^2$ for any $z_1, z_2 \in Z$.

This assumption is standard in the analysis of distributed optimization and variational inequality methods (Scutari et al., 2013; Koshal et al., 2016) for foundational uses of strong monotonicity in multi-agent and game-theoretic settings.

Now we move to the algorithm, which holds unbiased compression operators, i.e., compressors from Definition 1.3.

2.1 Extragradient DIANA for unbiased compressors

In this section, we present the Extragradient DIANA algorithm (Algorithm 1).

Originally, the idea of DIANA (Mishchenko et al., 2024) originated from the problems of classical QSGD, where naive compression of the local gradient/operator is performed. Indeed, $F_i(z^*)$ is not equal to zero, then the variance of $Q(F_i(z^*))$ does not reduce in the neighborhood of the optimum, and the method does not converge (Beznosikov et al., 2023). Therefore, DIANA compresses not the operator itself, but its difference with a specially constructed memory sequence h_i^k that is striving to $F_i(z^*)$. In this way, we apply a compression operator to the value that is going to zero, allowing the method to converge.

Algorithm 1 Extragradient DIANA

```
1: Input: initial point $z^0 \in Z$, initial vectors $\{h_i^0\}_{i=1}^M$, each $h_i^0 \in Z$, $h^0 = \frac{1}{M}\sum_{i=1}^M h_i^0$, compressor $Q$ is under Definition 1.3, number or workers $M$, number of iterations $K$
2: Parameters: Stepsize $\gamma > 0$
3: for $k = 0, 1, 2, ..., K-1$ do
4:     $z^{k+\frac{1}{2}} = z^k - \gamma h^k$
5:     The server broadcast $z^{k+\frac{1}{2}}$ to all workers
6:     for $i = 1, 2, \dots, M$ in parallel do
7:         Compute $F_i(z^{k+\frac{1}{2}})$
8:         $\Delta_i^k = F_i(z^{k+\frac{1}{2}}) - h_i^k$
9:         Compress $\hat{\Delta}_i^k \sim Q(\Delta_i^k)$
10:        Grad correction $h_i^{k+1} = h_i^k + \frac{1}{1+\omega}\hat{\Delta}_i^k$
11:        Broadcast $\hat{\Delta}_i^k$ to the server
12:    end for
13:    $\hat{\Delta}^k = \frac{1}{M}\sum_{i=1}^M \hat{\Delta}_i^k$
14:    $g^k = h^k + \hat{\Delta}^k$
15:    $z^{k+1} = z^k - \gamma g^k$
16:    $h^{k+1} = h^k + \frac{1}{1+\omega}\hat{\Delta}^k$
17: end for
18: Output: $z^K$
```

To better understand how this idea integrates with the Extragradient method, let us recall its original update steps:

$$
\begin{aligned}
z^{k+\frac{1}{2}} &= z^k - \gamma F(z^k), \\
z^{k+1} &= z^k - \gamma F(z^{k+\frac{1}{2}}).
\end{aligned}
$$

In original DIANA, the step is performed using g^k, which serves as an unbiased estimator of the full operator (Line 14). We maintain this idea for the main step of the method (Line 15), and the server retains this estimation through the entire computing process. The question arises regarding what to employ for the step to the extrapolation point $z^{k+1/2}$. One can try to use a similar estimator of the full gradient at a different point, e.g. z^k, but this complicates the analysis. In our method, we propose a more natural solution, namely to complete this step using an extra vector h_k (Line 4). In this case, the difference $(F_i(z^{k+1/2}) - h_i^k)$ appears in the proof and can be easily estimated due to its convergence to zero. These modifications allow us to leverage both the h^k and g^k sequences while preserving the core concept of variance reduction.

It is also important to highlight the communication aspects of our approach. Given a set of vectors $\Delta_1, \dots, \Delta_M \in Z$, each device computes in parallel and transmits a compressed version $Q(\Delta_i)$ to the central server at each aggregation round k. Notably, we do not apply compression when the server sends the intermediate point $z^{k+1/2}$ back to the devices (Line 5), as our primary focus is on reducing the communication cost associated with the exchange of gradients (Kairouz et al., 2021).

We present the following definition to further quantify the communication efficiency of our approach.

Definition 2.4. Let us define the expected density of the transmission with an unbiased compressor as

$$
q_\omega^{-1} = {}^1\!/_{bd} \cdot \mathbb{E}\, \|Q(z)\|_{\text{bits}}\,,
$$

where $\|z\|_{\text{bits}}$ represents the number of bits required to encode the vector z, b denotes the number of bits per floating-point value, and d is the dimensionality of the problem (i.e., $bd = \|z\|_{\text{bits}}$). Similarly, we can define q_α for the contractive compressors.

Formally, we assume a synchronous setup with equivalent devices and communication channels. Transmitting one unit of data from devices to the server takes τ time units. If clients start communicating simultaneously, channel initialization is negligible. Thus, one round of communication takes τN, where N is the amount of data sent per device. Consequently, we can infer that the communication complexity at each iteration scales as $\mathcal{O}\left({}^{1}/_{q_\omega}\right)$ or $\mathcal{O}\left({}^{1}/_{q_\alpha}\right)$. Note that for the aforementioned compressor operators TOPn and RANDn, according to the introduced Definition 2.4, the following equality holds: $\omega = \frac{1}{\alpha} = q_\omega = q_\alpha = \frac{d}{n}$.

Having established the methodological framework, we now turn our attention to the theoretical analysis of our approach.

Theorem 2.5. *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then for Algorithm 1 with*

$$\gamma \leq \min\left\{\frac{1}{4\omega\mu}, \frac{1}{8\sqrt{30}\omega L}, \frac{\sqrt{M}}{40\sqrt{6}\cdot\omega^{3/2}\hat{L}}\right\}$$

and $\beta = \frac{1}{\omega}$ the following estimate is fulfilled:

$$\mathbb{E}V^K \leq (1-\gamma\mu)^K V^0,$$

where

$$V^k \;\;=\;\; \left\|z^k - z^*\right\|^2 + \frac{200\omega^2\gamma^2}{M}\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 + 40\omega\gamma^2\left\|F(z^{k+\frac{1}{2}}) - h^k\right\|.$$

Corollary 2.6. *Suppose the conditions of Theorem 2.5 hold. Then Algorithm 1 has*

$$\mathcal{O}\left(\left(\frac{\omega}{q_\omega} + \frac{\omega L}{\mu q_\omega} + \frac{\omega^{3/2}\hat{L}}{\mu q_\omega\sqrt{M}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

communication complexity, where q_ω is the expected density and ε represents the accuracy of the solution.

Remark 2.7. One can note, that for practical compressors (Beznosikov et al., 2023) it implies $q_\omega \geq \omega$. In that way, our estimate can be written as

$$\widetilde{\mathcal{O}}\left(1 + \frac{L}{\mu} + \frac{\omega^{1/2}\hat{L}}{\mu\sqrt{M}}\right).$$

We obtained a convergence estimate without requiring full gradient transmission, significantly enhancing the applicability of compression methods for VIs. Although the estimate includes a non-optimal $\omega^{1/2}$ term in the numerator, we do not consider it critical, as the $\sqrt{M}$ term in the denominator dominates the convergence behavior.

Now we proceed with our second algorithm, which handles contractive compressors.

2.2 EXTRAGRADIENT DIANA FOR CONTRACTIVE COMPRESSORS

The analysis of distributed EXTRAGRADIENT with contractive compressors (see Definition 1.5) raises specific challenges. There arises a dot product of the following form $\langle g^k, z^{k+\frac{1}{2}} - z^*\rangle$, where g^k is the full gradient estimator. In this way, we cannot address it straightforwardly using the variance reduction technique DIANA, due to the compressor's bias and non-fulfillment of $\mathbb{E}_Q\left[\left\langle g^k, z^{k+\frac{1}{2}} - z^*\right\rangle\right] = \left\langle F\left(z^{k+\frac{1}{2}}\right), z^{k+\frac{1}{2}} - z^*\right\rangle$. To adapt our method to this setting, one can apply the EF21 approach; however, estimating a dot product of the form $\langle F(z^{k+\frac{1}{2}}) - g^{k-1}, z^{k+\frac{1}{2}} - z^*\rangle$, complicates the analysis and does not offer the prospect of an optimal final convergence rate.

Therefore, we employ a more standard method in our approach, specifically the ERROR FEEDBACK scheme. In this section, we present the EXTRAGRADIENT DIANA WITH ERROR FEEDBACK method (Algorithm 2).

Let us focus on several details. We apply the classical ERROR FEEDBACK scheme, i.e., add the error compensation term to the compressed vector (Lines 9,10). It is important to note that such compensation is necessary only for the full gradient estimators g^k that directly perform the main step of the method. For updating the auxiliary sequence h^k, this technique is redundant. In this way, we additionally compress $\left(F_i(z^{k+1/2}) - h_i^k\right)$ (Lines 11, 12). Such a modification leads to broadcasting both $\hat{\Delta}_i^k$ and $\widetilde{\Delta}_i^k$ from devices to the server at each iteration to update g^k and h^k (Lines 17, 19). Mention it is not critical, since the communication complexity at each iteration is still $\mathcal{O}\left({}^{1}/_{q}\right)$.

Now we move to the theoretical convergence results of Algorithm 2.

Theorem 2.8. *Suppose Assumptions 2.1, 2.3 hold. Then for Algorithm 2 with*

$$\gamma \leq \min\left\{\frac{\alpha}{8\mu}, \frac{\alpha^2}{16\sqrt{1772}\hat{L}}\right\}$$

and $\beta = 1 - \frac{\alpha}{8}$ the following estimate is fulfilled:

$$\mathbb{E}V^K \leq (1-\gamma\mu)^K V^0,$$

where

$$\begin{aligned} V^k \quad = \quad & \left\|z^k - z^*\right\|^2 \\ & + \frac{7088\gamma^2}{\alpha^3}\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 \\ & + \frac{276\gamma^2}{\alpha}\frac{1}{M}\sum_{i=1}^{M}\left\|e_i^k\right\|. \end{aligned}$$

Corollary 2.9. *Suppose the conditions of Theorem 2.8 hold. Then Algorithm 2 has*

$$\mathcal{O}\left(\left(\frac{1}{q_\alpha \alpha} + \frac{\hat{L}}{\mu\alpha^2 q_\alpha}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

communication complexity, where q_α is the expected density and ε represents accuracy of the solution.

Remark 2.10. One can note, that for practical compressors (Beznosikov et al., 2023) it implies $q_\alpha \geq \frac{1}{\alpha}$. In that way, our estimate can be written as

$$\widetilde{\mathcal{O}}\left(1 + \frac{\hat{L}}{\mu\alpha}\right).$$

Algorithm 2 EXTRAGRADIENT DIANA WITH ERROR FEEDBACK

```
Input: initial point $z^0 \in Z$, initial vectors $\{h_i^0\}_{i=1}^M, \{e_i^0\}_{i=1}^M$, each $h_i^0, e_i^0 \in Z$, $h^0 = \frac{1}{M}\sum_{i=1}^M h_i^0$, compressor $Q$ is under Definition 1.5, number or workers $M$, number of iterations $K$
Parameters: Stepsize $\gamma > 0$, bias $0 < \beta < 1$
for $k = 0, 1, 2, ..., K-1$ do
    $z^{k+\frac{1}{2}} = z^k - \gamma h^k$
    The server broadcast $z^{k+\frac{1}{2}}$ to all workers
    for $i = 1, 2, \ldots, M$ in parallel do
        Compute $F_i(z^{k+\frac{1}{2}})$
        $\Delta_i^k = F_i(z^{k+\frac{1}{2}}) - h_i^k$
        Compress $\hat{\Delta}_i^k \sim Q(\Delta_i^k + e_i^k)$
        $e_i^{k+1} = e_i^k + \left(\Delta_i^k - \hat{\Delta}_i^k\right)$
        Compress $\widetilde{\Delta}_i^k \sim Q(\Delta_i^k)$
        Grad correction $h_i^{k+1} = h_i^k + \beta\widetilde{\Delta}_i^k$
        Broadcast $\hat{\Delta}_i^k, \widetilde{\Delta}_i^k$ to the server
    end for
    $\hat{\Delta}^k = \frac{1}{M}\sum_{i=1}^M \hat{\Delta}_i^k$
    $\widetilde{\Delta}^k = \frac{1}{M}\sum\limits_{i=1}^M \widetilde{\Delta}_i^k$
    $g^k = h^k + \hat{\Delta}^k$
    $z^{k+1} = z^k - \gamma g^k$
    $h^{k+1} = h^k + \beta\widetilde{\Delta}^k$
end for
Output: $z^K$
```

Comparing this rate with the result of Theorem 2.5 one can note that a deterioration occurs in $1/\alpha$ times. A similar effect appeared in the work (Beznosikov et al., 2022) when matching MASHA1 and MASHA2.

Providing a final discussion of the theoretical results, we note that our approach yields a worse estimate than the MASHA1 method (see Table 1). Comparing with OPTIMISTIC MASHA (Beznosikov & Gasnikov, 2022) and THREE PILLARS ALGORITHM (Beznosikov et al., 2024), we observe that these methods exploit the data similarity assumption. Without it, their estimates revert to those of the MASHA1 method. However, all these methods utilize a momentum parameter, as in the work (Alacaoglu & Malitsky, 2022) and require the transmission of the full gradients. In this regard, our method represents a significant improvement.

3 EXPERIMENTS

In this section, we present an empirical study of our methods, focusing on the image classification problem and generative network training. Our code is open-sourced[1].

3.1 CLASSIFICATION PROBLEM

Setup. The first part of the numerical experiments is conducted on the CIFAR-10 dataset Krizhevsky et al. (2009), which is widely used as a benchmark in the optimization community, consisting of 50,000 training and 10,000 test samples. Each sample is an 32×32 RGB image associated with one of ten class labels. The experiments are implemented in Python using the PyTorch library Paszke et al. (2017), leveraging both a single CPU (Intel Xeon 2.20 GHz) and a single GPU (NVIDIA Tesla P100) for computation. To emulate a distributed environment, we split batches

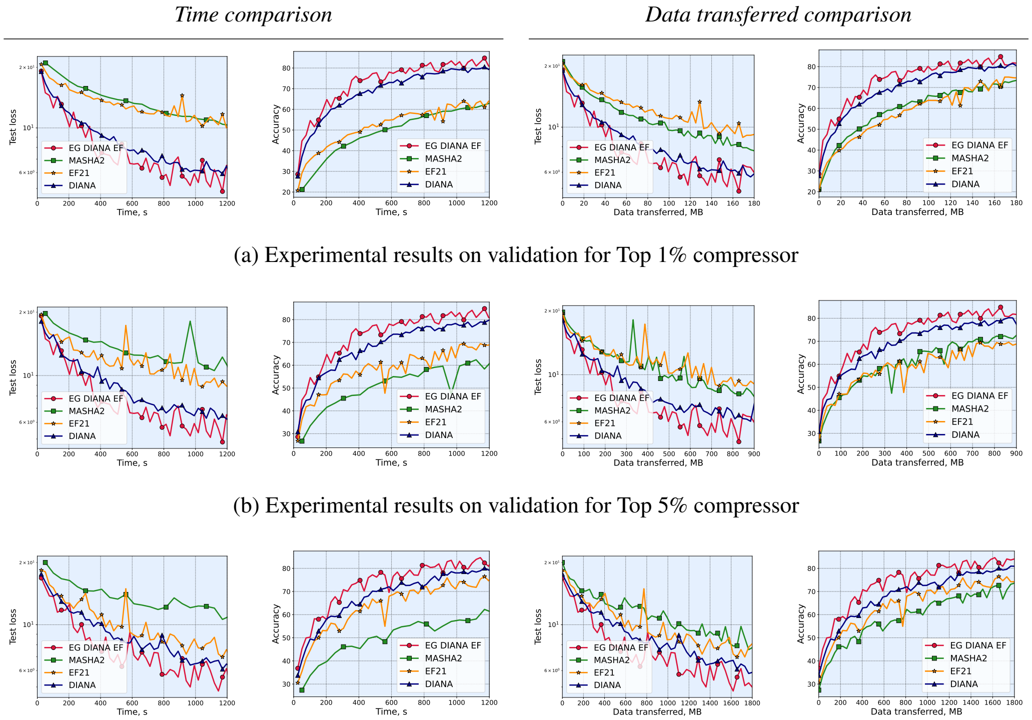
[1] `https://anonymous.4open.science/r/EG-DIANA-exp`

across multiple workers, simulating a decentralized optimization setting. The total runtime for all experiments is approximately 40 hours.

We address a classification task on the CIFAR-10 dataset using the ResNet18 architecture Meng et al. (2019), a standard benchmark model for evaluating optimization algorithms due to its balance of complexity and performance. To explore the robustness of the optimizers, we reformulate the standard minimization problem into a min-max optimization framework as described in equation 6. Experimental details are available in Appendix A.

3.2 Results and Discussion

Performance Comparison. We compare four algorithms: Extragradient DIANA with Error Feedback, MASHA2, EF21 and DIANA with Top$n\%$ compressor. Figure 1 represents the convergence of methods at compression levels $1\%, 5\%, 10\%$, respectively. To facilitate comprehension, all graphs have been aligned according to the comparison and compression factors. We evaluate the convergence rate in terms of time and the total amount of information exchanged between workers, providing a pairwise comparison of the methods. The results are presented in the left and the right columns of Figure 1, demonstrating the superior convergence behavior regarding communication efficiency for our approach. We employ an unconventional time metric for convergence comparison to emphasize a critical aspect: the SOTA method MASHA2 requires full gradient computation and transmission, which effectively doubles the actual wall-clock time needed for model training compared to methods that rely on stochastic gradient estimates. This distinction is crucial for understanding the practical efficiency of the algorithms, as it highlights the computational overhead associated with MASHA2's approach.



(a) Experimental results on validation for Top 1% compressor

(b) Experimental results on validation for Top 5% compressor

(c) Experimental results on validation for Top 10% compressor

Figure 1: Performance comparison of ResNet18 performance on CIFAR-10 with respect to the training time and amount of data transferred. Results for each type of compression value are presented in the rows.

Discussion. As anticipated by the theoretical analysis, MASHA2 demonstrates strong communication efficiency due to its use of full gradient information. However, in practice, Algorithm 2 achieves comparable and often superior performance in terms of both test accuracy and training speed, despite not relying on full gradient transmission. Moreover, our experimental results show that Algorithm 1 consistently outperforms distributed optimization methods such as EF21 and DIANA. This confirms the practical effectiveness of our approach.

3.3 GAN TRAINING

In this section, we present the experiments conducted using the GAN architecture, for which the loss functions are naturally $\min\max$ problems. Thus, our variational inequality approach serves as a common method to solve them. We implement STYLEGAN (introduced by Karras (2019)): a generative adversarial network that enables the generation of high-quality images while controlling the style of the generated images at various levels of detail.

Setup. Our experiments were conducted using the basic STYLEGAN implementation (. The model was trained on the I'M SOMETHING OF A PAINTER MYSELF dataset, which comprises 300 Monet paintings sized 256×256 and 7028 photos in the same format. The training process took place on the same hardware as in the previous section. We train the generator (specifically, the parameters of 2 generators for our architecture) with an extrapolation step (see Line 9) and subsequently make the «whole» step (see Line 18) for both discriminators. Note that, to maintain a variational inequality representation of the loss, an equal number of steps were taken for both generators and discriminators.

Optimizers and Experimental Design. Training a GAN is highly computationally expensive, thus we assumed in our experiments $M = 5$ number of clients and trained the model for only 160 epochs (however, this did not have a significant impact on the final results). Once again, we choose MASHA2 and EXTRAGRADIENT DIANA WITH ERROR FEEDBACK as optimizers, and TOP10% as a compressor. They demonstrated the best performance in terms of time in the previous section and exhibited nearly the same behavior regarding transmitted information. To maintain fairness, we set the following hyperparameters for both optimizers: i) learning rate: $5 \cdot 10^{-4}$; ii) exponential moving average (EMA) decay rate 0.99.

3.4 RESULTS AND DISCUSSION

Firstly, we present the generators and discriminators loss function graphs, as well as the image generation results for the best model in Figure 2. Secondly, we evaluate the models using the Fréchet Inception Distance (FID) metric (lower - better): i) **10.32** for EXTRAGRADIENT DIANA WITH ERROR FEEDBACK; ii) **24.47** for MASHA2.

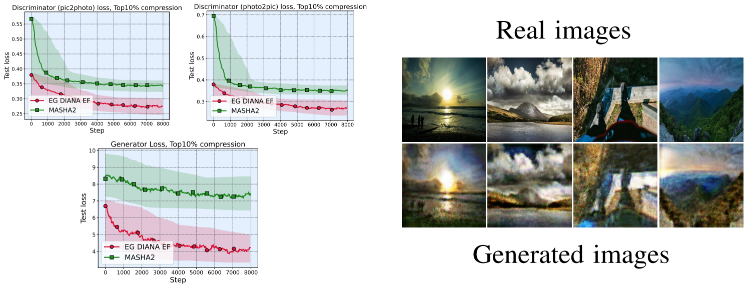


Figure 2: Comparison of train losses for generator (averaged generators loss) and discriminators for STYLEGAN architecture under TOP10% compression (left half) and obtained results for generation on some examples pictures (right half).

Performance comparison. On Figure 2 we observe that Algorithm 2 exhibits more stable training dynamics compared to MASHA2. The generator and discriminator losses for DIANA converge more smoothly, indicating a balanced training process. In contrast, MASHA2 fluctuates around a local optimum due to its full gradient updates (see generator graph). Moreover, it can be observed from the FID metric evaluation that the STYLEGAN with full gradient updates does not achieve the desired quality of restored images.

Discussion. Despite the fact that I'M SOMETHING OF A PAINTER MYSELF... is 7 times smaller than CIFAR-10 dataset in terms of the number of classes, it comprises images of significantly higher resolution (by a factor of 16), making it an excellent benchmark for evaluating optimizer performance on more complex, high-dimensional data.
GAN architectures are well-known for their highly complex loss function landscapes, which makes them especially sensitive updates based on of full gradients. Although the error-feedback mechanism helps MASHA2 cope with compressed gradients, the variance reduction scheme inevitably drives the algorithm into local optima, making it inapplicable for real-world applications. In contrast, proposed Algorithm 2 eliminates the need for full gradient computation entirely. This not only leads to more stable training dynamics but also enables the achievement of superior metric values, demonstrating its effectiveness in handling the challenging optimization landscape inherent to GANs.

REFERENCES

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816. PMLR, 2022.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

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

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28, 2015.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Aleksandr Beznosikov and Alexander Gasnikov. Compression and data similarity: Combination of two techniques for communication-efficient solving of distributed variational inequalities. In *International Conference on Optimization and Applications*, pp. 151–162. Springer, 2022.

Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.

Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov. Distributed saddle-point problems under data similarity. *Advances in Neural Information Processing Systems*, 34:8172–8184, 2021.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. *Advances in Neural Information Processing Systems*, 35:14013–14029, 2022.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Aleksandr Beznosikov, Martin Takác, and Alexander Gasnikov. Similarity, compression and local steps: three pillars of efficient communications for distributed variational inequalities. *Advances in Neural Information Processing Systems*, 36, 2024.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pp. 1387–1395. PMLR, 2021.

Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 3573–3579. IEEE, 2020.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. Marina: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pp. 3788–3798. PMLR, 2021a.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 3556–3564. PMLR, 2021b.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 7865–7901. PMLR, 2022.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Peter Richtárik, and Sebastian Stich. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, 38(1):91–106, 2023.

Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

Yujia Jin and Aaron Sidford. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pp. 4890–4900. PMLR, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Jakub Konečnỳ and Peter Richtárik. Randomized distributed mean estimation: Accuracy vs. communication. *Frontiers in Applied Mathematics and Statistics*, 4:62, 2018.

Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Distributed algorithms for aggregative games on graphs. *Operations Research*, 64(3):680–704, 2016.

Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. *Advances in Neural Information Processing Systems*, 35:33494–33507, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915. PMLR, 2019.

Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070, 2020.

Ruichen Luo, Sebastian U Stich, Samuel Horváth, and Martin Takáč. Revisiting localsgd and scaffold: Improved rates and missing analysis. *arXiv preprint arXiv:2501.04443*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yu Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

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

Daniil Medyakov, Gleb Molodtsov, Evseev Grigoriy, Egor Petrov, and Aleksandr Beznosikov. Shuffling heuristic in variational inequalities: Establishing new convergence guarantees. In *International Conference on Computational Optimization*, 2024.

Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pp. 3866–3870. IEEE, 2019.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, pp. 1–16, 2024.

Soham Mukherjee and Mrityunjoy Chakraborty. A decentralized algorithm for large scale min-max problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2967–2972. IEEE, 2020.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pp. 2681–2690. PMLR, 2017.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34: 4384–4396, 2021.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, Dmitry Kovalev, Pavel Dvurechensky, and Alexander Gasnikov. Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*, 2021.

Gesualdo Scutari, Francisco Facchinei, Peiran Song, Daniel P Palomar, and Jong-Shi Pang. Decomposition by partial linearization: Parallel optimization of multi-agent systems. *IEEE Transactions on Signal Processing*, 62(3):641–656, 2013.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pp. 1058–1062. Singapore, 2014.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR, 2014.

Kunal Srivastava, Angelia Nedić, and Dušan Stipanović. Distributed min-max optimization in networks. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–8. IEEE, 2011.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

Nikko Ström. Scalable distributed dnn training using commodity gpu cloud computing. 2015.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

Alexander Tyurin and Peter Richtárik. Dasha: Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *arXiv preprint arXiv:2202.01268*, 2022.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2): 1–33, 2020.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior: by J. Von Neumann and O. Morgenstern*. Princeton university press, 1953.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. *Advances in neural information processing systems*, 17, 2004.

Deming Yuan, Qian Ma, and Zhen Wang. Dual averaging method for solving multi-agent saddle-point problems with quantized information. *Transactions of the Institute of Measurement and Control*, 36(1):38–46, 2014.

A EXPERIMENTAL DETAILS

The implementation of the stochastic versions of extragradient-based optimization algorithms discussed in this work is publicly available in an anonymized repository. The full source code can be accessed by link: `https://anonymous.4open.science/r/EG-DIANA-exp`.

A.1 CLASSIFICATION PROBLEM

We compare two algorithms: EXTRAGRADIENT DIANA WITH ERROR FEEDBACK and MASHA WITH ERROR FEEDBACK. Both algorithms are evaluated in terms of information transferred (communication efficiency) and training time. The experiments are conducted with the following setup:

- number of workers (M): 10;
- learning rate: 0.01 for both optimizers;
- regularization parameters $\lambda_1 = \lambda_2 = 0.0005$.

Compressor. Due to different sizes of inputs, we implement TOP$k\%$, which retains the largest $k\%$ values (in absolute) of a tensor. For MASHA WITH ERROR FEEDBACK, the server compressor is set to the identity one to ensure a fair comparison with EXTRAGRADIENT DIANA WITH ERROR FEEDBACK, which does not employ server compression.

Hyperparameter Tuning To ensure a fair comparison, we conducted a grid search for key hyperparameters:

- for EXTRAGRADIENT DIANA WITH ERROR FEEDBACK, we tuned β over the range $[0, 1]$;
- for MASHA WITH ERROR FEEDBACK, we tuned τ over a similar range.

The optimal values were found to be $\beta = 0.5$ for EXTRAGRADIENT DIANA WITH ERROR FEEDBACK and $\tau = 0.1$ for MASHA WITH ERROR FEEDBACK. These values were selected based on methods ability to achieve the best convergence and stability during training.

B NOTATION AND GENERAL INEQUALITIES

• **Notation.** There we present notation used in the work: $\mathbb{E}\left[\xi\right]$ denotes the expected value of a random vector $\xi \in \mathbb{R}^d$, $\mathbb{E}_\eta\left[\xi\right]$ refers to the expected value of a random vector $\xi \in \mathbb{R}^d$ conditioned on a random vector $\eta \in \mathbb{R}^d$, $\|x\| = \|x\|_2 = \sqrt{\langle x, x\rangle}$ represents the Euclidean norm of the vector $x \in \mathbb{R}^d$.

• **General inequalities.** We introduce important inequalities that are used in further proofs. Let $a_1, a_2 \in \mathbb{R}^d$, then

$$\|a_1 + a_2\|^2 \leqslant (1+s)\|a_1\|^2 + (1+s^{-1})\|a_2\|^2\,, \tag{Jensen}$$

$$2\langle a_1, b_1\rangle \leqslant \|a_1\|^2\|b_1\|^2\,. \tag{CB}$$

C PROOF OF THEOREM 2.5

Theorem 2.5 *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then for Algorithm 1 with* $\gamma \leq \min\left\{\frac{1}{4\omega\mu}, \frac{1}{8\sqrt{30}\omega L}, \frac{\sqrt{M}}{40\sqrt{6}\cdot\omega^{3/2}\sqrt{\frac{1}{M}\sum\limits_{i=1}^{M} L_i^2}}\right\}$ *and* $\beta = \frac{1}{\omega}$ *the following estimate is fulfilled:*

$$\mathbb{E}V^K \leq (1-\gamma\mu)^K V^0,$$

where $V^k = \left\|z^k - z^*\right\|^2 + \frac{200\omega^2\gamma^2}{M}\frac{1}{M}\sum\limits_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 + 40\omega\gamma^2\left\|F(z^{k+\frac{1}{2}}) - h^k\right\|$.

Proof. We start with the classicalal for EXTRAGRADIENT proofs equation for arbitrary z, v and u:

$$\|z-u\|^2 = \|v-u\|^2 + 2\langle z-v, z-u\rangle - \|z-v\|^2\,.$$

For $z = z^{k+\frac{1}{2}}$, $u = z^*$ and $v = z^k$:

$$\left\|z^{k+\frac{1}{2}} - z^*\right\|^2 = \left\|z^k - z^*\right\|^2 + 2\left\langle z^{k+\frac{1}{2}} - z^k, z^{k+\frac{1}{2}} - z^*\right\rangle - \left\|z^{k+\frac{1}{2}} - z^k\right\|^2.$$

Using the same for $z = z^{k+1}$ with $u = z^*$ and $v = z^{k+\frac{1}{2}}$,

$$\left\|z^{k+1} - z^*\right\|^2 = \left\|z^{k+\frac{1}{2}} - z^*\right\|^2 + 2\left\langle z^{k+1} - z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}} - z^*\right\rangle + \left\|z^{k+\frac{1}{2}} - z^{k+1}\right\|^2.$$

After the summation with the previous equality, we obtain

$$\begin{aligned}\left\|z^{k+1} - z^*\right\|^2 &= \left\|z^k - z^*\right\|^2 + 2\left\langle z^{k+1} - z^k, z^{k+\frac{1}{2}} - z^*\right\rangle \\ &\quad + \left\|z^{k+\frac{1}{2}} - z^{k+1}\right\|^2 - \left\|z^{k+\frac{1}{2}} - z^k\right\|^2. \end{aligned} \tag{7}$$

Let's work with the term $\left\|z^{k+1} - z^{k+\frac{1}{2}}\right\|^2$:

$$\begin{aligned}\left\|z^{k+1} - z^{k+\frac{1}{2}}\right\|^2 &= \gamma^2 \left\|g^k - h^k\right\|^2 \\ &= \gamma^2 \left\|\frac{1}{M}\sum_{i=1}^{M} Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right)\right\|^2 \\ &= \gamma^2 \frac{1}{M^2}\sum_{i=1}^{M}\left\|Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right)\right\|^2 \\ &\quad + \gamma^2 \frac{1}{M^2}\sum_{i\neq j}^{M}\left\langle Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right), Q\left(F_j(z^{k+\frac{1}{2}}) - h_j^k\right)\right\rangle \\ &= \gamma^2 \frac{\omega}{M^2}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 \\ &\quad + \gamma^2 \frac{1}{M^2}\sum_{i\neq j}^{M}\left\langle Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right), Q\left(F_j(z^{k+\frac{1}{2}}) - h_j^k\right)\right\rangle. \end{aligned} \tag{8}$$

Due to unbiasdness of the operator Q we can take the expectation with respect to Q and get

$$\begin{aligned} &\mathbb{E}_Q\left[\sum_{i\neq j}^{M}\left\langle Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right), Q\left(F_j(z^{k+\frac{1}{2}}) - h_j^k\right)\right\rangle\right] \\ &= \sum_{i\neq j}^{M}\left\langle F_i(z^{k+\frac{1}{2}}) - h_i^k, F_j(z^{k+\frac{1}{2}}) - h_j^k\right\rangle \\ &= \left\|\sum_{i=1}^{M}\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right)\right\|^2 - \sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 \\ &= M^2\left\|F(z^{k+\frac{1}{2}}) - h^k\right\|^2 - \sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2. \end{aligned} \tag{9}$$

Now we insert this result into equation 8 and simplify:

$$\begin{aligned}\mathbb{E}\left\|z^{k+\frac{1}{2}} - z^{k+1}\right\|^2 &\leq \gamma^2 \frac{\omega}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 \\ &\quad + \gamma^2\mathbb{E}\left\|F(z^{k+\frac{1}{2}}) - h^k\right\|^2 - \frac{\gamma^2}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 \\ &= \gamma^2\frac{\omega - 1}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 + \gamma^2\mathbb{E}\left\|F(z^{k+\frac{1}{2}}) - h^k\right\|^2. \end{aligned} \tag{10}$$

Now we work with dot product term in equation 7 and take the expectation with respect to Q:

$$\begin{aligned} &2\mathbb{E}_Q\left[\left\langle z^{k+1} - z^k, z^{k+\frac{1}{2}} - z^*\right\rangle\right] \\ &= \quad -2\gamma\mathbb{E}_Q\left[\left\langle g^k, z^{k+\frac{1}{2}} - z^*\right\rangle\right] \end{aligned}$$

$$
\begin{aligned}
&= -2\gamma \mathbb{E}_Q \left[\left\langle h^k + \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right), z^{k+\frac{1}{2}} - z^* \right\rangle \right] \\
&= -2\gamma \left\langle F(z^{k+\frac{1}{2}}), z^{k+\frac{1}{2}} - z^* \right\rangle \\
&= -2\gamma \left\langle F(z^{k+\frac{1}{2}}) - F(z^*), z^{k+\frac{1}{2}} - z^* \right\rangle - 2\gamma \underbrace{\left\langle F(z^*), z^{k+\frac{1}{2}} - z^* \right\rangle}_{\leq 0} \\
&\overset{\text{Ass. 2.3}}{\leq} -2\gamma\mu \left\| z^{k+\frac{1}{2}} - z^* \right\|^2 .
\end{aligned}
$$

Substituting this result with equation 10 into equation 7 we get

$$
\begin{aligned}
\mathbb{E}\left\| z^{k+1} - z^* \right\|^2 &= \mathbb{E}\left\| z^k - z^* \right\|^2 - 2\gamma\mu\mathbb{E}\left\| z^{k+\frac{1}{2}} - z^* \right\|^2 \\
&\quad + \gamma^2 \frac{\omega - 1}{M^2} \sum_{i=1}^M \mathbb{E}\left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 + \gamma^2 \mathbb{E}\left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 \\
&\quad - \mathbb{E}\left\| z^{k+\frac{1}{2}} - z^k \right\|^2 \\
&\overset{CB}{\leq} (1 - \gamma\mu)\mathbb{E}\left\| z^k - z^* \right\|^2 + \gamma^2 \frac{\omega - 1}{M^2} \sum_{i=1}^M \mathbb{E}\left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
&\quad + \gamma^2 \mathbb{E}\left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 - (1 - \gamma\mu)\mathbb{E}\left\| z^{k+\frac{1}{2}} - z^k \right\|^2 . \qquad (11)
\end{aligned}
$$

Firstly, we derive a recursion on the $\left\| F(z^{k+\frac{3}{2}}) - h^{k+1} \right\|^2$:

$$
\begin{aligned}
\left\| F(z^{k+\frac{3}{2}}) - h^{k+1} \right\|^2 &= \left\| F(z^{k+\frac{3}{2}}) - h^k - \beta \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right) \right\|^2 \\
&\overset{Jensen}{\leq} (1 + c_1^{-1}) \left\| F(z^{k+\frac{3}{2}}) - F(z^{k+\frac{1}{2}}) \right\|^2 \\
&\quad + (1 + c_1) \left\| F(z^{k+\frac{1}{2}}) - h^k - \beta \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right) \right\|^2 \\
&\overset{\text{Ass. 2.2}}{\leq} (1 + c_1^{-1}) L^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
&\quad + (1 + c_1) \left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 \\
&\quad + (1 + c_1)\beta^2 \left\| \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right) \right\|^2 \\
&\quad - 2(1 + c_1)\beta \left\langle F(z^{k+\frac{1}{2}}) - h^k, \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right) \right\rangle \\
&\overset{\substack{\text{Def. 1.3} \\ (i)}}{\leq} (1 + c_1^{-1}) L^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
&\quad + (1 + c_1)(1 - 2\beta) \left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 \\
&\quad + (1 + c_1)\beta^2 \left\| \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right) \right\|^2 \\
&\overset{8,9}{\leq} (1 + c_1^{-1}) L^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
&\quad + (1 + c_1)(1 - 2\beta) \left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
& +(1+c_1)\beta^2 \frac{\omega-1}{M^2} \sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& +(1+c_1)\beta^2 \left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 \\
= \quad & (1+c_1^{-1})L^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
& +(1+c_1)(1-\beta)^2 \left\| F(z^{k+\frac{1}{2}}) - h^k \right\|^2 \\
& +(1+c_1)\beta^2 \frac{\omega-1}{M^2} \sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 ,
\end{aligned} \tag{12}
$$

where in (i) we take expectation $\mathbb{E}_Q$ with respect to Q. Secondly, we obtain a recursion on the $\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{3}{2}}) - h_i^{k+1} \right\|^2$:

$$
\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{3}{2}}) - h_i^{k+1} \right\|^2 \quad = \quad & \frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{3}{2}}) - h_i^k - \beta Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k \right) \right\|^2 \\
\overset{Jensen}{\leq} \quad & (1+c_2^{-1})\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{3}{2}}) - F_i(z^{k+\frac{1}{2}}) \right\|^2 \\
& +(1+c_2)\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right. \\
& \left. -\beta Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k \right) \right\|^2 \\
\overset{\text{Ass. 2.1}}{\leq} \quad & (1+c_2^{-1})\frac{1}{M}\sum_{i=1}^{M} L_i^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
& +(1+c_2)\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& +(1+c_2)\beta^2\frac{1}{M}\sum_{i=1}^{M} \left\| Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k \right) \right\|^2 \\
& -2(1+c_2)\beta \\
& \cdot\frac{1}{M}\sum_{i=1}^{M} \left\langle F_i(z^{k+\frac{1}{2}}) - h_i^k, Q(F_i(z^{k+\frac{1}{2}}) - h_i^k) \right\rangle \\
\overset{\substack{\text{Def. 1.3} \\ (i)}}{=} \quad & (1+c_2^{-1})\frac{1}{M}\sum_{i=1}^{M} L_i^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
& +(1+c_2)\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& +(1+c_2)\beta^2\omega\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& -2(1+c_2)\beta\frac{1}{M}\sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
= \quad & (1+c_2^{-1})\frac{1}{M}\sum_{i=1}^{M} L_i^2 \left\| z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
& +(1+c_2)\left[1 - 2\beta + \beta^2\omega \right]
\end{aligned}
$$

$$\cdot \frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2, \tag{13}$$

where in (*i*) we take expectation $\mathbb{E}_Q$ with respect to Q. Finally, we denote

$$\begin{aligned}
\mathcal{Z}^k &= \left\|z^k - z^*\right\|^2, \\
\mathcal{F}^k &= \frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2, \\
\mathcal{G}^k &= \left\|F(z^{k+\frac{1}{2}}) - h^k\right\|^2,
\end{aligned}$$

and sum up the estimates, obtained in equation 12, equation 13 (with coefficients $\frac{A\gamma^2}{M}$ and $B\gamma^2$, which we define later, respectively) with equation 11:

$$\begin{aligned}
&\mathbb{E}\mathcal{Z}^{k+1} + \frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{k+1} + B\gamma^2\mathbb{E}\mathcal{G}^{k+1} \\
&\leq (1-\gamma\mu)\mathbb{E}\mathcal{Z}^k + \gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k \\
&\quad +\gamma^2\mathbb{E}\mathcal{G}^k - (1-\gamma\mu)\mathbb{E}\left\|z^{k+\frac{1}{2}} - z^k\right\|^2 \\
&\quad +A\gamma^2(1+c_2^{-1})\frac{1}{M^2}\sum_{i=1}^{M}L_i^2\mathbb{E}\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2 \\
&\quad +A\gamma^2(1+c_2)\left[1-2\beta+\beta^2\omega\right]\frac{1}{M}\mathbb{E}\mathcal{F}^k \\
&\quad +B\gamma^2(1+c_1^{-1})L^2\mathbb{E}\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2 \\
&\quad +B\gamma^2(1+c_1)(1-\beta)^2\mathbb{E}\mathcal{G}^k \\
&\quad +B\gamma^2(1+c_1)\beta^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k \\
&= (1-\gamma\mu)\mathbb{E}\mathcal{Z}^k \\
&\quad +\gamma^2\left[\frac{1}{M^2}\sum_{i=1}^{M}L_i^2A(1+c_2^{-1}) + L^2B(1+c_1^{-1})\right]\mathbb{E}\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2 \\
&\quad +\gamma^2\left[(\omega-1) + A(1+c_2)\left[1-2\beta+\beta^2\omega\right] + B(1+c_1)\beta^2(\omega-1)\right]\frac{1}{M}\mathbb{E}\mathcal{F}^k \\
&\quad +\gamma^2\left[1 + B(1+c_1)(1-\beta)^2\right]\mathbb{E}\mathcal{G}^k \\
&\quad -(1-\gamma\mu)\gamma^2\mathbb{E}\left\|h^k\right\|^2.
\end{aligned} \tag{14}$$

Note that in the previous equation we used the fact, that $\left\|z^{k+\frac{1}{2}} - z^k\right\|^2 = \gamma^2\left\|h^k\right\|^2$. Next, consider the term $\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2$:

$$\begin{aligned}
\mathbb{E}\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2 &= \mathbb{E}\left\|(z^{k+\frac{3}{2}} - z^{k+1}) - (z^{k+\frac{1}{2}} - z^{k+1})\right\|^2 \\
&\overset{CB}{\leq} 2\mathbb{E}\left\|z^{k+1} - z^{k+\frac{1}{2}}\right\|^2 + 2\mathbb{E}\left\|z^{k+\frac{3}{2}} - z^{k+1}\right\|^2 \\
&= 2\gamma^2\mathbb{E}\left\|g^k - h^k\right\|^2 + 2\gamma^2\mathbb{E}\left\|h^{k+1}\right\|^2 \\
&\overset{10}{\leq} 2\gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k + 2\gamma^2\mathbb{E}\mathcal{G}^k \\
&\quad +2\gamma^2\mathbb{E}\left\|h^k + \frac{1}{M}\sum_{i=1}^{M}Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{CB}{\leq} && 2\gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k + 2\gamma^2\mathbb{E}\mathcal{G}^k + 4\gamma^2\mathbb{E}\left\|h^k\right\|^2 \\
& && +4\gamma^2\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}Q\left(F_i(z^{k+\frac{1}{2}})-h_i^k\right)\right\|^2 \\
&\overset{8,9}{\leq} && 2\gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k + 2\gamma^2\mathbb{E}\mathcal{G}^k \\
& && +4\gamma^2\mathbb{E}\left\|h^k\right\|^2 + 4\gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k + 4\gamma^2\mathbb{E}\mathcal{G}^k \\
&= && 6\gamma^2\frac{\omega-1}{M}\mathbb{E}\mathcal{F}^k + 6\gamma^2\mathbb{E}\mathcal{G}^k + 4\gamma^2\mathbb{E}\left\|h^k\right\|^2.
\end{aligned}
$$

We substitute this estimate into equation 14:

$$
\begin{aligned}
\mathbb{E}\mathcal{Z}^{k+1} + \frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{k+1} + B\gamma^2\mathbb{E}\mathcal{G}^{k+1} \quad \leq \quad & (1-\gamma\mu)\mathbb{E}\mathcal{Z}^k \\
& +\gamma^2\left[7(\omega-1) + A(1+c_2)\left[1-2\beta+\beta^2\omega\right]\right. \\
& \left.+B(1+c_1)\beta^2(\omega-1)\right]\frac{1}{M}\mathbb{E}\mathcal{F}^k \\
& +\gamma^2\left[7 + B(1+c_1)(1-\beta)^2\right]\mathbb{E}\mathcal{G}^k \\
& +\gamma^2\left[4\gamma^2\left[\frac{1}{M^2}\sum_{i=1}^{M}L_i^2A(1+c_2^{-1}) + L^2B(1+c_1^{-1})\right]\right. \\
& \left.-(1-\gamma\mu)\right]\mathbb{E}\left\|h^k\right\|^2. \qquad (15)
\end{aligned}
$$

Due to the arbitrary choice for parameters c_1 and c_2, we can set them equal to $\frac{1}{2\omega}$, and $\beta = \frac{1}{\omega}$, we instantly derive that

$$
\begin{aligned}
7(\omega-1) + A(1+c_2)\left[1-2\beta+\beta^2\omega\right] + B(1+c_1)\beta^2(\omega-1) \quad &= \quad 7(\omega-1) \\
&+ \quad A\left(1+\frac{1}{2\omega}\right)\left(1-\frac{1}{\omega}\right) \\
&+ \quad B\left(1+\frac{1}{2\omega}\right)\left(1-\frac{1}{\omega}\right) \\
&\leq \quad 7(\omega-1) \\
&+ \quad A\left(1-\frac{1}{2\omega}\right) \\
&+ \quad B\left(1-\frac{1}{2\omega}\right),
\end{aligned}
$$

and also

$$
\begin{aligned}
7 + B(1+c_1)(1-\beta)^2 \quad &= \quad 7 + B\left(1+\frac{1}{2\omega}\right)\left(1-\frac{1}{\omega}\right)^2 \\
&\leq \quad 7 + B\left(1-\frac{1}{2\omega}\right).
\end{aligned}
$$

Summing up equation 15 with $p^k = (1-\gamma\mu)^{-k}$ and substituting the obtained estimations,

$$
\begin{aligned}
\sum_{k=0}^{K-1}p^k\mathbb{E}\mathcal{Z}^{k+1} + \frac{A\gamma^2}{M}\sum_{k=0}^{K-1}p^k\mathbb{E}\mathcal{F}^{k+1} \quad + \quad & B\gamma^2\sum_{k=0}^{K-1}p^k\mathbb{E}\mathcal{G}^{k+1} \\
\leq \quad & \sum_{k=0}^{K-1}p^k(1-\gamma\mu)\mathbb{E}\mathcal{Z}^k \\
& +\gamma^2\left[7(\omega-1) + A\left(1-\frac{1}{2\omega}\right) + B\left(1-\frac{1}{2\omega}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\cdot\frac{1}{M}\sum_{k=0}^{K-1} p^k \mathbb{E}\mathcal{F}^k \\
&+\gamma^2\left[7+B\left(1-\frac{1}{2\omega}\right)\right] \\
&\cdot\sum_{k=0}^{K-1} p^k \mathbb{E}\mathcal{G}^k \\
&+\gamma^2\left[4\gamma^2\left[\frac{1}{M^2}\sum_{i=1}^{M} L_i^2 A(1+2\omega)+L^2 B(1+2\omega)\right]\right. \\
&\left.-(1-\gamma\mu)\right]\sum_{k=0}^{K-1} p^k \mathbb{E}\left\|h^k\right\|^2 .
\end{aligned}
$$

Next, we derive the following estimates with $p^{-1}=1-\gamma\mu$:

$$
\begin{aligned}
p\left[7(\omega-1)+A\left(1-\frac{1}{2\omega}\right)+B\left(1-\frac{1}{2\omega}\right)\right] &= \frac{7(\omega-1)+A\left(1-\frac{1}{2\omega}\right)+B\left(1-\frac{1}{2\omega}\right)}{1-\gamma\mu} \\
&\leq \frac{7(\omega-1)+A\left(1-\frac{1}{2\omega}\right)+B\left(1-\frac{1}{2\omega}\right)}{1-\frac{1}{4\omega}} \\
&\leq 10(\omega-1)+A\left(1-\frac{1}{4\omega}\right)+B\left(1-\frac{1}{4\omega}\right).
\end{aligned}
$$

In order to fulfill the last inequality, we choose $\gamma \leq \frac{1}{4\omega\mu}$. One can also note, that

$$
\begin{aligned}
p\left[7+B\left(1-\frac{1}{2\omega}\right)\right] &= \frac{7+B\left(1-\frac{1}{2\omega}\right)}{1-\gamma\mu} \\
&\leq \frac{7+B\left(1-\frac{1}{2\omega}\right)}{1-\frac{1}{4\omega}} \\
&\leq 10+B\left(1-\frac{1}{4\omega}\right).
\end{aligned}
$$

We choose A and B to satisfy

$$
\begin{cases}
A \geq 10(\omega-1)+A\left(1-\frac{1}{4\omega}\right)+B\left(1-\frac{1}{4\omega}\right), \\
B \geq 10+B\left(1-\frac{1}{4\omega}\right),
\end{cases}
$$

$$
\begin{cases}
A \geq 200\omega^2, \\
B \geq 40\omega.
\end{cases}
$$

If we choose

$$
\gamma \leq \min\left\{\frac{1}{4\omega\mu}, \frac{1}{8\sqrt{30}\omega L}, \frac{\sqrt{M}}{40\sqrt{6}\cdot\omega^{3/2}\sqrt{\frac{1}{M}\sum\limits_{i=1}^{M} L_i^2}}\right\},
$$

then it is obvious, that

$$
4\gamma^2\frac{1}{M^2}\sum_{i=1}^{M} L_i^2 A(1+2\omega)+4\gamma^2 L^2 B(1+2\omega)-(1-\gamma\mu) \leq \frac{1}{2}-(1-\gamma\mu)<0.
$$

Lastly, we get

$$
\begin{aligned}
\sum_{k=0}^{K-1} p^k \mathbb{E}\mathcal{Z}^{k+1}+\sum_{k=0}^{K-1} p^k \frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{k+1}+\sum_{k=0}^{K-1} p^k B\gamma^2\mathbb{E}\mathcal{G}^{k+1} &\leq \sum_{k=0}^{K-1} p^k(1-\gamma\mu)\mathbb{E}\mathcal{Z}^k \\
&+\sum_{k=0}^{K-1} p^{k-1}\frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^k
\end{aligned}
$$

$$+\sum_{k=0}^{K-1} p^{k-1}\gamma^2 B\mathbb{E}\mathcal{G}^k.$$

Using the choice of p we derive the final estimate:

$$\begin{aligned}
\sum_{k=0}^{K-1}(1-\gamma\mu)^{-k}\mathbb{E}\mathcal{Z}^{k+1} + \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k}\frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{k+1} &+ \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k}B\gamma^2\mathbb{E}\mathcal{G}^{k+1}\\
&\leq \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}\mathbb{E}\mathcal{Z}^{k}\\
&\quad+\sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}\frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{k}\\
&\quad+\sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}B\gamma^2\mathbb{E}\mathcal{G}^{k};\\
(1-\gamma\mu)^{-K+1}\mathbb{E}\mathcal{Z}^{K} + (1-\gamma\mu)^{-K+1}\frac{A\gamma^2}{M}\mathbb{E}\mathcal{F}^{K} &+ (1-\gamma\mu)^{-K+1}B\gamma^2\mathbb{E}\mathcal{G}^{K}\\
&\leq (1-\gamma\mu)\mathcal{Z}^0\\
&\quad+(1-\gamma\mu)\gamma^2\frac{A\gamma^2}{M}\mathcal{F}^0\\
&\quad+(1-\gamma\mu)\gamma^2\mathcal{G}^0;\\
\mathbb{E}V^K &\leq (1-\gamma\mu)^K\mathbb{E}V^0,
\end{aligned}$$

where we denote $V^k = \mathcal{Z}^k + \frac{A\gamma^2}{M}\mathcal{F}^k + B\gamma^2\mathcal{G}^k$. This ends the proof of the theorem. □

Corollary 2.6 *Suppose the conditions of Theorem 2.5 holds. Then Algorithm 1 has*

$$\mathcal{O}\left(\left(\frac{\omega}{q_\omega} + \frac{\omega L}{\mu q_\omega} + \frac{\omega^{3/2}\hat{L}}{\mu q_\omega\sqrt{M}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

communication complexity, where q_ω is the expected density and ε represents accuracy of the solution.

Proof. Having result of Theorem 2.5, we substitute the choice of γ and obtain the following iteration complexity:

$$\mathcal{O}\left(\left(\omega + \frac{\omega L}{\mu} + \frac{\omega^{3/2}\hat{L}}{\mu\sqrt{M}}\right)\log\left(\frac{1}{\varepsilon}\right)\right).$$

Now assuming, that at each iteration we transmit $\frac{1}{q_\omega}$ bits of information, where q_ω is the expected density, we ends the proof of the corollary. □

D Proof of Theorem 2.8

Theorem 2.8 *Suppose Assumptions 2.1, 2.3 hold. Then for Algorithm 2 with $\gamma \leq \min\left\{\frac{\alpha}{8\mu}, \frac{\alpha^2}{16\sqrt{1772}\sqrt{\frac{1}{M}\sum\limits_{i=1}^{M}L_i^2}}\right\}$ and $\beta = 1-\frac{\alpha}{8}$ the following estimate is fulfilled:*

$$\mathbb{E}V^K \leq (1-\gamma\mu)^K V^0,$$

where $V^k = \left\|z^k - z^*\right\|^2 + \frac{7088\gamma^2}{\alpha^3}\frac{1}{M}\sum\limits_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}}) - h_i^k\right\|^2 + \frac{276\gamma^2}{\alpha}\frac{1}{M}\sum\limits_{i=1}^{M}\left\|e_i^k\right\|.$

Proof. Firstly, we introduce new sequences $\hat{z}^k$ and $\hat{z}^{k+\frac{1}{2}}$:

$$\hat{z}^k := z^k - \gamma e^k, \tag{16}$$

$$\hat{z}^{k+\frac{1}{2}} := z^{k+\frac{1}{2}} - \gamma e^k. \tag{17}$$

Note then, that the following equal transition is held:

$$\begin{aligned}
\hat{z}^{k+1} &\overset{16}{=} z^{k+1} - \gamma e^{k+1} \\
&= (z^k - \gamma g^k) - \gamma \frac{1}{M} \sum_{i=1}^{M} \left[e_i^k + \Delta_i^k - \hat{\Delta}_i^k \right] \\
&= z^k - \gamma e^k - \gamma g^k - \gamma \frac{1}{M} \sum_{i=1}^{M} \left[F_i^k(z^{k+\frac{1}{2}}) - h_i^k + h_i^k - g_i^k \right] \\
&= \hat{z}^k - \gamma F(z^{k+\frac{1}{2}}).
\end{aligned}$$

We use a classical for EXTRAGRADIENT proofs equation for arbitrary z, v and u:

$$\|z-u\|^2 = \|v-u\|^2 + 2\langle z-v, z-u\rangle - \|z-v\|^2 .$$

For $z = z^{k+\frac{1}{2}}$, $u = z^*$ and $v = \hat{z}^k$:

$$\left\| z^{k+\frac{1}{2}} - z^* \right\|^2 = \left\| \hat{z}^k - z^* \right\|^2 + 2\left\langle z^{k+\frac{1}{2}} - \hat{z}^k, z^{k+\frac{1}{2}} - z^* \right\rangle - \left\| z^{k+\frac{1}{2}} - \hat{z}^k \right\|^2 .$$

Using the same for $z = \hat{z}^{k+1}$ with $u = z^*$ and $v = z^{k+\frac{1}{2}}$,

$$\left\| \hat{z}^{k+1} - z^* \right\|^2 = \left\| z^{k+\frac{1}{2}} - z^* \right\|^2 + 2\left\langle \hat{z}^{k+1} - z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}} - z^* \right\rangle + \left\| z^{k+\frac{1}{2}} - \hat{z}^{k+1} \right\|^2 .$$

After the summation with the previous equality, we obtain

$$\begin{aligned}
\left\| \hat{z}^{k+1} - z^* \right\|^2 &= \left\| \hat{z}^k - z^* \right\|^2 + 2\left\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+\frac{1}{2}} - z^* \right\rangle \\
&\quad + \left\| z^{k+\frac{1}{2}} - \hat{z}^{k+1} \right\|^2 - \left\| z^{k+\frac{1}{2}} - \hat{z}^k \right\|^2 . \qquad (18)
\end{aligned}$$

Let's work with the term $\left\| \hat{z}^{k+1} - z^{k+\frac{1}{2}} \right\|^2$:

$$\begin{aligned}
\left\| \hat{z}^{k+1} - z^{k+\frac{1}{2}} \right\|^2 &= 2\left\| \hat{z}^{k+1} - \hat{z}^{k+\frac{1}{2}} \right\|^2 + 2\left\| \hat{z}^{k+\frac{1}{2}} - z^{k+\frac{1}{2}} \right\|^2 \\
&= 2\gamma^2 \left\| e^{k+1} - e^k + g^k - h^k \right\|^2 + 2\gamma^2 \left\| e^k \right\|^2 \\
&\overset{CB}{\leq} 4\gamma^2 \left\| \frac{1}{M} \sum_{i=1}^{M} \left[\Delta_i^k - \hat{\Delta}_i^k \right] \right\|^2 + 4\gamma^2 \left\| g^k - h^k \right\| + 2\gamma^2 \left\| e^k \right\|^2 \\
&\overset{CB}{\leq} 8\gamma^2 \\
&\quad \cdot \left\| \frac{1}{M} \sum_{i=1}^{M} \left[\left(F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right) - Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right) \right] \right\|^2 \\
&\quad + 8\gamma^2 \left\| e^k \right\|^2 + 4\gamma^2 \left\| g^k - h^k \right\| + 2\gamma^2 \left\| e^k \right\|^2 \\
&\overset{\text{Def. 1.5}}{\overset{CB}{\leq}} 8\gamma^2(1-\alpha) \frac{1}{M} \sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right\|^2 + 10\gamma^2 \left\| e^k \right\|^2 \\
&\quad + 4\gamma^2 \left\| g^k - h^k \right\|^2 \\
&\overset{CB}{\leq} 16\gamma^2(1-\alpha) \frac{1}{M} \sum_{i=1}^{M} \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
&\quad + (26 - 16\alpha)\gamma^2 \frac{1}{M} \sum_{i=1}^{M} \left\| e_i^k \right\|^2 + 4\gamma^2 \left\| g^k - h^k \right\|^2 . \qquad (19)
\end{aligned}$$

The term $2\left\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+\frac{1}{2}} - z^* \right\rangle$ can be bounded in the following way, using the definition of $\hat{z}^{k+1}$ and μ-strongly monotonicity of operator F:

$$\begin{aligned}
2\left\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+\frac{1}{2}} - z^* \right\rangle &\overset{16}{=} -2\gamma \left\langle F(z^{k+\frac{1}{2}}), z^{k+\frac{1}{2}} - z^* \right\rangle \\
&= -2\gamma \left\langle F(z^{k+\frac{1}{2}}) - F(z^*), z^{k+\frac{1}{2}} - z^* \right\rangle
\end{aligned}$$

$$
\begin{aligned}
& -2\gamma \underbrace{\left\langle F(z^*), z^{k+\frac{1}{2}} - z^* \right\rangle}_{\leq 0} \\
\overset{\text{Ass. 2.3}}{\leq} \quad & -2\gamma\mu \left\| z^{k+\frac{1}{2}} - z^* \right\|^2 .
\end{aligned}
$$

Substituting this result with equation 19 into equation 18 we get

$$
\begin{aligned}
\left\| \hat{z}^{k+1} - z^* \right\|^2 \quad & \leq \quad \left\| \hat{z}^k - z^* \right\|^2 - 2\gamma\mu \left\| z^{k+\frac{1}{2}} - z^* \right\|^2 \\
& \quad + 16\gamma^2(1-\alpha)\frac{1}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& \quad + (26 - 16\alpha)\gamma^2 \frac{1}{M}\sum_{i=1}^M \left\| e_i^k \right\|^2 \\
& \quad + 4\gamma^2 \left\| g^k - h^k \right\|^2 - \left\| z^{k+\frac{1}{2}} - \hat{z}^k \right\|^2 \\
& = \quad (1-\gamma\mu) \left\| \hat{z}^k - z^* \right\|^2 \\
& \quad + 16\gamma^2(1-\alpha)\frac{1}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& \quad + \gamma^2 \left(26 - 16\alpha\right) \frac{1}{M}\sum_{i=1}^M \left\| e_i^k \right\|^2 \\
& \quad + 4\gamma^2 \left\| g^k - h^k \right\|^2 - (1-\gamma\mu) \left\| z^{k+\frac{1}{2}} - \hat{z}^k \right\|^2 \\
& \overset{Jensen}{\leq} \quad (1-\gamma\mu) \left\| \hat{z}^k - z^* \right\|^2 \\
& \quad + 16\gamma^2(1-\alpha)\frac{1}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 \\
& \quad + \gamma^2 \left(27 - 16\alpha - \gamma\mu\right) \frac{1}{M}\sum_{i=1}^M \left\| e_i^k \right\|^2 \\
& \quad + 4\gamma^2 \left\| g^k - h^k \right\|^2 - \frac{1-\gamma\mu}{2} \left\| z^{k+\frac{1}{2}} - z^k \right\|^2 .
\end{aligned}
$$

Once again using the Jensen inequality, we deal with the term $\left\| g^k - h^k \right\|^2$:

$$
\begin{aligned}
\left\| g^k - h^k \right\|^2 \quad & = \quad \left\| \hat{\Delta}^k \right\| = \left\| \frac{1}{M}\sum_{i=1}^M \hat{\Delta}_i^k \right\|^2 \\
& \overset{CB}{\leq} \quad \frac{2}{M}\sum_{i=1}^M \left\| Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right) - \left(F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right) \right\|^2 \\
& \quad + \frac{2}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right\|^2 \\
& \overset{\text{Def. 1.5}}{\leq} \quad \frac{4-2\alpha}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k + e_i^k \right\|^2 \\
& \overset{CB}{\leq} \quad \frac{8-4\alpha}{M}\sum_{i=1}^M \left\| F_i(z^{k+\frac{1}{2}}) - h_i^k \right\|^2 + \frac{8-4\alpha}{M}\sum_{i=1}^M \left\| e_i^k \right\|^2 . \qquad (20)
\end{aligned}
$$

In that way, we achieve

$$
\left\| \hat{z}^{k+1} - z^* \right\|^2 \quad \leq \quad (1-\gamma\mu) \left\| \hat{z}^k - z^* \right\|^2
$$

$$
\begin{aligned}
&+\gamma^2(48-32\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2\\
&+\gamma^2\left(59-32\alpha\right)\frac{1}{M}\sum_{i=1}^{M}\left\|e_i^k\right\|^2\\
&-\frac{1-\gamma\mu}{2}\left\|z^{k+\frac{1}{2}}-z^k\right\|^2.
\end{aligned}
\tag{21}
$$

Let's derive the recursion on the terms $\frac{1}{M}\sum\limits_{i=1}^{M}\left\|F_i(z^{k+\frac{3}{2}})-h_i^{k+1}\right\|^2$ and $\frac{1}{M}\sum\limits_{i=1}^{M}\left\|e_i^{k+1}\right\|^2$. Firstly, one can note that

$$
\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{3}{2}})-h_i^{k+1}\right\|^2\\
&= \quad \frac{1}{M}\sum_{i=1}^{M}\left\|\left(F_i(z^{k+\frac{3}{2}})-F_i(z^{k+\frac{1}{2}})\right)+\left(F_i(z^{k+\frac{1}{2}})-h_i^{k+1}\right)\right\|^2\\
&\overset{Jensen}{\leq} (1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{3}{2}})-F_i(z^{k+\frac{1}{2}})\right\|^2\\
&\qquad +(1+s_1)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^{k+1}\right\|^2\\
&\overset{\text{Ass. 2.1}}{\leq} (1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2\\
&\qquad +(1+s_1)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k-\beta Q(F_i(z^{k+\frac{1}{2}})-h_i^k)\right\|^2\\
&\overset{Jensen}{\leq} (1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2\\
&\qquad +(1+s_1)(1+s_2)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k-Q(F_i(z^{k+\frac{1}{2}})-h_i^k)\right\|^2\\
&\qquad +(1+s_1)(1+s_2^{-1})(1-\beta)^2\frac{1}{M}\sum_{i=1}^{M}\left\|Q\left(F_i(z^{k+\frac{1}{2}})-h_i^k\right)\right\|^2\\
&\overset{\substack{\text{Def. 1.5}\\ CB}}{\leq} (1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2\\
&\qquad +(1+s_1)(1+s_2)(1-\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2\\
&\qquad +2(1+s_1)(1+s_2^{-1})(1-\beta)^2\frac{1}{M}\sum_{i=1}^{M}\left\|Q\left(F_i(z^{k+\frac{1}{2}})-h_i^k\right)-\left(F_i(z^{k+\frac{1}{2}})-h_i^k\right)\right\|^2\\
&\qquad +2(1+s_1)(1+s_2^{-1})(1-\beta)^2\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2\\
&\overset{\text{Def. 1.5}}{\leq} (1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2\\
&\qquad +(1+s_1)(1+s_2)(1-\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2
\end{aligned}
$$

$$+(1+s_1)(1+s_2^{-1})(1-\beta)^2(4-2\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2. \tag{22}$$

Now we proceed to $\frac{1}{M}\sum\limits_{i=1}^{M}\left\|e_i^{k+1}\right\|^2$:

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\left\|e_i^{k+1}\right\|^2 &= \frac{1}{M}\sum_{i=1}^{M}\left\|e_i^k+\Delta_i^k-\hat{\Delta}_i^k\right\|^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k+e_i^k-Q\left(F_i(z^{k+\frac{1}{2}})-h_i^k+e_i^k\right)\right\|^2 \\
&\overset{\substack{\text{Def. 1.5}\\ Jensen}}{\leq} (1+s_3^{-1})(1-\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2 \\
&\quad +(1+s_3)(1-\alpha)\frac{1}{M}\sum_{i=1}^{M}\left\|e_i^k\right\|^2.
\end{aligned} \tag{23}$$

Finally, we denote

$$\begin{aligned}
\mathcal{Z}^k &= \left\|\hat{z}^{k+1}-z^*\right\|^2, \\
\mathcal{F}^k &= \frac{1}{M}\sum_{i=1}^{M}\left\|F_i(z^{k+\frac{1}{2}})-h_i^k\right\|^2, \\
\mathbb{E}^k &= \frac{1}{M}\sum_{i=1}^{M}\left\|e_i^k\right\|^2
\end{aligned}$$

and sum up the estimates, obtained in equation 22, equation 23 (with coefficients $A\gamma^2$ and $C\gamma^2$, which we define later, respectively) with equation 21:

$$\begin{aligned}
\mathcal{Z}^{k+1}+A\gamma^2\mathcal{F}^{k+1}+C\gamma^2\mathbb{E}^{k+1} &\leq (1-\gamma\mu)\mathcal{Z}^k \\
&\quad +\gamma^2(48-32\alpha)\mathcal{F}^k \\
&\quad +\gamma^2\left(59-32\alpha\right)\mathbb{E}^k \\
&\quad -\frac{1-\gamma\mu}{2}\left\|z^{k+\frac{1}{2}}-z^k\right\|^2 \\
&\quad +A\gamma^2(1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2 \\
&\quad +A\gamma^2(1+s_1)(1+s_2)(1-\alpha)\mathcal{F}^k \\
&\quad +A\gamma^2(1+s_1)(1+s_2^{-1})(1-\beta)^2(4-2\alpha)\mathcal{F}^k \\
&\quad +C\gamma^2(1+s_3^{-1})(1-\alpha)\mathcal{F}^k \\
&\quad +C\gamma^2(1+s_3)(1-\alpha)\mathbb{E}^k \\
&= (1-\gamma\mu)\mathcal{Z}^k \\
&\quad +\gamma^2\left[(48-32\alpha)+A(1+s_1)(1+s_2)(1-\alpha)\right. \\
&\quad + A(1+s_1)(1+s_2^{-1})(1-\beta)^2(4-2\alpha) \\
&\quad \left.+C(1+s_3^{-1})(1-\alpha)\right]\mathcal{F}^k \\
&\quad +\gamma^2\left[(59-32\alpha)+C(1+s_3)(1-\alpha)\right]\mathbb{E}^k \\
&\quad -\frac{1-\gamma\mu}{2}\gamma^2\left\|h^k\right\|^2 \\
&\quad +A\gamma^2(1+s_1^{-1})\frac{1}{M}\sum_{i=1}^{M}L_i^2\left\|z^{k+\frac{3}{2}}-z^{k+\frac{1}{2}}\right\|^2.
\end{aligned} \tag{24}$$

Note that in the previous equation we used the fact, that $\left\|z^{k+\frac{1}{2}} - z^k\right\|^2 = \gamma^2 \left\|h^k\right\|^2$. Next, consider the last term:

$$\begin{aligned}
\left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}}\right\|^2 &= \left\|z^{k+\frac{3}{2}} - z^{k+\frac{1}{2}} \pm z^{k+1}\right\|^2 \\
&\overset{CB}{\leq} 2\left\|z^{k+1} - z^{k+\frac{1}{2}}\right\|^2 + 2\left\|z^{k+\frac{3}{2}} - z^{k+1}\right\|^2 \\
&= 2\gamma^2 \left\|g^k - h^k\right\|^2 + 2\gamma^2 \left\|h^{k+1}\right\|^2 \\
&\overset{20}{\leq} (16-8\alpha)\gamma^2 \mathcal{F}^k + (16-8\alpha)\gamma^2 \mathbb{E}^k + 2\gamma^2 \left\|h^{k+1}\right\|^2 \\
&\overset{CB}{\leq} (16-8\alpha)\gamma^2 \mathcal{F}^k + (16-8\alpha)\gamma^2 \mathbb{E}^k \\
&\quad +2\gamma^2 \left\|h^k + \beta \frac{1}{M}\sum_{i=1}^M Q\left(F_i(z^{k+\frac{1}{2}}) - h_i^k\right)\right\|^2 \\
&\overset{CB}{\leq} (16-8\alpha)\gamma^2 \mathcal{F}^k + (16-8\alpha)\gamma^2 \mathbb{E}^k \\
&\quad +4\gamma^2 \left\|h^k\right\|^2 + (16-8\alpha)\gamma^2\beta^2 \mathcal{F}^k,
\end{aligned}$$

which we substitute this term into the equation 24:

$$\begin{aligned}
\mathcal{Z}^{k+1} + A\gamma^2\mathcal{F}^{k+1} + C\gamma^2\mathbb{E}^{k+1} \leq\ & (1-\gamma\mu)\mathcal{Z}^k \\
& +\gamma^2 \left[(48-32\alpha) + A(1+s_1)(1+s_2)(1-\alpha)\right. \\
& + A(1+s_1^{-1})(16-8\alpha)(1+\beta^2)\gamma^2 \frac{1}{M}\sum_{i=1}^M L_i^2 \\
& + A(1+s_1)(1+s_2^{-1})(1-\beta)^2(4-2\alpha) \\
& \left.+C(1+s_3^{-1})(1-\alpha)\right] \mathcal{F}^k \\
& +\gamma^2 \left[(59-32\alpha) + C(1+s_3)(1-\alpha)\right. \\
& \left.+A(1+s_1^{-1})\gamma^2 \frac{1}{M}\sum_{i=1}^M L_i^2(16-8\alpha)\right] \mathbb{E}^k \\
& +\gamma^2 \left[4A(1+s_1^{-1})\gamma^2 \frac{1}{M}\sum_{i=1}^M L_i^2 - \frac{1-\gamma\mu}{2}\right] \left\|h^k\right\|^2.
\end{aligned}$$

Due to the arbitrary choice for parameters s_1 and s_2, we can set them equal to $\frac{1}{\sqrt[4]{1-\alpha}} - 1$, from which instantly follows that

$$1 + s_{1,2} = \frac{1}{\sqrt[4]{1-\alpha}} \qquad \text{and} \qquad 1 + s_{1,2}^{-1} = \frac{1}{1-\sqrt[4]{1-\alpha}} \leq \frac{4}{\alpha},$$

similarly, for products:

$$\begin{aligned}
(1+s_1)(1+s_2) &= \frac{1}{\sqrt{1-\alpha}}, \\
(1+s_1)(1+s_2^{-1}) &= \frac{1}{\sqrt[4]{1-\alpha}\cdot\left(1-\sqrt[4]{1-\alpha}\right)} \leq \frac{4}{\alpha}.
\end{aligned}$$

Inserting these results into the coefficient in front of $\mathcal{F}^k$ and assuming $s_3^{-1} = \frac{2}{\alpha}$, one can obtain:

$$\begin{aligned}
(48-32\alpha) \quad &+ \quad A(1+s_1)(1+s_2)(1-\alpha) \\
&+ \quad A(1+s_1^{-1})(16-8\alpha)(1+\beta^2\gamma^2 \frac{1}{M}\sum_{i=1}^M L_i^2 \\
&+ \quad A(1+s_1)(1+s_2^{-1})(1-\beta)^2(4-2\alpha) \\
&+ \quad C(1+s_3^{-1})(1-\alpha) \\
&= \quad (48-32\alpha) + A\sqrt{1-\alpha}
\end{aligned}$$

$$
\begin{aligned}
&+ \quad A\left(\frac{1}{1-\sqrt[4]{1-\alpha}}\right)(16-8\alpha)(1+\beta^2)\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2 \\
&+ \quad A\frac{1}{\sqrt[4]{1-\alpha}\cdot\left(1-\sqrt[4]{1-\alpha}\right)}(1-\beta)^2(4-2\alpha)+C\left(1+\frac{2}{\alpha}\right)(1-\alpha) \\
&\leq \quad (48-32\alpha)+A\left(1-\frac{\alpha}{2}\right)+16A\left(\frac{4}{\alpha}\right)\left(1-\frac{\alpha}{2}\right)(1+\beta^2)\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2 \\
&+ \quad A\cdot\left(\frac{4}{\alpha}\right)(1-\beta)^2(4-2\alpha)+C\cdot\frac{2}{\alpha}
\end{aligned}
$$

Consider $\beta = 1-\frac{\alpha}{8}$, therefore

$$
\begin{aligned}
(48-32\alpha)+A\left(1-\frac{\alpha}{2}\right) \quad &+ \quad 16A\left(\frac{4}{\alpha}\right)\left(1-\frac{\alpha}{2}\right)(1+\beta^2)\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2 \\
&+ \quad A\left(\frac{4}{\alpha}\right)(1-\beta)^2(4-2\alpha)+C\cdot\frac{2}{\alpha} \\
&\leq \quad (48-32\alpha)+A\left(1-\frac{\alpha}{2}\right)+128\frac{A}{\alpha}\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2 \\
&+ \quad A\cdot\frac{\alpha}{4}+C\cdot\frac{2}{\alpha}.
\end{aligned}
$$

And finally, if $\gamma \leq \dfrac{1}{8\sqrt{\frac{1}{M}\sum\limits_{i=1}^{M}L_i^2}}\sqrt{\dfrac{\alpha}{A}}$ we get

$$
(48-32\alpha)+A\left(1-\frac{\alpha}{2}\right)+2+A\cdot\frac{\alpha}{4}+C\cdot\frac{2}{\alpha} = (50-32\alpha)+A\left(1-\frac{\alpha}{4}\right)+C\cdot\frac{2}{\alpha}.
$$

Further, we do the similar thing for the coefficient before $\mathbb{E}^k$:

$$
\begin{aligned}
(59-32\alpha)+C(1+s_3)(1-\alpha) \quad &+ \quad A(1+s_1^{-1})\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2(16-8\alpha) \\
&\leq \quad (59-32\alpha)+C\left(1+\frac{\alpha}{2}\right)(1-\alpha) \\
&+ \quad A\frac{64}{\alpha}\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2\left(1-\frac{\alpha}{2}\right) \\
&\leq \quad (60-32\alpha)+C\left(1-\frac{\alpha}{2}\right).
\end{aligned}
$$

Summing with coefficients $p^k = (1-\gamma\mu)^{-k}$:

$$
\begin{aligned}
&\sum_{k=0}^{K-1}p^k\mathcal{Z}^{k+1}+\sum_{k=0}^{K-1}p^kA\gamma^2\mathcal{F}^{k+1}+\sum_{k=0}^{K-1}p^kC\gamma^2\mathbb{E}^{k+1} \\
&\leq \sum_{k=0}^{K-1}p^k(1-\gamma\mu)\mathcal{Z}^k \\
&\quad+\sum_{k=0}^{K-1}p^k\gamma^2\left[(50-32\alpha)+A\left(1-\frac{\alpha}{4}\right)+C\cdot\frac{2}{\alpha}\right]\mathcal{F}^k \\
&\quad+\sum_{k=0}^{K-1}p^k\gamma^2\left[(60-32\alpha)+C\left(1-\frac{\alpha}{2}\right)\right]\mathbb{E}^k \\
&\quad+\sum_{k=0}^{K-1}p^k\gamma^2\left[\frac{16A}{\alpha}\gamma^2\frac{1}{M}\sum_{i=1}^{M}L_i^2-\frac{1-\gamma\mu}{2}\right]\left\|h^k\right\|^2.
\end{aligned}
$$

Now we found the following estimates with $p^{-1} = 1 - \gamma\mu$:

$$p\left[(50 - 32\alpha) + A\left(1 - \frac{\alpha}{4}\right) + C \cdot \frac{2}{\alpha}\right] \quad = \quad \frac{(50 - 32\alpha) + A\left(1 - \frac{\alpha}{4}\right) + C \cdot \frac{2}{\alpha}}{1 - \gamma\mu}.$$

Now we choose $\gamma \leq \dfrac{\alpha}{8\mu}$ to fulfill next inequality:

$$\frac{1}{1 - \gamma\mu} \leq \frac{1}{1 - \frac{\alpha}{8}}.$$

In that way,

$$\begin{aligned} \frac{(50 - 32\alpha) + A\left(1 - \frac{\alpha}{4}\right) + C \cdot \frac{2}{\alpha}}{1 - \gamma\mu} \quad &\leq \quad \frac{(50 - 32\alpha) + A\left(1 - \frac{\alpha}{4}\right) + C \cdot \frac{2}{\alpha}}{1 - \frac{\alpha}{8}} \\ &\leq \quad (58 - 37\alpha) + A\left(1 - \frac{\alpha}{8}\right) + C \cdot \frac{3}{\alpha}. \end{aligned}$$

Analogically, we can estimate

$$\begin{aligned} p\left[(60 - 32\alpha) + C\left(1 - \frac{\alpha}{2}\right)\right] \quad &= \quad \frac{(60 - 32\alpha) + C\left(1 - \frac{\alpha}{2}\right)}{1 - \frac{\alpha}{8}} \\ &\leq \quad (69 - 37\alpha) + C\left(1 - \frac{\alpha}{4}\right). \end{aligned}$$

Next we choose

$$\begin{cases} A & \geq (58 - 37\alpha) + A\left(1 - \frac{\alpha}{8}\right) + C \cdot \frac{3}{\alpha}; \\ C & \geq (69 - 37\alpha) + C\left(1 - \frac{\alpha}{4}\right); \end{cases}$$

$$\begin{cases} A & \geq \dfrac{464}{\alpha} + \dfrac{6624}{\alpha^3} \geq \dfrac{7088}{\alpha^3}; \\ C & \geq \dfrac{276}{\alpha}. \end{cases}$$

One can observe, that the coefficient in front of the $\left\|h^k\right\|^2$ is less, than zero:

$$4A(1 + s_1^{-1})(\gamma L)^2 - \frac{1 - \gamma\mu}{2} \leq \frac{1}{4} - \frac{1 - \gamma\mu}{2} = -\frac{1}{4} - \frac{\gamma\mu}{2} < 0,$$

with

$$\gamma \leq \min\left\{\frac{\alpha}{8\mu}, \frac{\alpha^2}{16\sqrt{1772}\sqrt{\frac{1}{M}\sum\limits_{i=1}^{M} L_i^2}}\right\}.$$

Lastly, we get

$$\begin{aligned} \sum_{k=0}^{K-1} p^k \mathcal{Z}^{k+1} + \sum_{k=0}^{K-1} p^k A\gamma^2 \mathcal{F}^{k+1} + \sum_{k=0}^{K-1} p^k C\gamma^2 \mathbb{E}^{k+1} \quad \leq \quad & \sum_{k=0}^{K-1} p^k (1 - \gamma\mu) \mathcal{Z}^k \\ & + \sum_{k=0}^{K-1} p^{k-1} \gamma^2 A \mathcal{F}^k \\ & + \sum_{k=0}^{K-1} p^{k-1} \gamma^2 C \mathbb{E}^k. \end{aligned}$$

Using the choice of p we derive the final estimate:

$$\begin{aligned} \sum_{k=0}^{K-1} (1 - \gamma\mu)^{-k} \mathcal{Z}^{k+1} \quad & + \quad \sum_{k=0}^{K-1} (1 - \gamma\mu)^{-k} A\gamma^2 \mathcal{F}^{k+1} \\ & + \quad \sum_{k=0}^{K-1} (1 - \gamma\mu)^{-k} C\gamma^2 \mathbb{E}^{k+1} \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}\mathcal{Z}^k \\
&+ \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}\gamma^2 A\mathcal{F}^k \\
&+ \sum_{k=0}^{K-1}(1-\gamma\mu)^{-k+1}\gamma^2 C\mathbb{E}^k; \\
(1-\gamma\mu)^{-K+1}\mathcal{Z}^K + (1-\gamma\mu)^{-K+1}A\gamma^2\mathcal{F}^K &+ (1-\gamma\mu)^{-K+1}C\gamma^2\mathbb{E}^K \\
&\leq (1-\gamma\mu)\mathcal{Z}^0 \\
&+ (1-\gamma\mu)\gamma^2 A\mathcal{F}^0 \\
&+ (1-\gamma\mu)\gamma^2 C\mathbb{E}^0; \\
V^K &\leq (1-\gamma\mu)^K V^0,
\end{aligned}$$

where we denote $V^k = \mathcal{Z}^k + A\gamma^2\mathcal{F}^k + C\gamma^2\mathbb{E}^k$. Taking full expectation ends the proof of the theorem. □

Corollary 2.9 *Suppose the conditions of Theorem 2.8 hold. Then Algorithm 2 has*

$$\mathcal{O}\left(\left(\frac{\omega}{q_\alpha} + \frac{\hat{L}}{\mu\alpha^2 q_\alpha}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

communication complexity, where q_α is the expected density and ε represents accuracy of the solution.

Proof. Proof repeats the proof of Corollary 2.6. □

The Use of Large Language Models (LLMs)

In this work, large language models (LLMs) were used exclusively for spelling edits.