# Distributed Online Convex Optimization with Compressed Communication: Optimal Regret and Applications

**Sifan Yang**[1,2], **Danyue Li**[1,2], **Lijun Zhang**[1,2]*
[1]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]School of Artificial Intelligence, Nanjing University, Nanjing, China
{yangsf, lidy, zhanglj}@lamda.nju.edu.cn

## Abstract

Distributed online convex optimization (D-OCO) is a powerful paradigm for modeling distributed scenarios with streaming data. However, the communication cost between local learners and the central server is substantial in large-scale applications. To alleviate this bottleneck, we initiate the study of D-OCO with compressed communication. Firstly, to quantify the compression impact, we establish the $\Omega(\delta^{-1/2}\sqrt{T})$ and $\Omega(\delta^{-1}\log T)$ lower bounds for convex and strongly convex loss functions, respectively, where $\delta \in (0, 1]$ is the compression ratio. Secondly, we propose an optimal algorithm, which enjoys regret bounds of $O(\delta^{-1/2}\sqrt{T})$ and $O(\delta^{-1}\log T)$ for convex and strongly convex loss functions, respectively. Our method innovatively incorporates the *error feedback mechanism* into the *Follow-the-Regularized-Leader* framework to address the coupling between the compression error and the projection error. Furthermore, we employ the *online compression strategy* to mitigate the accumulated error arising from the bidirectional compression. Our online method has great generality, and can be extended to the offline stochastic setting via online-to-batch conversion. We establish convergence rates of $O(\delta^{-1/2}T^{-1/2})$ and $O(\delta^{-1}T^{-1})$ for convex and strongly convex loss functions, respectively, providing the *first* guarantees for distributed non-smooth optimization with compressed communication and domain constraints.

## 1 Introduction

Distributed online convex optimization (D-OCO) [Patel et al., 2023] is a fundamental paradigm for modeling distributed sequential decision-making problems, such as mobile keyboard predictions [Hard et al., 2018, Chen et al., 2019], self-driving vehicles [Elbir et al., 2020], recommendation systems [Shi et al., 2021, Liang et al., 2021, Khan et al., 2021], among others. Specifically, it is formulated as a repeated game between an adversary and a set of $n$ local learners connected to a central server. At each round $t \in [T]$, each online learner plays a *global decision* $\mathbf{w}^t$ from a convex set $\mathcal{W} \subseteq \mathbb{R}^d$. After the decision $\mathbf{w}^t$ is committed, learner $i$ suffers a local loss $f_i^t(\mathbf{w}^t)$, where $f_i^t(\cdot) : \mathcal{W} \to \mathbb{R}$ is a convex function. Subsequently, learner $i$ sends its local information to the central server, which aggregates the received data and broadcasts the global information back to all learners. Upon receiving this message, local learners synchronously update the decision $\mathbf{w}^{t+1}$. To measure the performance of learners, we select the *standard regret*:

$$R_T = \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(f_i^t(\mathbf{w}^t) - f_i^t(\mathbf{w}^*)\right), \tag{1}$$

*Lijun Zhang is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

where $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w})$ is the best decision chosen in hindsight.

It is straightforward to verify that if each learner transmits its local gradient to the central server and updates the decision with the aggregated gradient from the server, the distributed setting reduces to the single-node case, and the procedure attains the optimal regret bound for single-node OCO [Shalev-Shwartz, 2012, Hazan, 2016]. In practice, however, transmitting full gradients between the learners and the server may incur substantial *communication overhead*, limiting the applicability of this approach in large-scale distributed problems. To alleviate this bottleneck in D-OCO, we conduct the *first* investigation of communication-efficient D-OCO, where learners and the server exchange compressed messages. We note that prior works have explored compressed communication in decentralized online convex optimization [Tu et al., 2022, Cao and Başar, 2023, Yang et al., 2026], but their methods are tailored for peer-to-peer topologies. Consequently, they are incompatible with our distributed setting due to the fundamental differences in network architecture.

Naively using the compressed gradients, e.g., updating with the sign of the gradients, may result in the unbounded accumulation of the compression error, which prevents convergence even in the offline optimization [Karimireddy et al., 2019]. To address this issue, the error feedback (EF) mechanism [Seide et al., 2014] is widely adopted to control the compression error in unconstrained settings [Alistarh et al., 2018, Karimireddy et al., 2019, Stich and Karimireddy, 2020, Richtárik et al., 2021, Huang et al., 2022]. At each round, EF records the accumulated error of the past compression steps and compresses the sum of the error and gradient, effectively bounding the total compression error over time. However, a critical challenge arises when extending the existing EF-based optimization methods to the distributed settings with domain constraints. Since each learner must project its decision onto the feasible domain per update, it results in an additional projection error, which *couples* with the compression error and leads to unbounded error accumulation. Unlike unconstrained optimization where EF alone suffices to control the compression error, the coupled error in constrained settings necessitates novel techniques for effective handling.

In this paper, we first develop Distributed Follow-the-Compressed-Leader (D-FTCL), which applies EF on both learner and server sides to enable bidirectional compressed communication. More importantly, D-FTCL *innovatively* adapts EF to the Follow-the-Regularized-Leader (FTRL) framework [Hazan, 2016, Orabona, 2019], which ensures that the projection error does not propagate to subsequent compression steps, thereby decoupling the projection error from the compression error. D-FTCL enjoys regret bounds of $O(\delta^{-1}\sqrt{T})$ and $O(\delta^{-2}\log T)$ for convex and strongly convex loss functions, respectively, where $\delta \in (0, 1]$ denotes the compression ratio that characterizes the quality of compression and $\delta = 1$ means no compression. Furthermore, to quantify the impact of compression, we establish the $\Omega(\delta^{-1/2}\sqrt{T})$ and $\Omega(\delta^{-1}\log T)$ lower bounds for convex and strongly convex loss functions, respectively. Thus, while D-FTCL successfully achieves efficient communication, there exists a gap between its regret guarantees and the lower bounds. This prompts a natural question: *is it possible to further improve the regret bound of D-FTCL?*

Through careful analysis, we find that the bidirectional compression in D-FTCL introduces the amplified compression error, which results in suboptimal regret bounds. Motivated by previous work [Huang et al., 2022, Yang et al., 2026], we employ the online compression strategy to reduce the compression error, which recursively compresses the *residual* for multiple rounds. However, since the online compression strategy leverages the blocking update mechanism to ensure the communication round remains one per round, the communication between the local learners and the server becomes *asynchronous* due to transmission latency. To tackle this issue, we rigorously characterize the deviation caused by the delayed updates in the analysis. Our method, named Distributed Follow-the-Fast-Compressed-Leader (D-FTFCL), enhances the regret bounds to $O(\delta^{-1/2}\sqrt{T})$ and $O(\delta^{-1}\log T)$ for convex and strongly convex loss functions, respectively, thereby achieving optimal regret bounds for D-OCO with compressed communication.

Our online methods have *broad applicability*, and thus can be extended to the distributed offline stochastic convex optimization. Previous work on distributed offline optimization with compressed communication typically assumes either smooth loss functions or an unconstrained optimization domain [Karimireddy et al., 2019, Richtárik et al., 2021, Huang et al., 2022, Gao et al., 2024, Islamov et al., 2025]. Consequently, the problem of distributed convex non-smooth optimization under domain constraints remains *unsolved*. To bridge this gap, we develop a distributed variant of anytime online-to-batch conversion [Cutkosky, 2019] and integrate our online method into this framework. Our proposed algorithm obtains a *last-iterate* convergence rate of $O(\delta^{-1/2}T^{-1/2})$ for convex loss

functions, matching the lower bound established by Islamov et al. [2025]. For strongly convex loss functions, we establish a faster rate of $O\left(\delta^{-1}T^{-1}\right)$, which is optimal with respect to $T$ [Agarwal et al., 2012]. To the best of our knowledge, these are the *first* guarantees for distributed convex non-smooth optimization with compressed communication and domain constraints. Experiment results in Appendix E demonstrate the effectiveness and efficiency of our methods.

**Notation.** For simplicity, we use $\|\cdot\|$ for $\|\cdot\|_2$ by default and adopt $\Pi_{\mathcal{W}}[\mathbf{x}] = \arg\min_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}-\mathbf{x}\|$ to denote the Euclidean projection onto the domain $\mathcal{W}$. Given sequences of weights $\{\alpha_t\}_t$, we use $\alpha_{1:t} = \sum_{k=1}^{t} \alpha_k$ to represent their summation.

## 2 Related Work

In this section, we briefly review the related work on distributed offline optimization and distributed online convex optimization. We provide additional discussions on related work in Appendix B.

### 2.1 Distributed Offline Optimization

Distributed offline optimization has attracted significant attention in the past decades due to its wide applicability in distributed machine learning tasks [Nedić and Ozdaglar, 2009, Dean et al., 2012, Yuan et al., 2016, Swenson et al., 2022]. Although prior studies have established solid theoretical guarantees, the massive information exchange between local learners and the central server may result in a communication bottleneck, particularly in high-dimensional and large-scale scenarios.

To mitigate errors induced by the sign compressor in distributed offline optimization, Seide et al. [2014] pioneer the error feedback (EF) mechanism, which tracks the past compression error and compresses the sum of the error and the gradient. Subsequently, Karimireddy et al. [2019] provide a unified theoretical framework for EF with general compressors. However, their analysis is limited to unidirectional compression. To achieve bidirectional compression, Stich [2020] and Beznosikov et al. [2023] provide the analysis for EF under the distributed setting. Richtárik et al. [2021] introduce a novel error feedback technique, termed EF21, which compresses the residual of the transmitted data and the gradient. To further mitigate the compression error, Huang et al. [2022] design the fast compressed communication (FCC), which recursively applies the standard compressor for multiple rounds. By utilizing FCC, they establish the nearly optimal convergence rates for non-convex smooth loss functions. More recently, Fatkhullin et al. [2025] extend EF21 into distributed stochastic optimization for smooth loss functions, achieving better convergence rates.

Despite the extensive research on incorporating EF into the offline optimization, existing methods typically require the loss function to be smooth or the feasible domain to be unconstrained [Seide et al., 2014, Stich et al., 2018, Karimireddy et al., 2019, Richtárik et al., 2021, Huang et al., 2022, Fatkhullin et al., 2025]. The only work that addresses offline distributed convex non-smooth optimization is Islamov et al. [2025], which investigates the safe constraint setting. They establish a convergence rate of $O(\delta^{-1}T^{-1/2})$, and a lower bound of $\Omega(\delta^{-1/2}T^{-1/2})$ for unconstrained optimization. However, their method relies on the specific structure of safe constraints and fails to cover domain constraints. Although Islamov et al. [2025] further provide an analysis for Projected-EF21 [Fatkhullin et al., 2025] in the convex smooth setting with domain constraints, they assume smooth loss functions and are restricted to unidirectional compression. It still remains an open problem whether EF is effective for convex non-smooth optimization with domain constraints.

A recent work [Gao et al., 2026] studies error feedback under constraints, but differs from ours in several key aspects. They consider offline distributed composite optimization with smooth loss functions and a non-smooth convex regularizer, whereas we study online convex optimization with non-smooth losses and domain constraints. Their analysis requires the loss to be smooth and the regularizer to admit a tractable proximal operator. Algorithmically, their method combines dual averaging with EControl [Gao et al., 2024] under unidirectional compression, while we incorporate EF into FTRL with bidirectional compression, and further employ an online compression strategy with blocking to attain the optimal regret.

### 2.2 Distributed Online Convex Optimization

To address the distributed optimization with streaming data, Patel et al. [2023] investigate distributed online convex optimization (D-OCO), where the loss function is time-varying and the local learners have to continuously update their decisions to minimize the cumulative regret. They consider a different regret defined as

$$R'_T = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T} f_i^t(\mathbf{w}_i^t) - f_i^t(\mathbf{w}^*)\right),$$

where $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w})$. $R'_T$ measures the performance of each learner by using the local loss functions. As noted by Patel et al. [2023], communication offers no benefit because learners can independently execute OGD [Zinkevich, 2003] to achieve the optimal regret bound without any information exchange. In contrast, the regret defined in (1) evaluates the performance of the decision against the global loss. This regret is more realistic for distributed learning tasks, where the ultimate goal is to train a consensus model. Moreover, when loss functions are fixed over rounds, our regret coincides with the offline optimization objective. Furthermore, they pivot to the bandit feedback setting, where the local learners can only access values of the loss functions [Shamir, 2017]. In this regime, they demonstrate that collaboration helps to achieve tighter regret bounds.

## 3 Preliminaries

In this section, we present necessary preliminaries including assumptions, definitions, and techniques used in this paper. Similar to previous studies on single-node OCO, we introduce the following standard assumptions [Hazan, 2016, Orabona, 2019].

**Assumption 3.1** (Convexity)**.** The loss function $f_i^t(\cdot)$ of each learner $i \in [n]$ is convex over the feasible domain $\mathcal{W}$.

**Assumption 3.2** (Strong convexity)**.** The loss function $f_i^t(\cdot)$ of each learner $i \in [n]$ is $\mu$-strongly convex over the domain $\mathcal{W}$, i.e., it holds that $f_i^t(\mathbf{y}) \geq f_i^t(\mathbf{x}) + \langle \nabla f_i^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2$, for $\forall \mathbf{x}, \mathbf{y} \in \mathcal{W}$.

**Assumption 3.3** (Bounded gradient norm)**.** The gradient of function $f_i^t(\cdot)$ of each learner $i \in [n]$ is bounded by $G$ over the domain $\mathcal{W}$, i.e., it holds that $\|\nabla f_i^t(\mathbf{w})\| \leq G$, for $\forall \mathbf{w} \in \mathcal{W}$. Consequently, the averaged gradient also satisfies $\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i^t(\mathbf{w})\right\| \leq G$, for $\forall \mathbf{w} \in \mathcal{W}$

**Assumption 3.4** (Bounded domain)**.** The convex set $\mathcal{W}$ contains the origin $\mathbf{0}$, i.e., $\mathbf{0} \in \mathcal{W}$, and it is bounded by $D$, i.e., it holds that $\|\mathbf{x}-\mathbf{y}\| \leq D$, for $\forall \mathbf{x}, \mathbf{y} \in \mathcal{W}$.

Next, we introduce an assumption governing the communication process in our distributed setting.

**Assumption 3.5** (Communication protocol)**.** At each round, each learner $i$ and the central server are allowed to communicate with each other exactly once.

A compressor $\mathcal{C}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is a mapping whose output can be efficiently transmitted. In this paper, we consider a general class of compressors with the following definition [Koloskova et al., 2019].

**Definition 3.6** (Compressor)**.** A compressor $\mathcal{C}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is $\delta$-contractive if the following holds

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2\right] \leq (1-\delta)\|\mathbf{x}\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

for some $\delta \in (0, 1]$, where the expectation $\mathbb{E}_{\mathcal{C}}[\cdot]$ is taken over the randomness of the compressor.

We provide two examples of compressor.

- Rand-$K$. Randomly selecting $k$ out of $d$ coordinates yields compressor with the compression ratio of $\delta = \frac{k}{d}$.
- (Scaled) Sign. Let $\text{Sign}(\mathbf{x})$ denote the sign of a vector $\mathbf{x} \in \mathbb{R}^d$. Outputting $\mathcal{C}(\mathbf{x}) = \frac{\|\mathbf{x}\|_1}{d}\text{Sign}(\mathbf{x})$ leads to a compression ratio of $\delta = \frac{1}{d}$.

To mitigate the error of compressors, Huang et al. [2022] design the *fast compressed communication* (FCC) technique, as summarized in Algorithm 1. The core idea is to recursively apply the standard compressor for $L$ rounds and transmit the compressed data at each round, which involves $L$ communication rounds. We state the following lemma to provide the guarantee of FCC.

**Algorithm 1** FCC($\mathbf{x}, \mathcal{C}(\cdot), L$, receiver)

**Input:** data $\mathbf{x}$, compressor $\mathcal{C}(\cdot)$, compression round $L$, receiver
Initialize $\mathbf{r}^1 = \mathbf{0}$
**for** $k = 1$ to $L$ **do**
  Send $\mathbf{c}^k = \mathcal{C}(\mathbf{x} - \mathbf{r}^k)$ to the receiver and calculate $\mathbf{r}^{k+1} = \mathbf{r}^k + \mathbf{c}^k$
**end for**
**Return** $\mathbf{r}^{L+1}$

**Lemma 3.7** (Lemma 2 in Huang et al. [2022]). *Given a $\delta$-contractive compressor $\mathcal{C}(\cdot)$ and for any compression rounds $L \geq 1$, Algorithm 1 ensures*

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{r}^{L+1} - \mathbf{x}\right\|^2\right] \leq (1-\delta)^L \left\|\mathbf{x}\right\|^2, \forall \mathbf{x} \in \mathbb{R}^d.$$

Lemma 3.7 demonstrates that the error in FCC decreases exponentially as $L$ grows, but such a result is achieved at the expense of $L$ communication rounds. When the compression round $L = 1$, FCC mechanism degenerates to the standard compressor.

## 4 Main Results

In this section, we first present an efficient algorithm for D-OCO with compressed communication, as well as the corresponding guarantees. Then, we derive the lower bounds for D-OCO with compressed communication. Finally, we develop an improved method that achieves the optimal regret bounds.

### 4.1 Follow-the-Compressed-Leader

Before presenting our algorithms, we briefly highlight the key challenges in applying EF mechanism to optimization problems with constrained domains. EF is a standard technique for addressing the convergence failures of gradient compression methods [Seide et al., 2014, Karimireddy et al., 2019, Richtárik et al., 2021, Huang et al., 2022]. In the single-node unconstrained setting, the learner initializes $\mathbf{e}^1 = \mathbf{0}$ to track compression errors across iterations and performs the update under the gradient descent framework:

$$\mathbf{e}^{t+1} = \mathbf{e}^t + \mathbf{g}^t - \mathcal{C}(\mathbf{e}^t + \mathbf{g}^t), \quad \mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathcal{C}(\mathbf{e}^t + \mathbf{g}^t),$$

where $\mathbf{g}^t$ is the gradient at round $t$ and $\eta$ is the learning rate. Let $\hat{\mathbf{w}}^{t+1}$ denote the sequence obtained using the exact gradient, i.e., $\hat{\mathbf{w}}^{t+1} = \hat{\mathbf{w}}^t - \eta \mathbf{g}^t$. EF ensures the deviation remains bounded:

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|\right] \leq \mathbb{E}_{\mathcal{C}}\left[\eta \left\|\mathbf{e}^t\right\|\right] \leq O(\delta^{-1}\eta). \tag{2}$$

Consequently, by setting an appropriate learning rate $\eta$, the compression error can be effectively controlled, ensuring that the algorithm converges despite the effect induced by compression. However, applying EF to D-OCO encounters a critical issue due to the *coupling* of the compression error and the projection error. Specifically, since each learner has to project its decision onto the feasible domain, i.e.,

$$\mathbf{w}^{t+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}^t - \eta \mathcal{C}(\mathbf{e}^t + \mathbf{g}^t)\right], \tag{3}$$

it incurs an additional projection error and the projection error will propagate to all subsequent compression steps, thereby invalidating the guarantee in (2). This interaction between projection and compression causes the total error to become *uncontrollable*.

To address the aforementioned limitation, we develop D-FTCL by integrating EF into the Follow-The-Regularized-Leader (FTRL) framework [Hazan et al., 2007, Hazan, 2016] to *decouple* the projection error from the compression error, as summarized in Algorithm 2. Additionally, we employ the EF mechanism on both learner and server sides to achieve bidirectional compression. Specifically, at each round $t$, the learner $i$ plays the global decision $\mathbf{w}^t$ and observes the gradient $\mathbf{g}_i^t$. Then it broadcasts the compressed message $\mathbf{v}_i^t = \mathcal{C}(\mathbf{e}_i^t + \mathbf{g}_i^t)$ to the server and updates $\mathbf{e}_i^{t+1} = \mathbf{e}_i^t + \mathbf{g}_i^t - \mathcal{C}(\mathbf{e}_i^t + \mathbf{g}_i^t)$. Upon receiving all messages from the learners, the server computes the aggregate information $\mathbf{v}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t$. Next, it transmits $\mathbf{s}^t = \mathcal{C}(\hat{\mathbf{e}}^t + \mathbf{v}^t)$ back to learners and updates $\hat{\mathbf{e}}^{t+1} = \hat{\mathbf{e}}^t + \mathbf{v}^t - \mathbf{s}^t$.

**Algorithm 2** Distributed Follow-the-Compressed-Leader (D-FTCL)

**Require:** learning rate $\eta$, compressor $\mathcal{C}(\cdot)$

Initialize $\mathbf{w}^1 = \mathbf{0}$, $\hat{\mathbf{e}}^1 = \mathbf{0}$ and $\mathbf{e}_i^1 = \mathbf{0}, \forall i \in [n]$
**for** $t = 1, \ldots, T$ **do**
  **for each learner** $i = 1, \ldots, n$ **in parallel**
    Play the decision $\mathbf{w}^t$ and receive gradient $\mathbf{g}_i^t$
    Send $\mathbf{v}_i^t = \mathcal{C}(\mathbf{e}_i^t + \mathbf{g}_i^t)$ to server and compute $\mathbf{e}_i^{t+1} = \mathbf{e}_i^t + \mathbf{g}_i^t - \mathbf{v}_i^t$
  **for server do**
    Compute $\mathbf{v}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t$ and send $\mathbf{s}^t = \mathcal{C}(\hat{\mathbf{e}}^t + \mathbf{v}^t)$ to all learners
    Compute $\hat{\mathbf{e}}^{t+1} = \hat{\mathbf{e}}^t + \mathbf{v}^t - \mathbf{s}^t$
  **for each learner** $i = 1, \ldots, n$ **in parallel**
    Update the decision $\mathbf{w}^{t+1}$ according to (4) or (5)
**end for**

In contrast to previous methods that update within the gradient descent framework, each learner updates the decision as follows:

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{s}^k, \mathbf{w} \right\rangle + \frac{1}{\eta} \|\mathbf{w}\|^2 . \tag{4}$$

Since the decision $\mathbf{w}^{t+1}$ is updated with the historical compressed information, the projection error is confined to the current update step. Thus, we can prevent the error from accumulating across rounds, thereby successfully avoiding the coupling between the compression error and the projection error. If the loss function is $\mu$-strongly convex, each learner updates the decision according to:

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{s}^k, \mathbf{w} \right\rangle + \frac{\mu}{2} \sum_{k=1}^{t} \left\|\mathbf{w} - \mathbf{w}^k\right\|^2 . \tag{5}$$

In the following, we establish the guarantees for D-FTCL.

**Theorem 4.1.** *Under the Assumptions 3.1, 3.3 and 3.4, by setting the learning rate* $\eta = \frac{\delta D}{G\sqrt{T}}$, *Algorithm 2 ensures*

$$\mathbb{E}_{\mathcal{C}}\left[R_T\right] \leq O(\delta^{-1} DG\sqrt{T}).$$

**Theorem 4.2.** *Under the Assumptions 3.2, 3.3 and 3.4, Algorithm 2 ensures*

$$\mathbb{E}_{\mathcal{C}}\left[R_T\right] \leq O\left(\mu^{-1}\delta^{-2}(G+\mu D)^2 \log T\right).$$

**Remark.** Our algorithm achieves the $O(\delta^{-1}\sqrt{T})$ and $O\left(\delta^{-2}\log T\right)$ regret bounds for convex and $\mu$-strongly convex loss functions, respectively, which are the first theoretical guarantees for D-OCO with compressed communication and domain constraints.

### 4.2 Lower Bounds

To verify the optimality of our proposed method, we establish the lower bounds for D-OCO with compressed communication. To derive lower bounds for D-OCO with compressed communication, the key challenge is characterizing the effect of compression. Inspired by Huang et al. [2022], we model the compressor as probabilistic communication failure, where transmitted information may not reach its destination at some rounds. We assume learners communicate exclusively with the central server and operate synchronously, ensuring all learners maintain identical decision. Since we focus on the online setting, distinct from the construction in Huang et al. [2022], we adopt the randomized gossip compressor $\mathcal{C}(\cdot): \mathbb{R}^d \to \mathbb{R}^d$ [Koloskova et al., 2019], which outputs $\mathcal{C}(\mathbf{x}) = \mathbf{x}$ with probability $\delta$ and $\mathcal{C}(\mathbf{x}) = \mathbf{0}$ otherwise. Thus, each learner successfully transmits information to the server with probability $\delta$ per round, implying that the expected number of rounds for a successful transmission is $\lceil 1/\delta \rceil$. In this way, we can characterize the effect of compression by analyzing the regret degradation caused by the intermittent communication. We formally state the lower bounds in the following theorems.

**Algorithm 3** Distributed Follow-the-Fast-Compressed-Leader (D-FTFCL)

**Require:** learning rate $\eta$, compressor $\mathcal{C}(\cdot)$, block size $L$

Initialize $\mathbf{w}^1 = \mathbf{0}$, $\hat{\mathbf{e}}^1 = \mathbf{0}$ and $\mathbf{e}_i^1 = \mathbf{0}, \forall i \in [n]$
**for** $b = 1, \ldots, T/L$ **do**
  **for each learner** $i = 1, \ldots, n$ **in parallel**
    **for** $t = (b-1)L + 1, \ldots, bL$ **do**
      Play the decision $\mathbf{w}^b$ and receive gradient $\mathbf{g}_i^t$
    **end for**
    Compute $\mathbf{z}_i^b = \sum_{t=(b-1)L+1}^{bL} \mathbf{g}_i^t$
    **if** $b \geq 2$ **do**
      $\mathbf{v}_i^{b-1}$ = FCC$(\mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1}, \mathcal{C}(\cdot), L,$ server$)$ ▷ Transmission distributed over $L$ rounds
      Compute $\mathbf{e}_i^b = \mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1} - \mathbf{v}_i^{b-1}$
  **for server and if** $b \geq 3$ **do**
    Compute $\mathbf{v}^{b-2} = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{b-2}$ ▷ Only receive gradients of block $b-2$
    $\mathbf{s}^{b-2}$ =FCC$(\mathbf{v}^{b-2} + \hat{\mathbf{e}}^{b-2}, \mathcal{C}(\cdot), L,$ learner$)$ ▷ Transmission distributed over $L$ rounds
    Compute $\hat{\mathbf{e}}^{b-1} = \hat{\mathbf{e}}^{b-2} + \mathbf{v}^{b-2} - \mathbf{s}^{b-2}$
  **for each learner** $i = 1, \ldots, n$ **in parallel and if** $b \geq 3$ **do**
    Update the decision $\mathbf{w}^{b+1}$ according to (6) or (7)
**end for**

**Theorem 4.3.** *Given the feasible domain $\mathcal{W} = [\frac{-D}{2\sqrt{d}}, \frac{D}{2\sqrt{d}}]^d$, for any D-OCO algorithm, if $T \geq (1-\delta)/\delta$, there exists a sequence of convex loss functions satisfying Assumption 3.3 and a compressor satisfying Definition 3.6 such that*

$$\mathbb{E}_{\mathcal{C}}\left[R_T\right] \geq \frac{DG\sqrt{T}}{2^3\sqrt{\delta}}.$$

**Theorem 4.4.** *Given the feasible domain $\mathcal{W} = [0, \frac{D}{\sqrt{d}}]^d$, for any D-OCO algorithm, if $T \geq (16+\delta)/\delta$, there exists a sequence of $\mu$-strongly convex loss functions satisfying Assumption 3.3 and a compressor satisfying Definition 3.6 such that*

$$\mathbb{E}_{\mathcal{C}}\left[R_T\right] \geq \frac{\mu D^2(\log_{16}(15\delta(T-1)) - 3)}{2^{18}\delta}.$$

**Remark.** We establish $\Omega(\delta^{-1/2}\sqrt{T})$ and $\Omega(\delta^{-1}\log T)$ lower bounds for convex and strongly convex loss functions, respectively. When there is no compression ($\delta = 1$), our results recover the classical lower bounds for single-node OCO [Abernethy et al., 2008, Hazan and Kale, 2014].

**Remark.** One might notice that the dependence on the compression ratio $\delta$ of upper bounds in Theorems 4.1 and 4.2 does not match the lower bounds. Therefore, a natural question arises: *is it possible to further improve the upper bounds?* We provide an affirmative answer by developing an improved algorithm in the next section.

### 4.3 Our Improved Algorithm

Through careful analysis, we find that the suboptimal regret bounds are primarily caused by the bidirectional compression scheme. The server re-compresses the aggregated messages that have already been compressed by learners, leading to an amplification of the compression error. In fact, if we apply the *unidirectional compression*, D-FTCL can obtain the optimal regret bounds with the detailed proof provided in Appendix C.9. Thus, to reduce the regret of D-FTCL, we need to control the error caused by compression. In offline optimization, Huang et al. [2022] design the fast compressed communication (FCC) technique to reduce the compression error, which recursively compresses the residual over $L$ rounds. This multi-round approach is not allowed in the online setting, where learners must update their decisions instantaneously to cope with the streaming data.

To resolve this dilemma, we utilize the online compression strategy [Yang et al., 2026]. By dividing the total $T$ rounds into blocks of size $L$ and updating the decision only at the end of each block, we can effectively distribute the communication cost of each update over the entire block. At round $t \in [(b-1)L + 1, bL]$ within the $b$-th block, each learner $i$ plays the fixed decision $\mathbf{w}^b$ and receives

gradient $\mathbf{g}_i^t$. Simultaneously, it transmits the compressed information to the server by performing FCC. Then it updates the local error $\mathbf{e}_i^b$. The server collects the aggregated information at the end of block $b$. In the subsequent block, server applies the EF and employs the online compression strategy to disseminate the compressed information back to the learners.

However, the block-wise transmission mechanism inherently introduces new challenges: since transmitting information between learners and the server requires $L$ rounds, the communication inevitably becomes *asynchronous* due to transmission latency. In particular, after learners send compressed information to the server through $L$ rounds, it takes another $L$ rounds for transmitting back the aggregated message. As a result, learner $i$ only receives the compressed information $\mathbf{s}^{b-2}$ of block $b-2$ at the end of block $b$. Reflecting the information lag, the learners update the decision using only the available information up to block $b-2$:

$$\mathbf{w}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{s}^k, \mathbf{w} \right\rangle + \frac{1}{\eta} \|\mathbf{w}\|^2 . \tag{6}$$

If the loss is $\mu$-strongly convex, learners update as follows:

$$\mathbf{w}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{s}^k, \mathbf{w} \right\rangle + \frac{\mu L}{2} \sum_{k=1}^{b} \left\|\mathbf{w} - \mathbf{w}^k\right\|^2 . \tag{7}$$

For the sake of clarity, we present a simplified version of our algorithm, termed Distributed Follow-the-Fast-Compressed-Leader (D-FTFCL), in Algorithm 3, with the comprehensive details provided in Appendix C.5. The asynchronous updates pose challenges to the theoretical analysis. To address this, we explicitly characterize the impact of the delayed information. In the following, we establish the guarantees for D-FTFCL.

**Theorem 4.5.** *Under the Assumptions 3.1, 3.3 and 3.4, by setting the learning rate* $\eta = \frac{D}{G\sqrt{LT}}$ *and the block size* $L = \lceil \frac{1}{\delta} \rceil$*, Algorithm 3 can ensure*

$$\mathbb{E}_{\mathcal{C}} [R_T] \leq O(\delta^{-1/2} DG\sqrt{T}).$$

**Theorem 4.6.** *Under the Assumptions 3.2, 3.3 and 3.4, by setting the block size* $L = \lceil \frac{1}{\delta} \rceil$*, Algorithm 3 can ensure*

$$\mathbb{E}_{\mathcal{C}} [R_T] \leq O\left(\mu^{-1}\delta^{-1}(G+\mu D)^2 \log T\right).$$

**Remark.** Our improved algorithm enjoys $O(\delta^{-1/2}\sqrt{T})$ and $O\left(\delta^{-1}\log T\right)$ regret bounds for convex and strongly convex loss functions, which match the lower bounds in Theorems 4.3 and 4.4.

**Remark.** Our theoretical guarantees require setting the block size to $L = \lceil 1/\delta \rceil$ to adequately suppress the compression error. It is worth noting that this choice is derived from the worst-case analysis. In practice, we can employ a smaller $L$ in real-world implementations.

**Remark.** Note that the upper regret bounds of D-OCO algorithms generally hold for all compressors satisfying Definition 3.6. Therefore, although lower bounds in Theorems 4.3 and 4.4 are constructed based on a specific compressor instantiation, they are sufficient to prove the tightness of the upper bounds in general.

## 5 Application to Distributed Stochastic Convex Optimization

In this section, we extend our online algorithm to distributed stochastic convex non-smooth optimization with domain constraints. Formally, we consider the following offline distributed optimization problem [Nedić and Ozdaglar, 2009]

$$\min_{\mathbf{x}\in\mathcal{W}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{8}$$

where $f_i(\cdot) : \mathcal{W} \to \mathbb{R}$ is the loss function of the learner $i$. Here, learners are assumed to have access solely to stochastic estimates of the function value $f_i(\cdot;\xi_i)$ and the local subgradient with sample $\xi_i$ from distribution $\mathcal{D}_i$. Following the previous work [Cutkosky, 2019, Islamov et al., 2025], we introduce the standard assumption outlined below.

**Algorithm 4** Distributed O2B Conversion with Compressed Communication

**Require:** Online method $\mathcal{A}$, weights $\alpha_1, \ldots, \alpha_{T/L}$, compression round $L$, communication round $T$

Initialize $\mathbf{w}^1 = \mathbf{0}$, $\mathcal{A}$ as Algorithm 2, compressor as Algorithm 1
**for** $t = 1$ to $T/L$ **do**
  **for each learner** $i = 1, ..., n$ **in parallel do**
    Compute the decision $\mathbf{x}^t = \frac{\sum_{k=1}^{t} \alpha_k \mathbf{w}^k}{\alpha_{1:t}}$ and query subgradient $\mathbf{g}_i^t$ at $\mathbf{x}^t$
    Construct the surrogate loss $\hat{\ell}_i^t(\mathbf{w})$ according to (9) or (10)
    Send $\hat{\ell}_i^t(\mathbf{w})$ to $\mathcal{A}$ as $t$-th loss
    Receive the updated decision $\mathbf{w}^{t+1}$ from $\mathcal{A}$
    Online method $\mathcal{A}$ updates the decision according to (4) or (11)
  **end for**
**end for**
**Return**: $\mathbf{x}^{T/L}$

**Assumption 5.1** (Stochastic subgradient)**.** The stochastic subgradient $\mathbf{g}_i^t$ of function $f_i(\cdot)$ with random sample $\xi_i^t$ at point $\mathbf{x}^t$ satisfies $\mathbb{E}[\mathbf{g}_i^t] \in \partial f_i(\mathbf{x}^t)$ and $\|\mathbf{g}_i^t\|^2 \leq G^2$.

As discussed in Section 2.1, existing studies on distributed optimization typically assume either smooth loss functions or an unconstrained optimization domain. There still remains a critical gap: no work has addressed distributed non-smooth optimization with compressed communication and domain constraints, despite its extensive applications in machine learning tasks [Agarwal et al., 2006, Hazan, 2016]. While a recent work by Islamov et al. [2025] investigates convex non-smooth optimization, it is limited to the specific setting of safe constraints. Crucially, their methods cannot handle standard domain constraints. Hence, a fundamental question arises: *is it possible to design an algorithm for distributed optimization with compressed communication and domain constraints?*

We provide an affirmative answer for this question by extending our online methods to offline stochastic optimization. An effective approach to solving the stochastic convex optimization problem is to leverage online learning algorithms through the anytime online-to-batch (O2B) conversion [Cesa-Bianchi et al., 2004, Shalev-Shwartz, 2012, Cutkosky, 2019], which constructs a sequence of surrogate loss functions using stochastic gradients and applies online learning algorithms to generate decisions. By minimizing the *regret*, we can achieve the optimal convergence rates for stochastic convex optimization problem.

In this work, we propose a distributed variant of the anytime O2B conversion [Cutkosky, 2019], as summarized in Algorithm 4. To mitigate the compression error in bidirectional compression, we again adopt the FCC technique [Huang et al., 2022]. A key distinction from the online setting is that local learners can communicate with the central server multiple times per update in the offline setting; it eliminates the need for the blocking update technique. Consequently, in the anytime O2B conversion, we can employ our D-FTCL (Algorithm 2) as the online method $\mathcal{A}$ and directly replace the standard compressor $\mathcal{C}(\cdot)$ with the FCC mechanism (Algorithm 1). At each round $t$, each learner computes a weighted average decision $\mathbf{x}^t = \frac{\sum_{k=1}^{t} \alpha_k \mathbf{w}^k}{\alpha_{1:t}}$ and queries the local subgradient $\mathbf{g}_i^t$ at $\mathbf{x}^t$. Then, it constructs the surrogate loss according to:

$$\hat{\ell}_i^t(\mathbf{w}) = \langle \alpha_t \mathbf{g}_i^t, \mathbf{w} \rangle, \tag{9}$$

and feeds it to D-FTCL as the $t$-th loss function. Then, the online method updates the decision according to (4). Let $T$ denote the total number of communication rounds. Since the FCC mechanism requires $L$ rounds per update, each learner only performs $K = T/L$ updates. We now establish the *last-iterate* convergence rate for Algorithm 4.

**Theorem 5.2.** *Under the Assumptions 3.1, 3.4 and 5.1, configured with the online method $\mathcal{A}$ as Algorithm 2, the compression method as Algorithm 1, $\alpha_t = 1$, $\hat{\ell}_i^t(\mathbf{w})$ according to (9), $\mathbf{w}^{t+1}$ according to (4), $\eta = \frac{D}{G\sqrt{K}}$ and compression round $L = \lceil 1/\delta \rceil$, for all $\mathbf{x} \in \mathcal{W}$, Algorithm 4 can ensure*

$$\mathbb{E}_{\mathcal{C}} \left[ f(\mathbf{x}^K) - f(\mathbf{x}) \right] \leq O\left( \frac{DG}{\sqrt{\delta T}} \right),$$

*where $T = KL$ is the total communication rounds of each local learner.*

It is not hard to verify that applying the uniform weighting scheme ($\alpha_t = 1$) to strongly convex functions only yields a convergence rate of $O(\delta^{-1}T^{-1}\log T)$, which is suboptimal with respect to $T$ [Agarwal et al., 2012]. Benefiting from the flexibility of the anytime O2B conversion, we can employ time-varying weights with $\alpha_t = t$ and construct the surrogate loss as:

$$\hat{\ell}_i^t(\mathbf{w}) = \langle \alpha_t \mathbf{g}_i^t, \mathbf{w} \rangle + \frac{\mu \alpha_t}{2} \left\| \mathbf{w} - \mathbf{x}^t \right\|^2. \tag{10}$$

Online method receives the subgradient $\alpha_t \mathbf{g}_i^t$ and updates the decision according to:

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w} \in \mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{s}^k, \mathbf{w} \right\rangle + \frac{\mu}{2} \sum_{k=1}^{t} \alpha_k \left\| \mathbf{w} - \mathbf{x}^k \right\|^2. \tag{11}$$

We formally state the convergence rate for strongly convex loss functions.

**Theorem 5.3.** *Under the Assumptions 3.2, 3.4 and 5.1, configured with the online method $\mathcal{A}$ as Algorithm 2, the compression method as Algorithm 1, $\alpha_t = t$, $\hat{\ell}_i^t(\mathbf{w})$ according to (10), $\mathbf{w}^{t+1}$ according to (11) and compression round $L = \lceil 1/\delta \rceil$, for all $\mathbf{x} \in \mathcal{W}$, Algorithm 4 can ensure*

$$\mathbb{E}_{\mathcal{C}} \left[ f(\mathbf{x}^K) - f(\mathbf{x}) \right] \leq O\left( \frac{(\mu D + G)^2}{\mu \delta T} \right),$$

*where $T = KL$ is the total communication rounds of each local learner.*

**Remark.** Islamov et al. [2025] establish a lower bound of $\Omega(\delta^{-1/2}T^{-1/2})$ for distributed convex non-smooth optimization in the unconstrained setting, where $T$ is the communication round. Our rate of $O(\delta^{-1/2}T^{-1/2})$ for convex objectives matches this lower bound. Furthermore, in the absence of domain and safety constraints, our method still achieves a more favorable rate than that of Islamov et al. [2025], demonstrating the inherent efficiency of our algorithmic design.

**Remark.** For $\mu$-strongly convex functions, our method achieves a convergence rate of $O(\delta^{-1}T^{-1})$, which is optimal with respect to $T$ [Agarwal et al., 2012]. Crucially, our results fill a significant gap in the literature by establishing the first convergence guarantees for distributed convex non-smooth optimization with compressed communication under domain constraints.

## 6 Conclusion and Future Work

In this paper, we study distributed online convex optimization (D-OCO) with compressed communication. Firstly, we propose a communication-efficient method that achieves the regret bounds of $O(\delta^{-1/2}\sqrt{T})$ and $O(\delta^{-1}\log T)$ for convex and strongly convex loss functions, respectively. Secondly, to certify the optimality of our algorithm, we establish lower bounds of $\Omega(\delta^{-1/2}\sqrt{T})$ and $\Omega(\delta^{-1}\log T)$ for convex and strongly convex loss functions, respectively. Finally, we extend our online methods to distributed offline stochastic optimization with domain constraints. We obtain the *first* convergence rates for both convex and strongly convex non-smooth objectives by developing a distributed anytime online-to-batch conversion. Experiments have validated the effectiveness and efficiency of our methods.

Two directions are worth future exploration. The first is to investigate whether our method can achieve tighter results for smooth loss functions. It is well-known that optimistic online learning methods [Rakhlin and Sridharan, 2013, Mohri and Yang, 2016] can leverage prior knowledge to construct the optimistic term and exploit smoothness to obtain better bounds. However, achieving such improvements in D-OCO with compressed communication presents significant technical challenges. Since each learner receives only compressed information per round, the compressed gradients alone are insufficient to construct effective optimistic terms. The second direction is to apply our online methods into distributed non-convex non-smooth optimization with compressed communication by using the online-to-non-convex conversion [Cutkosky et al., 2023].

## Acknowledgements

This work was partially supported by NSFC (62361146852), the "111 Center" (No. B26023), and the Fundamental Research Funds for the Central Universities (2026300271). Sifan Yang was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. 26CXJH0361). The authors would also like to thank the anonymous reviewers for their helpful comments.

## References


J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 415–423, 2008.

A. Agarwal, E. Hazan, S. Kale, and R. E. Schapire. Algorithms for portfolio management based on the newton method. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 560–569, 2018.

A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, pages 1–50, 2023.

X. Cao and T. Başar. Decentralized online convex optimization with compressed communications. *Automatica*, 156:111186, 2023.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):1–27, 2011.

M. Chen, R. Mathews, T. Ouyang, and F. Beaufays. Federated learning of out-of-vocabulary words. *arXiv preprint arXiv:1903.10635*, 2019.

A. Cutkosky. Anytime online-to-batch, optimism and acceleration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1446–1454, 2019.

A. Cutkosky, H. Mehta, and F. Orabona. Optimal, stochastic, non-smooth, non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, pages 592–606, 2011.

A. M. Elbir, B. Soner, and S. Coleri. Federated learning in vehicular networks. *arXiv preprint arXiv:2006.01412*, 2020.

I. Fatkhullin, I. Sokolov, E. A. Gorbunov, Z. Li, and P. Richtárik. EF21 with bells & whistles: Six algorithmic extensions of modern error feedback. *Journal of Machine Learning Research*, 26(189): 1–50, 2025.

Y. Gao, R. Islamov, and S. U. Stich. Econtrol: Fast distributed optimization with compression and error control. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Y. Gao, A. Rodomanov, J. Rack, and S. U. Stich. Composite optimization with error feedback: the dual averaging approach. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *SIAM Journal on Optimization*, pages 1493–1528, 2016.

E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtarik. Linearly converging error compensated sgd. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

E. Hazan and S. Kale. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1843–1850, 2012.

E. Hazan and S. Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

S. Hosseini, A. Chapman, and M. Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 1484–1489. IEEE, 2013.

X. Huang, Y. Chen, W. Yin, and K. Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 35:18955–18969, 2022.

X. Huang, P. Li, and X. Li. Stochastic controlled averaging for federated learning with communication compression. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

R. Islamov, Y. As, and I. Fatkhullin. Safe-ef: Error feedback for non-smooth constrained optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 26542–26585, 2025.

P. Jiang and G. Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Advances in neural information processing systems 31 (NeurIPS)*, 2018.

S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3252–3261, 2019.

F. K. Khan, A. Flanagan, K. E. Tan, Z. Alamgir, and M. Ammad-Ud-Din. A payload optimization method for federated recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 432–442, 2021.

A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3478–3487, 2019.

F. Liang, W. Pan, and Z. Ming. Fedrec++: Lossless federated recommendation with explicit feedback. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 4224–4231, 2021.

M. Mohri and S. Yang. Accelerating online convex optimization via adaptive prediction. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 848–856, 2016.

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

K. K. Patel, L. Wang, A. Saha, and N. Srebro. Federated online and bandit convex optimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 27439–27460, 2023.

X. Qian, P. Richtárik, and T. Zhang. Error compensated distributed sgd can be accelerated. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 30401–30413, 2021.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

P. Richtárik, I. Sokolov, and I. Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 4384–4396, 2021.

P. Richtárik, I. Sokolov, E. Gasanov, I. Fatkhullin, Z. Li, and E. Gorbunov. 3PC: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 18596–18648, 2022.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Conference of the International Speech Communication Association*, pages 1058–1062, 2014.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

C. Shi, C. Shen, and J. Yang. Federated multi-armed bandits with personalization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2917–2925, 2021.

S. U. Stich. On communication compression for distributed optimization on heterogeneous data. *arXiv preprint arXiv:2009.02388*, 2020.

S. U. Stich and S. P. Karimireddy. The error-feedback framework: Sgd with delayed gradients. *Journal of Machine Learning Research*, pages 1–36, 2020.

S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4447–4458, 2018.

B. Swenson, R. Murray, H. V. Poor, and S. Kar. Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima. *Journal of Machine Learning Research*, 23 (328):1–62, 2022.

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

Z. Tu, X. Wang, Y. Hong, L. Wang, D. Yuan, and G. Shi. Distributed online convex optimization with compressed communication. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 34492–34504, 2022.

Y. Wan. Black-box reductions for decentralized online convex optimization in changing environments. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, pages 5605–5631, 2025.

Y. Wan, W.-W. Tu, and L. Zhang. Projection-free distributed online convex optimization with $O(\sqrt{T})$ communication complexity. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9818–9828, 2020.

Y. Wan, G. Wang, W.-W. Tu, and L. Zhang. Projection-free distributed online learning with sublinear communication complexity. *Journal of Machine Learning Research*, 23(172):1–53, 2022.

Y. Wan, T. Wei, M. Song, and L. Zhang. Nearly optimal regret for decentralized online convex optimization. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT)*, pages 4862–4888, 2024.

Y. Wan, T. Wei, B. Xue, M. Song, and L. Zhang. Optimal and efficient algorithms for decentralized online convex optimization. *Journal of Machine Learning Research*, 26(135):1–43, 2025.

Y. Wang, Y. Wan, S. Zhang, and L. Zhang. Distributed projection-free online learning for smooth and convex losses. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 10226–10234, 2023.

J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, pages 2483–2493, 2013.

S. Yang, W. Yang, W. Jiang, and L. Zhang. Distributed online convex optimization with efficient communication: Improved algorithm and lower bounds. *arXiv preprint arXiv:2601.04907*, 2026.

K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 4035–4043, 2017a.

W. Zhang, P. Zhao, W. Zhu, S. C. H. Hoi, and T. Zhang. Projection-free distributed online learning in networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 4054–4062, 2017b.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions and scope are clearly written in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Appendix A.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Refer to Section Preliminaries and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the experimental results is provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to privacy concerns, we do not include the code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See the Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is mainly theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: This work is mainly theoretical.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the creators of the dataset used in our experiments properly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: We just use LLMs to polish this paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

# Appendix

## A Limitations

Our work focuses on convex and non-smooth losses. We believe this setting is important in its own right, since convex online learning provides a framework for understanding distributed decision-making problem with theoretical guarantees. Nevertheless, some practically important learning problems are non-convex, which fall outside the scope of our current analysis. In addition, we assume that each learner and the central server communicate exactly once per round. This is a standard assumption in distributed online optimization, but real-world systems may also involve communication latency and stale information.

## B Additional Discussions on Related Work

### B.1 Decentralized Online Convex Optimization

We also notice a parallel line of research investigating decentralized online convex optimization, where local learners interact over a network modeled by an undirected graph [Yan et al., 2013, Hosseini et al., 2013, Zhang et al., 2017b, Wan et al., 2020, 2022, Wang et al., 2023, Wan et al., 2024, 2025, Wan, 2025]. To reduce the communication cost among learners, recent efforts have further explored compressed communication within this setting [Tu et al., 2022, Cao and Başar, 2023, Yang et al., 2026]. However, in decentralized settings, each learner maintains an individual local decision and communicates in a peer-to-peer scheme without a central server for aggregation. This architecture fundamentally differs from the distributed setting we consider. Consequently, algorithms tailored for peer-to-peer topologies cannot be directly extended to our setting due to intrinsic differences in communication protocols and network structures.

### B.2 Compressed Communication

To reduce the communication cost in distributed learning, lots of work has investigated communication compression methods. Seide et al. [2014] propose SignSGD, which transmits only the sign of the information. Subsequent work [Stich et al., 2018, Wangni et al., 2018] transmits only a small number of entries to achieve sparsity in communication.

Stich et al. [2018] first study EF and use EF to develop sparsified SGD. The later work [Gorbunov et al., 2020, Qian et al., 2021] provides the analysis for unbiased compressors under the convex smooth setting. More recently, Gao et al. [2024] extend EF to develop EControl in both smooth convex and non-convex settings. In addition, Richtárik et al. [2022] establish the analysis of EF21 for 3PC compressors, which encompasses contractive compressors as a special case. Huang et al. [2024] further investigate EF in distributed optimization with local updates.

### B.3 Compressor

To alleviate communication overhead, extensive research has integrated compression mechanisms into distributed offline optimization algorithms [Seide et al., 2014, Stich et al., 2018, Bernstein et al., 2018, Karimireddy et al., 2019, Huang et al., 2022, Islamov et al., 2025]. Rather than broadcasting the full vector $\mathbf{x} \in \mathbb{R}^d$, they transmit compressed data $\mathcal{C}(\mathbf{x})$, where $\mathcal{C}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is an operator designed for efficient transmission. A quintessential example is the sign operator, which outputs the sign of each coordinate, thereby significantly reducing the communication cost.

Generally, compressors can be categorized into two classes: unbiased compressors [Jiang and Agrawal, 2018, Tang et al., 2018, Zhang et al., 2017a] and contractive compressors [Seide et al., 2014, Stich et al., 2018, Richtárik et al., 2022, Beznosikov et al., 2023]. An unbiased compressor satisfies $\mathbb{E}[\mathcal{C}(\mathbf{x})] = \mathbf{x}$ and $\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \omega \|\mathbf{x}\|$ for any input $\mathbf{x} \in \mathbb{R}^d$ and $\omega \geq 0$. The contractive compressor satisfies $\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2\right] \leq (1-\delta)\|\mathbf{x}\|$ and $\delta \in (0, 1]$.

# C Proof of Theorems

## C.1 Proof of Theorem 4.1

First, we introduce several notations. We denote $f_t(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i^t(\mathbf{w})$, $\mathbf{v}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t$ and $\mathbf{g}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i^t$. For $\hat{\mathbf{e}}^k$ and $\mathbf{e}^t$, by summing up from $k=1$ to $t$, we have

$$\mathbf{e}^t = \sum_{k=1}^{t-1} \mathbf{g}^k - \mathbf{v}^k, \qquad \hat{\mathbf{e}}^t = \sum_{k=1}^{t-1} \mathbf{v}^k - \mathbf{s}^k.$$

Then, we define several virtual decisions

$$\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{g}^k, \mathbf{w} \right\rangle + \frac{1}{\eta}\|\mathbf{w}\|^2, \quad \mathbf{y}^{t+1} = \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{1}{\eta}\|\mathbf{y}\|^2.$$

Note that $\hat{\mathbf{w}}^{t+1}$ represents the decision updated with the exact gradient, whereas $\mathbf{y}^{t+1}$ represents the decision updated with information that learners transmit to the server. Then we introduce some useful lemmas.

**Lemma C.1** (Lemma 6.6 in Garber and Hazan [2016]). *Let $\{\ell_t(\mathbf{w})\}_{t=1}^T$ be a sequence of functions and $\mathbf{w}_t^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{k=1}^t \ell_k(\mathbf{w})$ for any $t \in [T]$. Then, it holds that*

$$\sum_{t=1}^T \ell_t(\mathbf{w}_t^*) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^T \ell_t(\mathbf{w}) \le 0.$$

**Lemma C.2** (Lemma 5 in Duchi et al. [2011]). *Let $\Pi_{\mathcal{W}}(\mathbf{u},\eta) = \arg\min_{\mathbf{w}\in\mathcal{W}} \langle \mathbf{u}, \mathbf{w}\rangle + \frac{1}{\eta}\|\mathbf{w}\|^2$. For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we have*

$$\|\Pi_{\mathcal{W}}(\mathbf{u},\eta) - \Pi_{\mathcal{W}}(\mathbf{v},\eta)\| \le \frac{\eta}{2}\|\mathbf{u}-\mathbf{v}\|.$$

Now, we are ready to derive the regret bound.

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] &= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n f_i^t(\mathbf{w}^t) - \frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n f_i^t(\mathbf{w})\right] \\
&\le \frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}_i^t, \mathbf{w}^t - \mathbf{w}\rangle\right] \\
&= \sum_{t=1}^T \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \mathbf{w}\rangle\right] \\
&\le \sum_{t=1}^T \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t + \hat{\mathbf{w}}^t - \mathbf{w}\rangle\right] \\
&= \sum_{t=1}^T \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle + \langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t\rangle\right] \\
&\le \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle\right]}_{\texttt{TERM-A}} + \underbrace{\sum_{t=1}^T G\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|\right]}_{\texttt{TERM-B}},
\end{aligned}$$

where the last inequality is due to $\langle \mathbf{x}, \mathbf{y}\rangle \le \|\mathbf{x}\|\,\|\mathbf{y}\|$ and $\|\mathbf{g}^t\| = \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{g}_i^t\right\| \le G$. Remarkably, `TERM-A` is the regret bound of the decisions updated by using the uncompressed gradients and `TERM-B` is the error of the compression. In the following, we give the bound of this two term, respectively.

As for TERM-A, we have

$$\begin{aligned}\sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle &= \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1} + \hat{\mathbf{w}}^{t+1} - \mathbf{w}\rangle \\ &= \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\rangle + \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^{t+1} - \mathbf{w}\rangle \\ &\leq \sum_{t=1}^{T} G\left\|\hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\right\| + \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^{t+1} - \mathbf{w}\rangle.\end{aligned}$$

We define $\ell_1(\mathbf{w}) = \langle \mathbf{g}^1, \mathbf{w}\rangle + \frac{1}{\eta}\|\mathbf{w}\|^2$ and $\ell_t(\mathbf{w}) = \langle \mathbf{g}^t, \mathbf{w}\rangle$. We can verify that $\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{k=1}^{t} \ell_k(\mathbf{w})$. By using Lemma C.1, we can derive

$$\sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^{t+1} - \mathbf{w}\rangle + \frac{\left\|\hat{\mathbf{w}}^2\right\|^2 - \|\mathbf{w}\|^2}{\eta} \leq \sum_{t=1}^{T} \ell_t(\hat{\mathbf{w}}^{t+1}) - \ell_t(\mathbf{w}) \leq 0. \tag{12}$$

Therefore, we can obtain

$$\begin{aligned}\sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle &= \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\rangle + \sum_{t=1}^{T}\langle \mathbf{g}^t, \hat{\mathbf{w}}^{t+1} - \mathbf{w}\rangle \\ &\leq \sum_{t=1}^{T}\left\|\mathbf{g}^t\right\|\left\|\hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\right\| + \frac{\|\mathbf{w}\|^2 - \left\|\hat{\mathbf{w}}^2\right\|^2}{\eta} \\ &\leq \frac{\eta G^2 T}{2} + \frac{D^2}{\eta},\end{aligned} \tag{13}$$

where the first inequality is due to (12).

In the following, we give the bound of TERM-B.

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \mathbf{y}^t + \mathbf{y}^t - \hat{\mathbf{w}}^t\right\|\right] \\ &\leq \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \mathbf{y}^t\right\|\right] + \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{y}^t - \hat{\mathbf{w}}^t\right\|\right] \\ &\leq \frac{\eta}{2}\left(\mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t-1}\mathbf{s}^k - \mathbf{v}^k\right\|\right] + \mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t-1}\mathbf{v}^k - \mathbf{g}^k\right\|\right]\right) \\ &= \frac{\eta}{2}\left(\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}^t\right\|\right] + \mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^t\right\|\right]\right),\end{aligned}$$

where the second inequality is due to Lemma C.2. To give the bound of errors, we introduce the following lemma.

**Lemma C.3.** *Under the Assumptions 3.3 and 3.4, at any round $t$ of D-FTCL, the norm of the errors is bounded by*

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}^{t+1}\right\|^2\right] \leq \frac{4(1-\delta)G^2}{\delta^2}, \quad \mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] \leq \frac{160(1-\delta)G^2}{\delta^4}.$$

By using Lemma C.3, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|\right] &\leq \frac{\eta}{2}\left(\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t-1}\right\|\right] + \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}^{t-1}\right\|\right]\right) \\ &\leq \frac{\eta}{2}\left(\frac{13\sqrt{1-\delta}G}{\delta^2} + \frac{2\sqrt{1-\delta}G}{\delta}\right) \leq \frac{15\eta G}{2\delta^2}.\end{aligned}$$

By setting the learning rate $\eta = \frac{\delta D}{G\sqrt{T}}$, the final regret bound is

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] &\leq \sum_{t=1}^{T} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle\right] + \mathbb{E}_{\mathcal{C}}\left[G\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|\right] \\
&\leq \frac{\eta G^2 T}{2} + \frac{D^2}{\eta} + \frac{15\eta T G^2}{2\delta^2} \\
&\leq \frac{DG\delta\sqrt{T}}{2} + \frac{DG\sqrt{T}}{\delta} + \frac{15DG\sqrt{T}}{2\delta} \\
&\leq O(\delta^{-1}DG\sqrt{T}).
\end{aligned}$$

### C.2 Proof of Theorem 4.2

Similar to the previous proof, we define two decisions.

$$\begin{aligned}
\mathbf{y}^{t+1} &= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{\mu}{2}\sum_{k=1}^{t}\left\|\mathbf{y} - \mathbf{w}^k\right\|^2 \\
&= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k - \mu\mathbf{w}^k, \mathbf{y} \right\rangle + \frac{\mu t}{2}\left\|\mathbf{y}\right\|^2.
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{w}}^{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{g}^k, \mathbf{w} \right\rangle + \frac{\mu}{2}\sum_{k=1}^{t}\left\|\mathbf{w} - \mathbf{w}^k\right\|^2 \\
&= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{g}^k - \mu\mathbf{w}^k, \mathbf{w} \right\rangle + \frac{\mu t}{2}\left\|\mathbf{w}\right\|^2.
\end{aligned}$$

We begin to derive the regret bound, where the key difference is to utilize the strong convexity to improve the dependence on $T$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] &= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w}^t) - f_i^t(\mathbf{w})\right] \\
&\leq \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}_i^t, \mathbf{w}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t + \hat{\mathbf{w}}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t\rangle\right] + \mathbb{E}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right] \\
&\leq \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right]}_{\texttt{TERM-A}} + \underbrace{\sum_{t=1}^{T} G\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|}_{\texttt{TERM-B}},
\end{aligned}$$

where the first inequality is due to Assumption 3.2 and the last inequality is due to $\langle \mathbf{x}, \mathbf{y}\rangle \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$.

We first give the bound of TERM-A. We define $\ell_t(\mathbf{w}) = \langle \mathbf{g}^t, \mathbf{w} \rangle + \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^t\|^2$ and $\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{k=1}^{t} \ell_k(\mathbf{w})$ in the strongly convex case. Then, it is not hard to verify that

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w} \rangle - \frac{\mu}{2} \left\|\mathbf{w}^t - \mathbf{w}\right\| \\
=&\sum_{t=1}^{T} \langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1} + \hat{\mathbf{w}}^{t+1} - \mathbf{w} \rangle - \frac{\mu}{2} \left\|\mathbf{w}^t - \mathbf{w}\right\| \\
\leq&\sum_{t=1}^{T} \ell_t(\hat{\mathbf{w}}^{t+1}) - \ell_t(\mathbf{w}) + G \left\|\hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\right\| \\
\leq&\sum_{t=1}^{T} G \left\|\hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\right\|.
\end{aligned} \tag{14}$$

Next, we bound the term $G \left\|\hat{\mathbf{w}}^t - \hat{\mathbf{w}}^{t+1}\right\|$. We define $F_t(\mathbf{w}) = \sum_{k=1}^{t} \ell_k(\mathbf{w})$, and $F_t(\mathbf{w})$ is a $(t\mu)$-strongly convex function and $\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} F_t(\mathbf{w})$. We introduce the following lemma from Hazan and Kale [2012].

**Lemma C.4.** *For any $\mu$-strongly convex function $f(\mathbf{x}) : \mathcal{W} \to \mathbb{R}$ and any $\mathbf{x} \in \mathcal{W}$, it holds that*

$$\frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*), \tag{15}$$

*where $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{W}} f(\mathbf{x})$.*

For any $\mathbf{x}, \mathbf{y} \in \mathcal{W}$, we have

$$\begin{aligned}
|\ell_t(\mathbf{x}) - \ell_t(\mathbf{y})| &\leq |\langle \nabla \ell_t(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle| \\
&\leq \|\nabla \ell_t(\mathbf{x})\| \, \|\mathbf{x} - \mathbf{y}\| \\
&\leq (G + \mu D) \, \|\mathbf{x} - \mathbf{y}\| .
\end{aligned}$$

By using Lemma C.4, we can derive

$$\begin{aligned}
G \left\|\hat{\mathbf{w}}^{t+1} - \hat{\mathbf{w}}^t\right\|^2 \leq& \frac{2G}{t\mu}(F_t(\hat{\mathbf{w}}^t) - F_t(\hat{\mathbf{w}}^{t+1})) \\
\leq& \frac{2G}{t\mu}(F_{t-1}(\hat{\mathbf{w}}^t) - F_{t-1}(\hat{\mathbf{w}}^{t+1}) + \ell_t(\hat{\mathbf{w}}^t) - \ell_t(\hat{\mathbf{w}}^{t+1})) \\
\leq& \frac{2G}{t\mu}(\ell_t(\hat{\mathbf{w}}^t) - \ell_t(\hat{\mathbf{w}}^{t+1})) \\
\leq& \frac{2G}{t\mu}(G + \mu D) \left\|\hat{\mathbf{w}}^{t+1} - \hat{\mathbf{w}}^t\right\|,
\end{aligned}$$

which means

$$G \left\|\hat{\mathbf{w}}^{t+1} - \hat{\mathbf{w}}^t\right\| \leq \frac{2G}{t\mu}(G + \mu D). \tag{16}$$

By using Lemma C.2 and Lemma C.3, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}} \left[\left\|\mathbf{w}^{t+1} - \hat{\mathbf{w}}^{t+1}\right\|\right] &\leq \frac{1}{t\mu} \left(\mathbb{E}_{\mathcal{C}} \left[\left\|\hat{\mathbf{e}}^t\right\|\right] + \mathbb{E}_{\mathcal{C}} \left[\left\|\mathbf{e}^t\right\|\right]\right) \\
&\leq \frac{1}{t\mu} \left(\frac{13\sqrt{1-\delta}G}{\delta^2} + \frac{2\sqrt{1-\delta}G}{\delta}\right) \leq \frac{15G}{t\mu\delta^2}.
\end{aligned}$$

Finally, we can derive the regret bound

$$\mathbb{E}_{\mathcal{C}} \left[R_T\right] \leq \sum_{t=1}^{T} \frac{2G}{t\mu}(G + \mu D) + \frac{15G^2}{t\mu\delta^2} \leq \frac{15(G + \mu D)^2}{\mu\delta^2} \log T = O(\delta^{-2} \log T).$$

### C.3 Proof of Theorem 4.3

In this section, we present the proof of the lower bound for D-OCO with compressed communication. We follow Huang et al. [2022] and assume two protocols: (i) all learners can only communication with the central server and cannot exchange information with one another; (ii) all updates are synchronized, which means all learners update their decisions using the same information received from the central server, ensuring that the decisions across all learners are consistent.

To characterize the effect of the compressor, we employ the idea of maximizing communication delay by using a specific compressor [Huang et al., 2022]. In our construction, we consider the Random Gossip compressor $\mathcal{C}(\cdot)$ [Koloskova et al., 2019], which outputs $Q(\mathbf{x}) = \mathbf{x}$ with probability $\delta \in (0, 1]$ and $Q(\mathbf{x}) = \mathbf{0}$ otherwise. Under this scheme, the expected number of communication rounds required for a learner to successfully transmit information to the central server is $\lceil 1/\delta \rceil$. We assume the server does not use the compressor, which is unidirectional compression. As noted by Huang et al. [2022], lower bounds for bidirectional compression are at least as large as those for unidirectional compression. To simplify the analysis, we consider unidirectional compression where only learners compress their communications, while the server transmits without compression.

Let $K = \lceil 1/\delta \rceil, Z = \lfloor (T-1)/K \rfloor, c_0 = 0$ and $c_{Z+1} = T$. The total $T$ rounds can be divided into the following $Z+1$ intervals $[c_0+1,\, c_1]$, $[c_1+1,\, c_2]$, $\ldots$, $[c_Z+1,\, c_{Z+1}]$. To maximize the impact of the communication on the regret, we set the losses of each learner $i$ is same in each interval.

Specifically, we assume $m = \lfloor n/2 \rfloor$ and set the loss functions as

$$f_1^t(\mathbf{w}) = \cdots = f_m^t(\mathbf{w}) = 0,$$

for $t \in [c_i + 1,\, c_{i+1}], i \in \{0, ..., Z\}$, while the other loss functions are set $f_{m+1}^t(\mathbf{w}) = \cdots = f_n^t(\mathbf{w}) = h_i(\mathbf{w})$. In this way, the global loss function is

$$f_t(\mathbf{w}) = \frac{n-m}{n} h_i(\mathbf{w}).$$

We independently select $h_i(\mathbf{w}) = \langle \mathbf{z}_i, \mathbf{w} \rangle$ for any $i \in \{0, ..., Z\}$, where the coordinates of $\mathbf{z}_i$ are $\pm G/\sqrt{d}$ with probability 0.5 and the feasible domain $\mathcal{X} = [-D/2\sqrt{d}, D/2\sqrt{d}]^d$. According to the above discussion, the decisions $\mathbf{w}^t$ for any $t \in [c_i + 1,\, c_{i+1}], i \in \{0, ..., Z\}$ are made before the function $h_i(\mathbf{w})$ can be revealed to the local learner. Then we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] =& \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w}^t) - \min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w})\right] \\
=& \frac{n-m}{n}\mathbb{E}_{\mathcal{C}}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} h_i(\mathbf{w}^t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} h_i(\mathbf{w})\right] \\
=& \frac{n-m}{n}\mathbb{E}_{\mathcal{C}}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} \langle \mathbf{z}_i, \mathbf{w}^t \rangle - \min_{\mathbf{w}\in\mathcal{W}} \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} \langle \mathbf{z}_i, \mathbf{w} \rangle\right] \\
=& -\frac{n-m}{n}\mathbb{E}_{\mathcal{C}}\left[\min_{\mathbf{w}\in\mathcal{W}} \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} \langle \mathbf{z}_i, \mathbf{w} \rangle\right] \\
=& -\frac{n-m}{n}\mathbb{E}_{\mathcal{C}}\left[\min_{\mathbf{w}\in\mathcal{W}} \langle \sum_{i=0}^{Z}(c_{i+1} - c_i)\mathbf{z}_i, \mathbf{w} \rangle\right],
\end{aligned} \tag{17}$$

where the fourth equality is due to $\mathbb{E}[\langle \mathbf{z}_i, \mathbf{w}^t \rangle] = 0$ for $\forall t \in [c_i + 1, c_{i+1}]$.

Then, we denote $\epsilon_{01}, ..., \epsilon_{0d}, ..., \epsilon_{Z1}, ..., \epsilon_{Zd}$ be the coordinates of $\mathbf{z}_1, ..., \mathbf{z}_Z$, which are identically distributed variables with $\mathbb{P}(\epsilon_{ij} = \pm 1) = 1/2$ for $i \in \{0, ..., Z\}$ and $j \in \{1, ..., d\}$. By using the

Khintchine inequality on (17), we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] =& -\frac{n-m}{n}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Zd}}\left[\sum_{j=1}^{d}-\frac{D}{2\sqrt{d}}\left|\sum_{i=1}^{Z}(c_{i+1}-c_i)\frac{\epsilon_{ij}G}{\sqrt{d}}\right|\right] \\
=& \frac{n-m}{n}\frac{DG}{2}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Zd}}\left[\left|\sum_{i=0}^{Z}(c_{i+1}-c_i)\epsilon_{i1}\right|\right] \\
\geq& \frac{n-m}{n}\frac{DG}{2\sqrt{2}}\sqrt{\sum_{i=0}^{Z}(c_{i+1}-c_i)^2} \\
\geq& \frac{n-m}{n}\frac{DG}{2\sqrt{2}}\sqrt{\frac{(c_{Z+1}-c_0)^2}{Z+1}} \\
\geq& \frac{DGT}{4\sqrt{2Z+2}} \\
\geq& \frac{DGT}{4\sqrt{2\delta T+2-2\delta}},
\end{aligned} \tag{18}$$

where the second inequality is due to the Cauchy-Schwarz inequality.

If $2-2\delta \leq 2\delta T$, we can obtain

$$\mathbb{E}_{\mathcal{C}}\left[R_T\right] \geq \frac{DGT}{4\sqrt{2\delta T+2-2\delta}} \geq \frac{DG\sqrt{T}}{8\sqrt{\delta}}. \tag{19}$$

### C.4 Proof of Theorem 4.4

This proof is similar to the proof of Theorem 4.4. The key modification is to construct new loss functions to utilize the strong convexity and we choose the feasible domain $\mathcal{W} = [0, D/\sqrt{d}]^d$

Specifically, we set the loss functions as

$$f_1^t(\mathbf{w}) = \cdots = f_m^t(\mathbf{w}) = \frac{\mu}{2}\left\|\mathbf{w}\right\|^2,$$

for $t \in [c_i+1,\ c_{i+1}], i \in \{0,...,Z\}$, while the other loss functions are set

$$f_{m+1}^t(\mathbf{w}) = \cdots = f_n^t(\mathbf{w}) = \frac{\mu}{2}\left\|\mathbf{w}-\frac{D\mathbf{z}_i}{\sqrt{d}}\right\|^2,$$

where $\mathbf{z}_i$ is sampled from the distribution: $\Pr(\mathbf{z}_i = \mathbf{1}) = p$, $\Pr(\mathbf{z}_i = \mathbf{0}) = 1-p$ and $p \in [0,1]$. In this way, the global loss function in round $t \in [c_i+1, c_{i+1}], i \in \{0,...,Z\}$ is

$$\begin{aligned}
f_t(\mathbf{w}) =& \frac{\mu m}{2n}\left\|\mathbf{w}\right\|^2 + \frac{\mu(n-m)}{2n}\left\|\mathbf{w}-\frac{D\mathbf{z}_i}{\sqrt{d}}\right\|^2 \\
=& \frac{\mu}{2}\left\|\mathbf{w}\right\|^2 + \frac{\mu(n-m)D^2}{2nd}\left\|\mathbf{z}_i\right\|^2 - \frac{\mu(n-m)D}{n\sqrt{d}}\langle\mathbf{w},\mathbf{z}_i\rangle.
\end{aligned}$$

By taking the expectation on $\mathbf{z}_i$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}_i}[f_t(\mathbf{w})] =& \frac{\mu}{2}\left\|\mathbf{w}\right\|^2 + \frac{\mu(n-m)D^2}{2nd}\langle\mathbf{1},\mathbf{p}\rangle - \frac{\mu(n-m)D}{\sqrt{d}}\langle\mathbf{w},\mathbf{p}\rangle \\
=& \frac{\mu}{2}\left\|\mathbf{w}-\frac{(n-m)D\mathbf{p}}{n\sqrt{d}}\right\| + \frac{\mu(n-m)D^2}{2nd}\left\langle\mathbf{1}-\frac{n-m}{n}\mathbf{p},\mathbf{p}\right\rangle,
\end{aligned}$$

where $\mathbf{p} = [p,...,p]^\top \in \mathbb{R}^d$. We denote $H(\mathbf{w}) = \mathbb{E}_{\mathbf{z}_i}[f_i(\mathbf{w})]$ and $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \frac{(n-m)D\mathbf{p}}{n\sqrt{d}} \in \mathcal{W}$, which implies $H(\mathbf{w}) - H(\mathbf{w}^*) \geq 0$. Then we derive the lower bound for strongly convex loss functions.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] =& \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_t^i(\mathbf{w}^t) - \min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_t^i(\mathbf{w})\right] \\
=& \mathbb{E}_{\mathcal{C}}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} h_i(\mathbf{w}^t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} h_i(\mathbf{w})\right] \\
=& \mathbb{E}_{\mathcal{C}}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w})\right] \\
\geq& \mathbb{E}_{\mathcal{C}}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^t) - \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^*)\right],
\end{aligned} \tag{20}
$$

where the last inequality is due to $\mathbb{E}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^*)\right] \geq \mathbb{E}\left[\min_{\mathbf{w}\in\mathcal{W}}\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w})\right]$.

Now, our goal is to derive a lower bound of the right term of (20). We introduce a lemma from the proof of Theorem 4 in Wan et al. [2025], which establish the lower bound for decentralized online convex optimization.

**Lemma C.5.** *For an online convex optimization with the domain $\mathcal{W} = [0, D/\sqrt{d}]^d$, the length of interval is $K$ and the interval number $Z = \lfloor (T-1)/K \rfloor$, $M = \lfloor \log_{16}(15Z+16) - 1 \rfloor \geq 1$, the lower bound of (20) is*

$$
\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^t) - \sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}} H(\mathbf{w}^*) \geq \frac{\mu M K (n-m)^2 D^2}{16^4 n^2}.
$$

By using lemma C.5 and setting $16\delta^{-1} + 1 \leq T$, we can obtain the following lower bound

$$
\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] \geq& \frac{\mu M K (n-m)^2 D^2}{16^4 n^2} \geq \frac{\mu D^2 (\log_{16}(15Z+16) - 2)}{4 \cdot 16^4 \cdot \delta} \\
\geq& \frac{\mu D^2 (\log_{16}(15\lfloor (T-1)/K \rfloor + 16) - 2)}{4 \cdot 16^4 \cdot \delta} \geq \frac{\mu D^2 (\log_{16}(15(T-1)/K) - 2)}{4 \cdot 16^4 \cdot \delta} \\
\geq& \frac{\mu D^2 (\log_{16}(15\delta(T-1)/2) - 2)}{4 \cdot 16^4 \cdot \delta} \\
\geq& \frac{\mu D^2 (\log_{16}(15\delta(T-1)) - 3)}{4 \cdot 16^4 \cdot \delta},
\end{aligned}
$$

where the fifth inequality is due to $\lceil 1/\delta \rceil \leq 2/\delta$.

### C.5 Proof of Theorem 4.5

We first provide a more detailed version of D-FTFCL in Algorithm 5.

Due to the asynchronous communication between the learners and the server, each learner experiences a two-block delay during aggregation. Specifically, for the gradient computed in block $b$, it takes one block for the learner to transmit the information to the server, and another block for the server to aggregate and broadcast the updated information back to the learners. Consequently, at the end of block $b$, each learner updates its model using the gradient information from block $b-2$.

We first define several decisions.

$$
\hat{\mathbf{w}}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b} \mathbf{z}^k, \mathbf{w} \right\rangle + \frac{1}{\eta}\left\|\mathbf{w}\right\|^2,
$$

where $\mathbf{z}^b = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^b$. It is not hard to verify that $\hat{\mathbf{w}}^{b+1}$ is the decision with the exact gradient when there is no delay in transmission.

$$
\mathbf{y}^{b+1} = \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{1}{\eta}\left\|\mathbf{y}\right\|^2, \quad \tilde{\mathbf{w}}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{z}^k, \mathbf{w} \right\rangle + \frac{1}{\eta}\left\|\mathbf{w}\right\|^2.
$$

**Algorithm 5** Distributed Follow-the-Fast-Compressed-Leader (D-FTFCL)

**Require:** $\mathbf{w}^1 = \mathbf{0}$, $\mathbf{e}_i^1 = \mathbf{0}$, $\hat{\mathbf{e}}^1 = \mathbf{0}$, learning rate $\eta$, compressor $\mathcal{C}(\cdot)$, block size $L$

**for** $b = 1, \ldots, T/L$ **do**
  **for each learner** $i = 1, \ldots, n$ **in parallel**
    If $b \geq 2$, set $\mathbf{r}_{i,b-1}^k = \mathbf{0}, k = 1$
    **for** $t = (b-1)L + 1, \ldots, bL$ **do**
      Play the decision $\mathbf{w}^b$ and receive gradient $\mathbf{g}_i^t$
      If $b \geq 2$
        Compute $\mathcal{C}(\mathbf{z}_i^{b-1} - \mathbf{r}_{i,b-1}^k)$ and send to the central server
        Update $\mathbf{r}_{i,b-1}^{k+1} = \mathbf{r}_{i,b-1}^k + \mathcal{C}(\mathbf{z}_i^{b-1} - \mathbf{r}_{i,b-1}^k)$ and set $k = k + 1$
      **end if**
    **end for**
    Compute $\mathbf{z}_i^b = \sum_{t=(b-1)L+1}^{bL} \mathbf{g}_i^t$ and $\mathbf{v}_i^{b-1} = \mathbf{r}_{i,b-1}^{L+1}$
    Compute $\mathbf{e}_i^b = \mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1} - \mathbf{v}_i^{b-1}$
  **for server and if** $b \geq 3$ **do**
    Compute $\mathbf{v}^{b-2} = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^{b-2}$ ▷ Only receive gradients of block $b-2$ at block $b$
    Set $\mathbf{r}_{b-2}^k = \mathbf{0}, k = 1$
    **for** $t = (b-1)L + 1, \ldots, bL$ **do**
      Compute $\mathcal{C}(\mathbf{v}^{b-2} + \hat{\mathbf{e}}^{b-2} - \mathbf{r}_{b-2}^k)$ and send to all local learners
      Update $\mathbf{r}_{b-2}^{k+1} = \mathbf{r}_{b-2}^k + \mathcal{C}(\mathbf{v}^{b-2} + \hat{\mathbf{e}}^{b-2} - \mathbf{r}_{b-2}^k)$ and set $k = k + 1$
    **end for**
    Compute $\mathbf{s}^{b-2} = \mathbf{r}_{b-2}^{L+1}$
    Update $\hat{\mathbf{e}}^{b-1} = \hat{\mathbf{e}}^{b-2} + \mathbf{v}^{b-2} - \mathbf{s}^{b-2}$
  **for each learner** $i = 1, \ldots, n$ **in parallel and if** $b \geq 3$ **do**
    Receive total information $\mathbf{s}^{b-2}$ and update the decision $\mathbf{w}^{b+1}$ according to (6) or (7)
**end for**

We also have

$$\mathbf{e}^{b+1} = \sum_{k=1}^{b} \mathbf{z}^k - \mathbf{v}^k, \qquad \hat{\mathbf{e}}^{b+1} = \sum_{k=1}^{b} \mathbf{v}^k - \mathbf{s}^k.$$

In the following, we begin to derive the regret

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] &\leq \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}_i^t, \mathbf{w}^t - \mathbf{w} \rangle\right] \\
&= \frac{1}{n} \sum_{b=1}^{K} \sum_{t=(b-1)L}^{bL} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}_i^t, \mathbf{w}^b - \mathbf{w} \rangle\right] \\
&= \sum_{b=1}^{K} \sum_{t=(b-1)L}^{bL} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^b - \mathbf{w} \rangle\right] \\
&\leq \sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \mathbf{w}^b - \hat{\mathbf{w}}^b + \hat{\mathbf{w}}^b - \mathbf{w} \rangle\right] \\
&= \sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w} \rangle\right] + \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \mathbf{w}^b - \hat{\mathbf{w}}^b \right] \\
&\leq \underbrace{\sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w} \rangle\right]}_{\text{TERM-A}} + \underbrace{\sum_{b=1}^{K} GL\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^b - \hat{\mathbf{w}}^b\right\|\right]}_{\text{TERM-B}}.
\end{aligned}$$

We first give the bound of TERM-A. We define $\ell_1(\mathbf{w}) = \langle \mathbf{z}^1, \mathbf{w} \rangle + \frac{1}{\eta} \|\mathbf{w}\|^2$ and $\ell_b(\mathbf{w}) = \langle \mathbf{z}^b, \mathbf{w} \rangle$. By using Lemma C.1, we have

$$\sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^{b+1} - \mathbf{w} \rangle + \frac{\left\|\hat{\mathbf{w}}^2\right\|^2 - \|\mathbf{w}\|^2}{\eta} \leq \sum_{b=1}^{K} \ell_b(\hat{\mathbf{w}}^{b+1}) - \ell_b(\mathbf{w}) \leq 0. \tag{21}$$

Therefore, we can derive

$$\begin{aligned}
\sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w} \rangle =& \sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1} \rangle + \sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^{b+1} - \mathbf{w} \rangle \\
\leq& \sum_{b=1}^{K} \left\|\mathbf{z}^b\right\| \left\|\hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1}\right\| + \frac{\|\mathbf{w}\|^2 - \left\|\hat{\mathbf{w}}^2\right\|^2}{\eta} \\
\leq& \frac{\eta G^2 LT}{2} + \frac{D^2}{\eta},
\end{aligned}$$

where the last inequality is due to $\left\|\mathbf{z}^b\right\| \leq LG$.

The key difference in the proof is the bound of the norm of the errors. We introduce the following lemma.

**Lemma C.6.** *Under the Assumptions 3.3 and 3.4, at any block $b$ of D-FTFCL, the norm of the errors is bounded by*

$$\mathbb{E}_{\mathcal{C}} \left[\left\|\mathbf{e}^{b+1}\right\|^2\right] \leq 4e^2 L^2 G^2, \quad \mathbb{E}_{\mathcal{C}} \left[\left\|\hat{\mathbf{e}}^{b+1}\right\|^2\right] \leq 120 e^2 L^2 G^2.$$

In the following, we explicitly characterize the impact of the delayed updates and can get

$$\begin{aligned}
\left\|\mathbf{w}^b - \hat{\mathbf{w}}^b\right\| =& \left\|\mathbf{w}^b - \mathbf{y}^b + \mathbf{y}^b - \tilde{\mathbf{w}}^b + \tilde{\mathbf{w}}^b - \hat{\mathbf{w}}^b\right\| \\
\leq& \left\|\mathbf{w}^b - \mathbf{y}^b\right\| + \left\|\tilde{\mathbf{w}}^b - \mathbf{y}^b\right\| + \left\|\tilde{\mathbf{w}}^b - \hat{\mathbf{w}}^b\right\| \\
\leq& \frac{\eta}{2} \left( \left\| \sum_{k=1}^{b-2} \mathbf{s}^k - \mathbf{v}^k \right\| + \left\| \sum_{k=1}^{b-2} \mathbf{v}^k - \mathbf{z}^k \right\| + \left\|\mathbf{z}^{b-2} + \mathbf{z}^{b-1}\right\| \right) \\
=& \frac{\eta}{2} \left( \left\|\hat{\mathbf{e}}^{b-1}\right\| + \left\|\mathbf{e}^{b-1}\right\| + \left\|\mathbf{z}^{b-2} + \mathbf{z}^{b-1}\right\| \right) \\
\leq& \frac{\eta}{2} \left( 2eLG + \sqrt{120} eLG + 2LG \right) \leq 20 \eta LG,
\end{aligned} \tag{22}$$

where the second inequality is due to Lemma C.2 and the last inequality is due to Lemma C.6.

By setting the learning rate $\eta = \frac{D}{G\sqrt{LT}}$, the final regret bound is

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[R_T\right] \leq& \sum_{b=1}^{K} \langle \mathbb{E}_{\mathcal{C}}\left[\mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w}\right] \rangle + GL \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^b - \hat{\mathbf{w}}^b\right\|\right] \\
\leq& \frac{\eta G^2 LT}{2} + \frac{D^2}{\eta} + 20 \eta LT G^2 \\
\leq& O(DG\sqrt{LT}) = O(DG\delta^{-1/2}\sqrt{T}).
\end{aligned}$$

### C.6 Proof of Theorem 4.6

We first give some virtual decisions.

$$\begin{aligned}
\hat{\mathbf{w}}^{b+1} =& \arg\min_{\mathbf{w} \in \mathcal{W}} \left\langle \sum_{k=1}^{b} \mathbf{z}^k, \mathbf{w} \right\rangle + \frac{\mu L}{2} \sum_{k=1}^{b} \left\|\mathbf{w} - \mathbf{w}^k\right\|^2 \\
=& \arg\min_{\mathbf{w} \in \mathcal{W}} \left\langle \sum_{k=1}^{b} \mathbf{z}^k - \sum_{k=1}^{b} \mu L \mathbf{w}^k, \mathbf{w} \right\rangle + \frac{\mu b L}{2} \|\mathbf{w}\|^2 .
\end{aligned}$$

$$\begin{aligned}\mathbf{y}^{b+1} &= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{\mu L}{2} \sum_{k=1}^{b} \left\|\mathbf{y} - \mathbf{w}^k\right\|^2 \\ &= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{v}^k - \sum_{k=1}^{b} \mu L \mathbf{w}^k, \mathbf{y} \right\rangle + \frac{\mu b L}{2} \left\|\mathbf{y}\right\|^2 .\end{aligned}$$

$$\begin{aligned}\tilde{\mathbf{w}}^{b+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{z}^k, \mathbf{w} \right\rangle + \frac{\mu L}{2} \sum_{k=1}^{b} \left\|\mathbf{w} - \mathbf{w}^k\right\|^2 \\ &= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{b-2} \mathbf{z}^k - \sum_{k=1}^{b} \mu L \mathbf{w}^k, \mathbf{w} \right\rangle + \frac{\mu b L}{2} \left\|\mathbf{w}\right\|^2 .\end{aligned}$$

In the following, we begin to derive the regret bound

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_T\right] &= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{C}}\left[f_i^t(\mathbf{w}^t) - f_i^t(\mathbf{w})\right] \\ &\le \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}_i^t, \mathbf{w}^t - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^t - \mathbf{w}\right\|\right] \\ &= \sum_{b=1}^{K}\sum_{t=(b-1)L}^{bL} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^b - \mathbf{w}\rangle - \frac{\mu}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\|\right] \\ &= \sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \mathbf{w}^b - \hat{\mathbf{w}}^b + \hat{\mathbf{w}}^b - \mathbf{w}\rangle - \frac{\mu L}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\|\right] \\ &= \sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \mathbf{w}^b - \hat{\mathbf{w}}^b\rangle\right] + \mathbb{E}\left[\langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w}\rangle - \frac{\mu L}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\|\right] \\ &\le \underbrace{\sum_{b=1}^{K} \mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w}\rangle - \frac{\mu L}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\|\right]}_{\texttt{TERM-A}} + \underbrace{\sum_{b=1}^{K} GL\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{w}^b - \hat{\mathbf{w}}^b\right\|\right]}_{\texttt{TERM-B}},\end{aligned}$$

where the first inequality is due to Assumption 3.2 and the last inequality is due to $\left\|\mathbf{z}^b\right\| \le GL$.

We first give the bound of TERM-A. We define $\ell_b(\mathbf{w}) = \langle \mathbf{z}^b, \mathbf{w}\rangle + \frac{\mu L}{2}\left\|\mathbf{w} - \mathbf{w}^b\right\|^2$ and $\hat{\mathbf{w}}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{k=1}^{b} \ell_k(\mathbf{w})$ in the strongly convex case. Then, we have

$$\begin{aligned}&\sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \mathbf{w}\rangle - \frac{\mu L}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\| \\ &= \sum_{b=1}^{K} \langle \mathbf{z}^b, \hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1} + \hat{\mathbf{w}}^{b+1} - \mathbf{w}\rangle - \frac{\mu L}{2}\left\|\mathbf{w}^b - \mathbf{w}\right\| \\ &\le \sum_{b=1}^{K} \ell_b(\hat{\mathbf{w}}^{b+1}) - \ell_b(\mathbf{w}) + GL\left\|\hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1}\right\| \\ &\le \sum_{b=1}^{K} GL\left\|\hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1}\right\| .\end{aligned} \tag{23}$$

Next, we give the bound of $GL\left\|\hat{\mathbf{w}}^b - \hat{\mathbf{w}}^{b+1}\right\|$. We define $F_b(\mathbf{w}) = \sum_{k=1}^{b} \ell_k(\mathbf{w})$. It is not hard to verify that $F_b(\mathbf{w})$ is $(bL\mu)$-strongly convex function and $\hat{\mathbf{w}}^{b+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} F_b(\mathbf{w})$.

By using Lemma C.4, for any $\mathbf{x}, \mathbf{y} \in \mathcal{W}$, we have

$$\begin{aligned}|\ell_b(\mathbf{x}) - \ell_b(\mathbf{y})| &\leq |\langle \nabla \ell_b(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle| \\ &\leq \|\nabla \ell_b(\mathbf{x})\| \, \|\mathbf{x} - \mathbf{y}\| \\ &\leq (GL + \mu LD) \, \|\mathbf{x} - \mathbf{y}\| \, .\end{aligned}$$

Therefore, we can derive

$$\begin{aligned}GL \left\|\hat{\mathbf{w}}^{b+1} - \hat{\mathbf{w}}^{b}\right\|^2 \leq & \frac{2GL}{bL\mu}(F_b(\hat{\mathbf{w}}^b) - F_b(\hat{\mathbf{w}}^{b+1})) \\ \leq & \frac{2GL}{bL\mu}(F_{b-1}(\hat{\mathbf{w}}^b) - F_{b-1}(\hat{\mathbf{w}}^{b+1}) + \ell_b(\hat{\mathbf{w}}^b) - \ell_b(\hat{\mathbf{w}}^{b+1})) \\ \leq & \frac{2GL}{bL\mu}(\ell_b(\hat{\mathbf{w}}^b) - \ell_b(\hat{\mathbf{w}}^{b+1})) \\ \leq & \frac{2GL}{bL\mu}(GL + \mu DL) \left\|\hat{\mathbf{w}}^{b+1} - \hat{\mathbf{w}}^{b}\right\| ,\end{aligned}$$

which means

$$GL \left\|\hat{\mathbf{w}}^{b+1} - \hat{\mathbf{w}}^{b}\right\| \leq \frac{2GL}{bL\mu}(GL + \mu DL). \tag{24}$$

Moreover, according to (22), we have

$$\begin{aligned}\left\|\mathbf{w}^b - \hat{\mathbf{w}}^b\right\| &= \left\|\mathbf{w}^b - \mathbf{y}^b + \mathbf{y}^b - \tilde{\mathbf{w}}^b + \tilde{\mathbf{w}}^b - \hat{\mathbf{w}}^b\right\| \\ &\leq \left\|\mathbf{w}^b - \mathbf{y}^b\right\| + \left\|\tilde{\mathbf{w}}^b - \mathbf{y}^b\right\| + \left\|\tilde{\mathbf{w}}^b - \hat{\mathbf{w}}^b\right\| \\ &\leq \frac{1}{\mu bL}\left(\left\|\hat{\mathbf{e}}^{b-1}\right\| + \left\|\mathbf{e}^{b-1}\right\| + \left\|\mathbf{z}^{b-2} + \mathbf{z}^{b-1}\right\|\right) \\ &\leq \frac{1}{\mu bL}\left(2eLG + \sqrt{120}eLG + 2LG\right) \leq \frac{40LG}{\mu bL}.\end{aligned} \tag{25}$$

Therefore, by using Lemma C.6, (23) and (24), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_T\right] &\leq \sum_{b=1}^{K} \frac{2GL}{bL\mu}(GL + \mu DL) + \frac{GL}{\mu bL}(40LG) \\ &= L\sum_{b=1}^{K} \frac{2G}{b\mu}(G + \mu D) + \frac{40G^2}{\mu b} \\ &\leq O\left(L\frac{(G + \mu D)^2}{\mu}\log T\right),\end{aligned}$$

where the last inequality is due to $\sum_{b=1}^{T/L} \frac{1}{b} \leq \log(T/L) + 1$.

### C.7 Proof of Theorem 5.2

We first recall the property of the subgradient. For a subgradient $\mathbf{g}_i \in \partial f_i(\mathbf{x})$ of a convex function $f_i(\cdot)$ at $\mathbf{x}$, we have $f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \mathbf{g}_i, \mathbf{y} - \mathbf{x} \rangle$. Then, we first give the theoretical guarantee of the anytime online-to-batch conversion.

**Lemma C.7.** *(Theorem 2 in Cutkosky [2019]) We assume $\hat{\ell}^t(\mathbf{x})$ is convex and satisfies $f(\mathbf{x}^t) - f(\mathbf{x}) \leq \mathbb{E}[\hat{\ell}^t(\mathbf{x}^t) - \hat{\ell}^t(\mathbf{x})]$. Then for all $\mathbf{x} \in \mathcal{W}$ and $\mathbf{x}^T = \frac{\sum_{t=1}^{T} \alpha_t \mathbf{w}^t}{\alpha_{1:T}}$, the online-to-batch conversion guarantees:*

$$\mathbb{E}_{\mathcal{C}}[f(\mathbf{x}^T) - f(\mathbf{x})] \leq \mathbb{E}_{\mathcal{C}}\left[\frac{\sum_{t=1}^{T} \hat{\ell}^t(\mathbf{w}^t) - \hat{\ell}^t(\mathbf{x})}{\alpha_{1:T}}\right]$$

By setting $\alpha_t = 1$, $\hat{\ell}_i^t(\mathbf{w}) = \langle \mathbf{g}_i^t, \mathbf{w} \rangle$, $\hat{\ell}^t(\mathbf{w}) = \langle \mathbf{g}^t, \mathbf{w} \rangle$ and $\mathbf{g}^t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i^t$, we can obtain

$$\mathbb{E}_{\mathcal{C}}[f(\mathbf{x}^K) - f(\mathbf{x}^*)] = \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}^K) - f_i(\mathbf{x}^*)\right] \leq \mathbb{E}_{\mathcal{C}}\left[\frac{R_K}{K}\right].$$

We further have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_K\right] &= \sum_{t=1}^{K}\mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \mathbf{w}^t - \mathbf{w}\rangle\right] \\ &\leq \sum_{t=1}^{K}\langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t + \hat{\mathbf{w}}^t - \mathbf{w}\rangle \\ &= \sum_{t=1}^{K}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle + \langle \mathbf{g}^t, \mathbf{w}^t - \hat{\mathbf{w}}^t\rangle \\ &\leq \sum_{t=1}^{K}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle + G\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\|.\end{aligned}$$

Similar to the previous proof, we give the bound of the compression error.

**Lemma C.8.** *Under the Assumptions 3.4 and 5.1, at any round $t$ of D-O2B with compression communication, the norm of the errors is bounded by*

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}^{t+1}\right\|^2\right] \leq 4e^2\alpha_t^2G^2, \quad \mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] \leq 120e^2\alpha_t^2G^2.$$

According to the previous proof of Theorem 4.1, by using (13), Lemma C.2 and C.8, and setting $\alpha_t = 1$, we can obtain the following

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_K\right] &\leq \sum_{t=1}^{K}\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle + G\left\|\mathbf{w}^t - \hat{\mathbf{w}}^t\right\| \\ &\leq \frac{\eta G^2K}{2} + \frac{D^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{K}\left(\left\|\hat{\mathbf{e}}^{t-1}\right\| + \left\|\mathbf{e}^{t-1}\right\|\right) \\ &\leq \frac{\eta G^2K}{2} + \frac{D^2}{\eta} + \frac{13e\eta KG}{2} \leq O(DG\sqrt{K}),\end{aligned}$$

where the last inequality due to $\eta = \frac{D}{G\sqrt{K}}$ and $\mathbf{e}^0 = \hat{\mathbf{e}}^0 = \mathbf{0}$.

Furthermore, we can obtain

$$\mathbb{E}_{\mathcal{C}}[f(\mathbf{x}^K) - f(\mathbf{x}^*)] \leq \mathbb{E}\left[\frac{O(DG\sqrt{K})}{K}\right] = O\left(\frac{DG}{\sqrt{\delta T}}\right),$$

where the last equality is due to $T = KL$.

### C.8 Proof of Theorem 5.3

Following Cutkosky [2019], we set $\alpha_t = t$, $\hat{\ell}_i^t(\mathbf{w}) = \langle t\mathbf{g}_i^t, \mathbf{w}\rangle + \frac{\mu t}{2}\left\|\mathbf{w} - \mathbf{x}^t\right\|$, $\hat{\ell}^t(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_i^t(\mathbf{w})$, we can obtain

$$\mathbb{E}_{\mathcal{C}}[f(\mathbf{x}^K) - f(\mathbf{x}^*)] \leq \mathbb{E}_{\mathcal{C}}\left[\frac{2R_K}{K(K+1)}\right].$$

In each round $t$, each learner updates its decision as follows:

$$\begin{aligned}\mathbf{w}^{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}}\left\langle\sum_{k=1}^{t}\mathbf{s}^k, \mathbf{w}\right\rangle + \frac{\mu}{2}\sum_{k=1}^{t}k\left\|\mathbf{w} - \mathbf{x}^k\right\|^2 \\ &= \arg\min_{\mathbf{w}\in\mathcal{W}}\left\langle\sum_{k=1}^{t}\mathbf{s}^k - \sum_{k=1}^{t}k\mu\mathbf{x}^k, \mathbf{w}\right\rangle + \frac{\mu t(t+1)}{4}\left\|\mathbf{w}\right\|^2,\end{aligned}$$

and we define two variables:

$$\begin{aligned}\mathbf{y}^{t+1} &= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{\mu}{2} \sum_{k=1}^{t} k \left\|\mathbf{y} - \mathbf{x}^k\right\|^2 \\ &= \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k - \sum_{k=1}^{t} k\mu\mathbf{x}^k, \mathbf{y} \right\rangle + \frac{\mu t(t+1)}{4} \|\mathbf{y}\|^2 ,\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{w}}^{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} k\mathbf{g}^k, \mathbf{w} \right\rangle + \frac{\mu}{2} \sum_{k=1}^{t} k \left\|\mathbf{w} - \mathbf{x}^k\right\|^2 \\ &= \arg\min_{\mathbf{w}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} k\mathbf{g}^k - \sum_{k=1}^{t} k\mu\mathbf{x}^k, \mathbf{w} \right\rangle + \frac{\mu t(t+1)}{4} \|\mathbf{w}\|^2 .\end{aligned}$$

Next, we give the regret bound on the loss function $\hat{\ell}^t(\mathbf{w})$.

$$\begin{aligned}R_K =& \sum_{t=1}^{K} \hat{\ell}^t(\mathbf{w}^t) - \hat{\ell}^t(\mathbf{w}) \\ \leq& \sum_{t=1}^{K} \hat{\ell}^t(\hat{\mathbf{w}}^t) - \hat{\ell}^t(\mathbf{w}) + \hat{\ell}^t(\mathbf{w}^t) - \hat{\ell}^t(\hat{\mathbf{w}}^t) \\ \leq& \sum_{t=1}^{K} \hat{\ell}^t(\hat{\mathbf{w}}^t) - \hat{\ell}^t(\mathbf{w}) + (tG + \mu tD) \left\|\hat{\mathbf{w}}^t - \mathbf{w}^t\right\| .\end{aligned}$$

To bound the term $\sum_{t=1}^{K} \hat{\ell}^t(\hat{\mathbf{w}}^t) - \hat{\ell}^t(\mathbf{w})$, we introduce a lemma.

**Lemma C.9** (Corollary 5 in Cutkosky [2019]). *Under the Assumptions 3.2, 3.4 and 5.1, by setting the weight $\alpha_t = t$, $\hat{\ell}^t(\mathbf{w}) = \langle \alpha_t \mathbf{g}^t, \mathbf{w} \rangle + \frac{\mu\alpha_t}{2} \|\mathbf{w} - \mathbf{x}^t\|^2$ and $\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{k=1}^{t} \hat{\ell}^t(\mathbf{w})$, for any $\mathbf{w} \in \mathcal{W}$ anytime O2B can ensure*

$$\sum_{t=1}^{K} \hat{\ell}^t(\hat{\mathbf{w}}^t) - \hat{\ell}^t(\mathbf{w}) \leq \frac{K(\mu D + G)^2}{\mu}.$$

Then we bound the term $\|\mathbf{w}^t - \hat{\mathbf{w}}^t\|$. By using Lemma C.2 and C.8, and setting $\alpha_t = t$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}} \left[\left\|\mathbf{w}^{t+1} - \hat{\mathbf{w}}^{t+1}\right\|\right] &\leq \frac{2}{\mu t(t-1)} \left(\mathbb{E}_{\mathcal{C}} \left[\left\|\hat{\mathbf{e}}^{t-1}\right\|\right] + \mathbb{E}_{\mathcal{C}} \left[\left\|\mathbf{e}^{t-1}\right\|\right]\right) \\ &\leq \frac{30eG}{\mu t}.\end{aligned} \tag{26}$$

Finally, by using Lemma C.9 and (26), we can derive the regret bound

$$\begin{aligned}\mathbb{E}_{\mathcal{C}} \left[R_K\right] &\leq \frac{K(\mu D + G)^2}{\mu} + \sum_{t=2}^{K} \frac{30(G + \mu D)^2 et}{\mu(t-1)} \\ &\leq \frac{2K(G + \mu D)^2}{\mu} + \frac{60eK(G + \mu D)^2}{\mu} = O\left(\frac{K(G + \mu D)^2}{\mu}\right).\end{aligned}$$

The final convergence rate is

$$\mathbb{E}_{\mathcal{C}}[f(\mathbf{x}^K) - f(\mathbf{x}^*)] \leq \mathbb{E}_{\mathcal{C}} \left[\frac{2R_K}{K(K+1)}\right] \leq O\left(\frac{(G + \mu D)^2}{\mu K}\right) = O\left(\frac{(G + \mu D)^2}{\delta \mu T}\right).$$

**Additional discussion.** It is worth noting that our method involves only $K$ updates, with the total communication rounds amounting to $T = \lceil \frac{1}{\delta} \rceil K$. In contrast, standard compression-based

methods (e.g., O2B with Algorithm 2 and standard compressor) typically perform $T$ updates over $T$ communication rounds, yielding a convergence rate for convex loss functions of:

$$\mathbb{E}[f(\mathbf{x}^T) - f(\mathbf{x}^*)] \leq \mathbb{E}\left[\frac{R_T}{T}\right] = \frac{DG}{\delta\sqrt{T}}.$$

Compared to these methods, our approach achieves not only a faster convergence rate but also lower sample complexity under the same communication budget (i.e., the same number of communication rounds $T$). From a high-level perspective, although our algorithm performs fewer updates, the information incorporated in each update exhibits significantly lower bias compared to standard compression, thereby resulting in faster convergence.

Although using Algorithm 3 achieves the same convergence rate, it incurs redundant gradient queries.

### C.9 Proof of the Unidirectional Compression

It is worth noting that under the unidirectional compression, learner $i$ updates the decision according to

$$\mathbf{w}^{t+1} = \mathbf{y}^{t+1} = \arg\min_{\mathbf{y}\in\mathcal{W}} \left\langle \sum_{k=1}^{t} \mathbf{v}^k, \mathbf{y} \right\rangle + \frac{1}{\eta}\|\mathbf{y}\|^2.$$

Therefore, we can obtain the regret bound

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_T\right] &= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w}^t) - \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w})\right] \\ &\leq \sum_{t=1}^{T}\mathbb{E}_{\mathcal{C}}\left[\langle \mathbf{g}^t, \hat{\mathbf{w}}^t - \mathbf{w}\rangle\right] + G\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{y}^t - \hat{\mathbf{w}}^t\right\|\right].\end{aligned}$$

According to the proof of Theorem 4.1, we have

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{y}^t - \hat{\mathbf{w}}^t\right\|\right] \leq \frac{\eta}{2}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}^t\right\|\right] \leq \frac{\eta G}{\delta}.$$

By using (13) and setting $\eta = \frac{D\sqrt{\delta}}{\sqrt{T}G}$, the final regret bound is

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[R_T\right] &= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w}^t) - \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} f_i^t(\mathbf{w})\right] \\ &\leq \frac{\eta G^2 T}{2} + \frac{D^2}{\eta} + \frac{\eta T G}{\delta} \leq O(DG\sqrt{\delta^{-1}T}).\end{aligned}$$

Notably, the regret bound of $O(\delta^{-1}\log T)$ for strongly convex loss functions can be derived using an analogous proof under unidirectional compression.

# D Proof of Supporting Lemmas

## D.1 Proof of Lemma C.3

We first give the bound of the compression error of each learner. Following the proof in Karimireddy et al. [2019], we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}^{t+1}\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_i^{t+1}\right\|^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t+1}\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}+\mathbf{g}_i^{t}-\mathcal{C}(\mathbf{e}_i^{t}+\mathbf{g}_i^{t})\|^2\right] \\
&\leq \frac{1-\delta}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}+\mathbf{g}_i^{t}\|^2\right] \\
&\leq (1-\delta)(1+\alpha_1)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}\|^2\right] + (1-\delta)(1+\alpha_1^{-1})\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{g}_i^{t}\right\|^2 \\
&\leq (1-\delta)(1+\alpha_1)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}\|^2\right] + (1-\delta)(1+\alpha_1^{-1})G^2,
\end{aligned}$$

where the third inequality is due to $\|\mathbf{x}+\mathbf{y}\|^2 \leq (1+\alpha_1)\|\mathbf{x}\|^2 + (1+\alpha_1^{-1})\|\mathbf{y}\|^2$. By setting $\alpha_1 = \frac{\delta}{2-2\delta}$ and summing up, we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}^{t+1}\|^2\right] \leq& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t+1}\|^2\right] \leq \sum_{k=0}^{t-1}\left((1-\delta)(1+\alpha_1)\right)^{t-k}(1-\delta)(1+\alpha_1^{-1})G^2 \\
\leq& \frac{(1-\delta)(1+\alpha_1^{-1})G^2}{1-(1-\delta)(1+\alpha_1)} = \frac{(1-\delta)(1+\alpha_1^{-1})G^2}{\delta-\alpha_1+\alpha_1\delta} \\
\leq& \frac{4(1-\delta)G^2}{\delta^2}.
\end{aligned} \tag{27}$$

As for the upper bound of $\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right]$, we have the following

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t}\mathbf{v}^k - \mathbf{s}^k\right\|^2\right] = \mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t}\mathbf{v}^k - \sum_{k=1}^{t-1}\mathbf{s}^k - \mathcal{C}\left(\hat{\mathbf{e}}^t+\mathbf{v}^t\right)\right\|^2\right] \\
&= \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t+\hat{\mathbf{e}}^t-\mathcal{C}\left(\hat{\mathbf{e}}^t+\mathbf{v}^t\right)\right\|^2\right] \\
&\leq (1-\delta)\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t+\hat{\mathbf{e}}^t\right\|\right] \\
&\leq (1-\delta)(1+\alpha_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^t\right\|^2\right] + (1-\delta)(1+\alpha_2^{-1})\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right].
\end{aligned} \tag{28}$$

As for the second term $\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right]$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right] &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^t\right\|^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^t - (\mathbf{e}_i^t + \mathbf{g}_i^t) + (\mathbf{e}_i^t + \mathbf{g}_i^t)\right\|^2\right] \\
&\leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\left\|\mathcal{C}(\mathbf{e}_i^t + \mathbf{g}_i^t)) - (\mathbf{e}_i^t + \mathbf{g}_i^t)\right\|^2 + \frac{2}{n}\sum_{i=1}^n \left\|\mathbf{e}_i^t + \mathbf{g}_i^t\right\|^2\right] \\
&\leq \frac{2(1-\delta)}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t + \mathbf{g}_i^t\right\|^2\right] + \mathbb{E}_{\mathcal{C}}\left[\frac{2}{n}\sum_{i=1}^n \left\|\mathbf{e}_i^t + \mathbf{g}_i^t\right\|^2\right] \\
&= \frac{2(2-\delta)}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t + \mathbf{g}_i^t\right\|^2\right] \\
&\leq \frac{4(2-\delta)}{n}\sum_{i=1}^n \left(\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t\right\|^2\right] + \left\|\mathbf{g}_i^t\right\|^2\right) \\
&\leq 4(2-\delta)\left(\frac{4(1-\delta)G^2}{\delta^2} + G^2\right) \\
&\leq 8\left(\frac{4G^2}{\delta^2} + \frac{1}{\delta^2}G^2\right) \leq \frac{40G^2}{\delta^2}.
\end{aligned} \tag{29}$$

By setting $\alpha_2 = \frac{\delta}{2-2\delta}$ and plugging (29) into (28), we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] \leq& (1-\delta)(1+\alpha_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^t\right\|^2\right] + (1-\delta)(1+\alpha_2^{-1})\frac{40G^2}{\delta^2} \\
\leq& \sum_{k=0}^{t-1}\left((1-\delta)(1+\alpha_2)\right)^{t-k}(1-\delta)(1+\alpha_2^{-1})\frac{40G^2}{\delta^2} \\
\leq& \frac{(1-\delta)(1+\alpha_2^{-1})\frac{40G^2}{\delta^2}}{1-(1-\delta)(1+\alpha_2)} \leq \frac{160(1-\delta)G^2}{\delta^4}.
\end{aligned}$$

### D.2 Proof of Lemma C.6

According to Lemma 1, by choosing the compression round $L = \lceil 1/\delta \rceil$, we can ensure $\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{r}^{L+1} - \mathbf{x}\right\|^2\right] \leq \frac{1}{e}\left\|\mathbf{x}\right\|^2$, where $\mathbf{r}^{L+1}$ is the output of Algorithm 1. Therefore, we have

$$\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1} - \mathbf{v}_i^{b-1}\right\|^2\right] = \mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1} - \text{FCC}(\mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1}, \mathcal{C}(\cdot), L)\right\|^2\right] \leq \frac{1}{e}\left\|\mathbf{e}_i^{b-1} + \mathbf{z}_i^{b-1}\right\|^2. \tag{30}$$

Therefore, we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}^{b+1}\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{e}_i^{b+1}\right\|^2\right] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{b+1}\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^b + \mathbf{z}_i^b - \mathbf{v}_i^b\|^2\right] \\
&\leq \frac{1}{en}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^b + \mathbf{z}_i^b\|^2\right] \\
&\leq (1+\alpha_1)\frac{1}{en}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^b\|^2\right] + (1+\alpha_1^{-1})\frac{1}{en}\sum_{i=1}^n \left\|\mathbf{z}_i^b\right\|^2 \\
&\leq (1+\alpha_1)\frac{1}{en}\sum_{i=1}^n \mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^b\|^2\right] + \frac{1+\alpha_1^{-1}}{e}L^2G^2,
\end{aligned}$$

where the second equality is due to (30). By setting $\alpha_1 = \frac{e-1}{2}$ and summing up, we can derive

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{b+1}\|^2\right] \leq & \sum_{k=0}^{b-1}\left(\frac{1+\alpha_1}{e}\right)^{b-k}(1+\alpha_1^{-1})\frac{L^2G^2}{e} \\ \leq & \frac{(1+\alpha_1^{-1})L^2G^2}{e-(1+\alpha_1)} \leq \frac{4L^2G^2}{e(1-1/e)^2} \leq 4e^2L^2G^2.\end{aligned}$$

As for the term $\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b+1}\right\|^2\right]$, we have the following

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b+1}\right\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b}+\mathbf{v}^b-\text{FCC}\left(\hat{\mathbf{e}}^b+\mathbf{v}^b,\mathcal{C}(\cdot),L\right)\right\|^2\right] \\ &\leq \frac{1}{e}\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b}+\mathbf{v}^b\right\|^2\right] \\ &\leq \frac{1+\alpha_2}{e}\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b}\right\|^2\right]+\frac{1+\alpha_2^{-1}}{e}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^b\right\|^2\right],\end{aligned} \tag{31}$$

where the first equality is due to (30). To bound the term $\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^{b+1}\right\|^2\right]$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^{b}\right\|^2\right] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^{b}\right\|^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^{b}-(\mathbf{e}_i^b+\mathbf{z}_i^b)+(\mathbf{e}_i^b+\mathbf{z}_i^b)\right\|^2\right] \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^{b}-(\mathbf{e}_i^b+\mathbf{z}_i^b)\right\|^2\right]+\mathbb{E}_{\mathcal{C}}\left[\frac{2}{n}\sum_{i=1}^{n}\left\|\mathbf{e}_i^b+\mathbf{z}_i^b\right\|^2\right] \\ &\leq \frac{2/e}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^b+\mathbf{z}_i^b\right\|^2\right]+\mathbb{E}_{\mathcal{C}}\left[\frac{2}{n}\sum_{i=1}^{n}\left\|\mathbf{e}_i^b+\mathbf{z}_i^b\right\|^2\right] \\ &= \frac{2+2/e}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^b+\mathbf{z}_i^b\right\|^2\right] \\ &\leq \frac{4+4/e}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^b\right\|^2\right]+\left\|\mathbf{z}_i^b\right\|^2\right) \\ &\leq (4+4/e)\left(4e^2L^2G^2+L^2G^2\right)=30e^2L^2G^2.\end{aligned} \tag{32}$$

By setting $\alpha_2 = \frac{e-1}{2}$ and plugging (32) into (31), we can derive

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b+1}\right\|^2\right] \leq & \frac{1}{e}(1+\alpha_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b}\right\|^2\right]+\frac{1+\alpha_2^{-1}}{e}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^b\right\|^2\right] \\ \leq & \frac{1}{e}(1+\alpha_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{b}\right\|^2\right]+\frac{1+\alpha_2^{-1}}{e}30e^2L^2G^2 \\ \leq & \sum_{k=0}^{t-1}\left(\frac{1}{e}(1+\alpha_2)\right)^{t-k}\frac{1+\alpha_2^{-1}}{e}30e^2L^2G^2 \\ \leq & \frac{(1+\alpha_2^{-1})30e^2L^2G^2}{e-(1+\alpha_2)} \leq 120e^2L^2G^2,\end{aligned}$$

where the last inequality is due to $\frac{1+e}{(e-1)^2} \leq 2$.

### D.3 Proof of Lemma C.8

We first give the bound of the compression error of each learner. Following the proof in Karimireddy et al. [2019], we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}^{t+1}\|^2\right] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t+1}\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}+\alpha_t\mathbf{g}_i^t-\text{FCC}(\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t,\mathcal{C}(\cdot),L)\|^2\right] \\
&\leq \frac{1}{en}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}+\alpha_t\mathbf{g}_i^t\|^2\right] \\
&\leq (1+\beta_1)\frac{1}{en}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}\|^2\right]+(1+\beta_1^{-1})\frac{1}{en}\sum_{i=1}^{n}\left\|\alpha_t\mathbf{g}_i^t\right\|^2 \\
&\leq (1+\beta_1)\frac{1}{en}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t}\|^2\right]+(1+\beta_1^{-1})\alpha_t^2G^2.
\end{aligned}$$

By setting $\beta_1=\frac{e-1}{2}$ and summing up, we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}^{t+1}\|^2\right] \leq& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\|\mathbf{e}_i^{t+1}\|^2\right] \leq \alpha_t^2\sum_{k=0}^{t-1}\left(\frac{1}{e}(1+\beta_1)\right)^{t-k}\frac{1}{e}(1+\beta^{-1})G^2 \\
\leq& \frac{\alpha_t^2(1+\beta^{-1})G^2}{e-(1+\beta_1)} \leq 4e^2\alpha_t^2G^2,
\end{aligned} \tag{33}$$

where the third inequality is due to $\alpha_t \geq \alpha_{t-1}$.

As for the upper bound of $\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right]$, we similar inequality

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t}\mathbf{v}^k-\mathbf{s}^k\right\|^2\right] = \mathbb{E}_{\mathcal{C}}\left[\left\|\sum_{k=1}^{t}\mathbf{v}^k-\sum_{k=1}^{t}\mathbf{s}^k-\text{FCC}\left(\sum_{k=1}^{t}\mathbf{v}^k-\sum_{k=1}^{t}\mathbf{s}^k,\mathcal{C}(\cdot),L\right)\right\|^2\right] \\
&\leq \frac{1}{e}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t+\hat{\mathbf{e}}^t\right\|\right] \leq \frac{1}{e}(1+\beta_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^t\right\|^2\right]+\frac{1}{e}(1+\beta_2^{-1})\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right].
\end{aligned} \tag{34}$$

As for the second term $\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right]$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}^t\right\|^2\right] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^t-(\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t)+(\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t)\right\|^2\right] \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{v}_i^t-(\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t)\right\|^2+\frac{2}{n}\sum_{i=1}^{n}\left\|\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t\right\|^2\right] \\
&\leq \frac{2/e}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t\right\|^2\right]+\mathbb{E}_{\mathcal{C}}\left[\frac{2}{n}\sum_{i=1}^{n}\left\|\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t\right\|^2\right] \\
&= \frac{2+2/e}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t+\alpha_t\mathbf{g}_i^t\right\|^2\right] \\
&\leq \frac{4+4/e}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mathcal{C}}\left[\left\|\mathbf{e}_i^t\right\|^2\right]+\left\|\alpha_t\mathbf{g}_i^t\right\|^2\right) \\
&\leq (4+4/e)\left(4e^2\alpha_{t-1}^2G^2+\alpha_t^2G^2\right) \\
&\leq (4+4/e)\left(4e^2\alpha_t^2G+\alpha_t^2G^2\right) \leq 30e^2\alpha_t^2G^2.
\end{aligned} \tag{35}$$

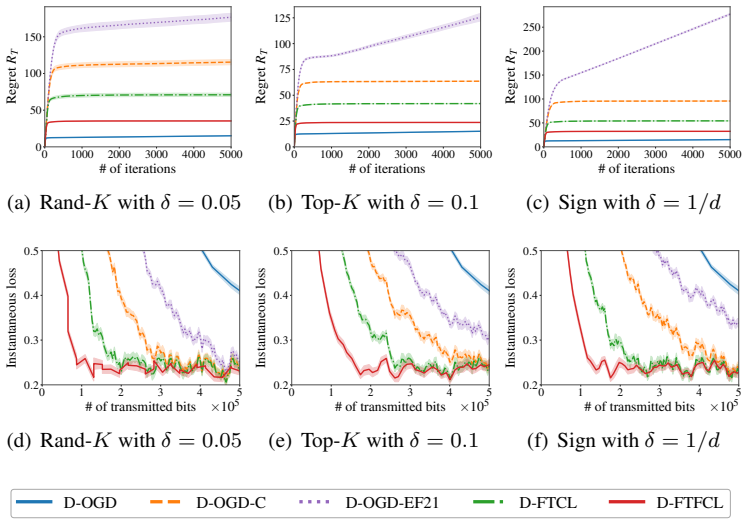
By setting $\beta_2 = \frac{e-1}{2}$ and plugging (35) into (34), we can derive

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t+1}\right\|^2\right] \leq & \frac{1}{e}(1+\beta_2)\mathbb{E}_{\mathcal{C}}\left[\left\|\hat{\mathbf{e}}^{t}\right\|^2\right] + \frac{1}{e}(1+\beta_2^{-1})30e^2\alpha_t^2G^2 \\ \leq & \alpha_t^2\sum_{k=0}^{t-1}\left(\frac{1}{e}(1+\beta_2)\right)^{t-k}\frac{1}{e}(1+\beta_2^{-1})30e^2G^2 \\ \leq & \frac{(1+\beta_2^{-1})30e^2\alpha_t^2G^2}{e-(1+\beta_2)} \leq 120e^2\alpha_t^2G^2.\end{aligned}$$

# E Experiments

In this section, we assess the performance of our methods through numerical experiments. All experiments are conducted on a server equipped with an Intel Xeon Platinum 8358P CPU (2.60 GHz).

## E.1 Online Classification

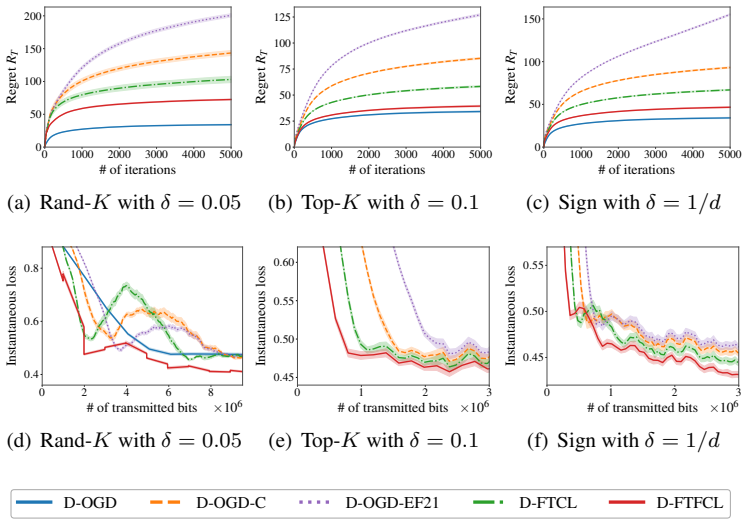



(a) Rand-$K$ with $\delta = 0.05$ (b) Top-$K$ with $\delta = 0.1$ (c) Sign with $\delta = 1/d$

(d) Rand-$K$ with $\delta = 0.05$ (e) Top-$K$ with $\delta = 0.1$ (f) Sign with $\delta = 1/d$

Figure 1: Results on the ijcnn1 dataset for convex loss functions.

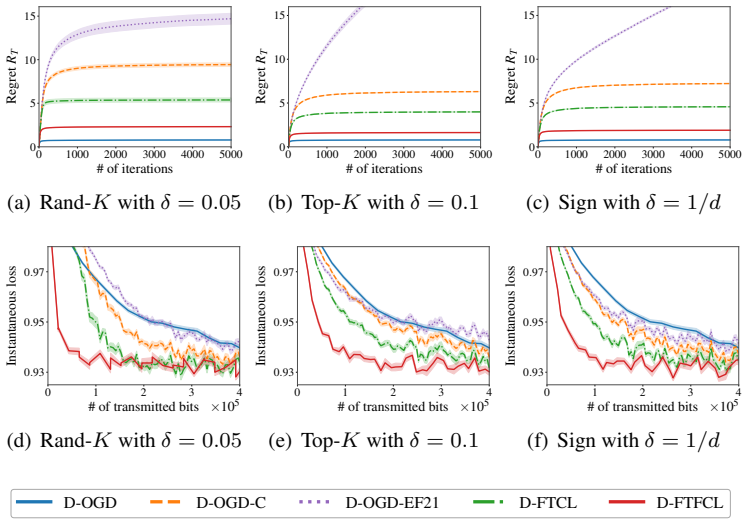
**Setup.** To validate our theoretical results, we perform the distributed online classification tasks with ijcnn1 and a9a datasets from LIBSVM [Chang and Lin, 2011]. Let $T$ denote the number of rounds and $n$ denote the number of learners. At each round $t \in [T]$, each learner $i$ receives a batch of sampled data $\{\mathbf{z}_{i,j}^t, y_{i,j}^t\}_{j\in[m_i]} \in \mathbb{R}^d \times \{-1, 1\}$, where $m_i$ is the batch size of each learner. Then, each learner plays the decision $\mathbf{w}^t$ from the ball $\mathcal{W}$ with the diameter $D$ and suffers a loss. We consider two types of functions. For convex functions, the learner $i$ suffers the non-smooth hinge loss $f_i^t(\mathbf{w}) = \sum_{j=1}^{m_i}\frac{1}{m_i}\max\{0, 1-y_{i,j}^t\mathbf{w}^\top\mathbf{z}_{i,j}^t\}$, and for $\mu$-strongly convex functions, the learner $i$ suffers the regularized loss $f_i^t(\mathbf{w}) = \sum_{j=1}^{m_i}\frac{1}{m_i}\max\{0, 1-y_{i,j}^t\mathbf{w}^\top\mathbf{z}_{i,j}^t\} + \frac{\mu}{2}\|\mathbf{w}\|^2$. In our experiments, we set the rounds $T = 5000$ and $\mu = 1$ for the strongly convex loss functions and each learner only receives positive or negative data at each round to ensure the difference of the local functions. We set domain diameter $D = 5$, the number of learners $n = 16$, batch size $m = \sum_{i=1}^n m_i = 64$ for the ijcnn1 dataset and number of learners $n = 128$, batch size $m = \sum_{i=1}^n m_i = 1024$, $D = 6$ for the a9a dataset. We conduct the experiments with 3 types of compressors: Rand-$K$ with $\delta = 0.05$, Top-$K$ with $\delta = 0.1$ and Sign with $\delta = 1/d$. As for the block size, we set $L = \lceil 1/\delta \rceil$ for Rand-$K$ compressor. Since the regret bounds are derived under worst-case analysis, we perform a grid search over the block size $L$ from the set $\{\lceil 1/(10\delta)\rceil, \lceil 1/(8\delta)\rceil, \lceil 1/(6\delta)\rceil, \lceil 1/(4\delta)\rceil, \lceil 1/(2\delta)\rceil, \lceil 1/\delta\rceil\}$ for Top-$K$ and Sign compressors. We compare our methods with distributed OGD without compressed communication (D-OGD), distributed OGD with bidirectional EF21 compression (D-OGD-EF21)




(a) Rand-$K$ with $\delta = 0.05$ (b) Top-$K$ with $\delta = 0.1$ (c) Sign with $\delta = 1/d$

(d) Rand-$K$ with $\delta = 0.05$ (e) Top-$K$ with $\delta = 0.1$ (f) Sign with $\delta = 1/d$

Figure 2: Results on the a9a dataset for convex loss functions.




(a) Rand-$K$ with $\delta = 0.05$ (b) Top-$K$ with $\delta = 0.1$ (c) Sign with $\delta = 1/d$

(d) Rand-$K$ with $\delta = 0.05$ (e) Top-$K$ with $\delta = 0.1$ (f) Sign with $\delta = 1/d$

Figure 3: Results on the ijcnn1 dataset for strongly convex loss functions.

[Fatkhullin et al., 2025] and distributed OGD with compressed communication (D-OGD-C). Each experiment is repeated 5 times and the average performance is reported.

**Results.** We plot the regret versus the iterations and the instantaneous loss versus the transmitted bits. As demonstrated in Figures 1, 2, 3 and 4, although D-OGD-C and D-OGD-EF21 converge, they suffer from worse regret due to the accumulation of projection error. Our D-FTFCL outperforms D-FTCL, a result attributed to its smaller compression error. D-FTFCL achieves performance comparable to OGD without compressed communication and shows better performance under the same communication budget. Additionally, we plot the total communication cost on the a9a dataset for D-FTFCL with Rand-$K$ (D-FTFCL(R)), Top-$K$ (D-FTFCL(T)) and Sign (D-FTFCL(S)) in Figure 5. As can be observed, our methods significantly reduce the communication cost compared to D-OGD.

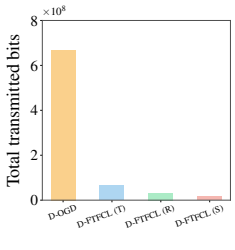
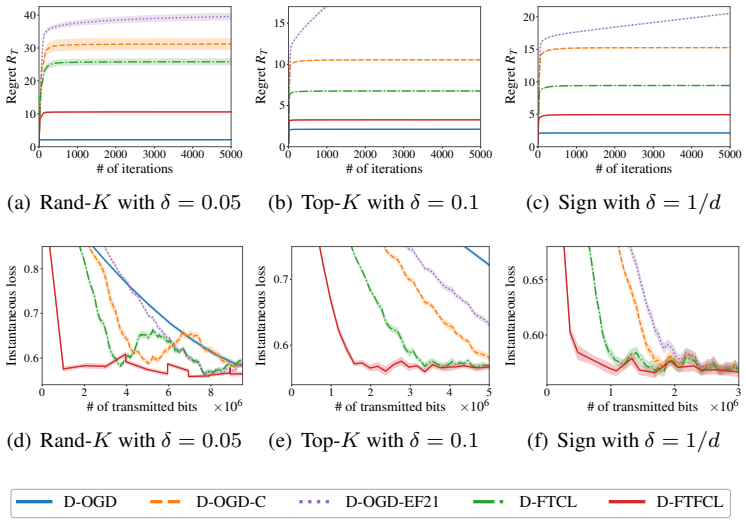



Figure 5: Transmitted bits on the a9a dataset.




(a) Rand-$K$ with $\delta = 0.05$ (b) Top-$K$ with $\delta = 0.1$ (c) Sign with $\delta = 1/d$

(d) Rand-$K$ with $\delta = 0.05$ (e) Top-$K$ with $\delta = 0.1$ (f) Sign with $\delta = 1/d$

Figure 4: Results on the a9a dataset for strongly convex loss functions.

### E.2 Stochastic Optimization

**Setup.** In this section, we perform the distributed stochastic optimization tasks with a9a and Gisette datasets from LIBSVM [Chang and Lin, 2011]. In particular, Gisette is a high-dimensional dataset with $d = 5000$. In each round $t$, each learner $i$ can only have access to a batch of $m_i$ samples $\{\mathbf{z}_{i,j}^t, y_{i,j}^t\}_{j\in[m_i]} \in \mathbb{R}^d \times \{-1, 1\}$. The goal of the learners is to minimize the regularized hinge loss

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{m_i}\max\{0, 1 - y_{i,j}^t \mathbf{x}^\top \mathbf{z}_{i,j}^t\} + \frac{\mu}{2}\|\mathbf{x}\|^2,$$

where $\mu$ is the regularization parameter. For convex loss functions, we set $\mu = 0$. For strongly convex loss functions, we set $\mu = 0.001$. In our experiments, each learner only receives one type of data at each round to ensure the difference of the local functions. For the Gisette dataset, we set $D = 50$, number of learners $n = 8$ and batch size $m_i = 3$. For the a9a dataset, we set $D = 6$, and we conduct experiments under two configurations: $n = 16, m_i = 2$ and $n = 128, m_i = 8$.

**Results.** We conduct the experiments with the scaled sign compressor. Each experiment is repeated 5 times and the average performance is reported. We compare our Algorithm 4 (FCC-O2B) with distributed O2B [Cutkosky, 2019], distributed O2B with EF21 [Fatkhullin et al., 2025] (O2B-EF21) and distributed O2B with standard compressors (C-O2B). As presented in Figures 6, 7, 8, 9, 10 and 11, our algorithms exhibit significantly faster convergence than the distributed O2B and distributed O2B with standard compressors under the same communication bits budget. This empirical evidence corroborates the superior communication efficiency of our proposed methods.

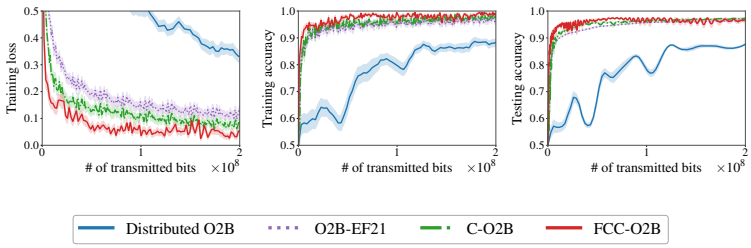



Figure 6: Results on the Gisette dataset for convex loss functions in stochastic optimization.

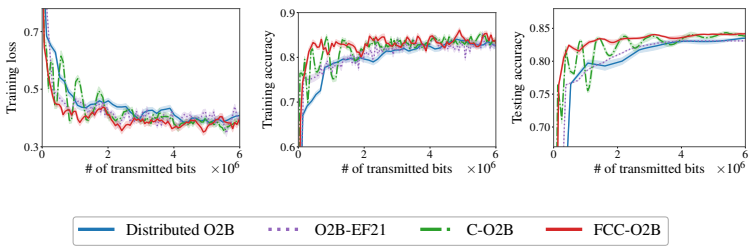



Figure 7: Results on the a9a dataset with $n = 16$ for convex loss functions in stochastic optimization.

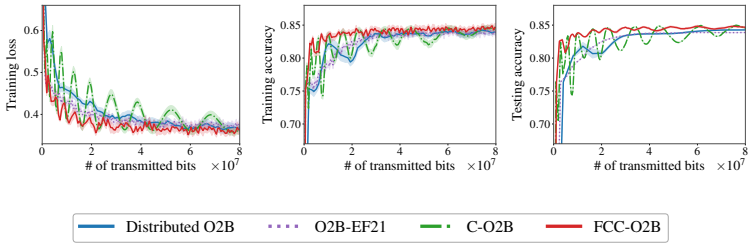



Figure 8: Results on the a9a dataset with $n = 128$ for convex loss functions in stochastic optimization.

### E.3 Sensitivity of the Block Size

In this section, we empirically validate the impact of different block sizes on the experimental results. The setup follows the same methodology as in the online classification experiment. We conduct the experiments on the a9a dataset with the Rand-$K$ compressor with $\delta = 0.05$. We test the block sizes $L \in \{1, \lceil 1/\delta \rceil, \lceil 5/\delta \rceil\}$ and plot the regret and the instantaneous loss.

**Results.** All experiments are repeated 5 times, and we report the instantaneous loss, and the regret against the number of rounds in Figure 12 for the general convex case. From the experimental results, it is evident that different choices of block size lead to distinct performance behaviors. A small block size results in frequent updates, but the corresponding update information tends to be inaccurate. In contrast, a large block size yields more accurate update information but suffers from delayed updates, which causes the loss of information. Consequently, only an appropriately chosen block size can lead to the best regret performance.

### E.4 Sensitivity of the Batch Size

In this section, we present additional convex optimization experiments on the a9a dataset to validate whether the proposed method is sensitive to batch size with the sign compressor. The setup follows the same methodology as in the stochastic classification experiment. Specifically, we consider the convex loss functions and adopt the scaled sign compressor. We set $D = 6$ and the number of learners $n = 16$. We try different total batch sizes $m \in \{16, 32, 64, 128\}$, and the results are shown in Figure 13. It can be seen that our algorithm is not sensitive to variations in batch size.

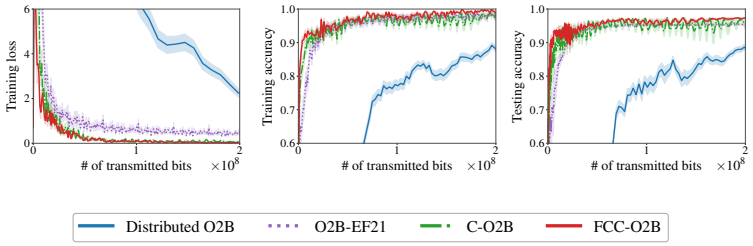



Figure 9: Results on the Gisette dataset for strongly convex loss functions in stochastic optimization.

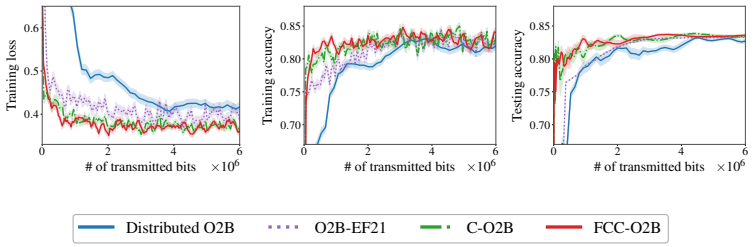



Figure 10: Results on the a9a dataset with $n = 16$ for strongly convex loss functions in stochastic optimization.

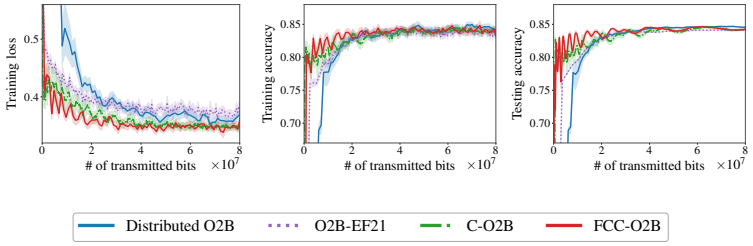



Figure 11: Results on the a9a dataset with $n = 128$ for strongly convex loss functions in stochastic optimization.

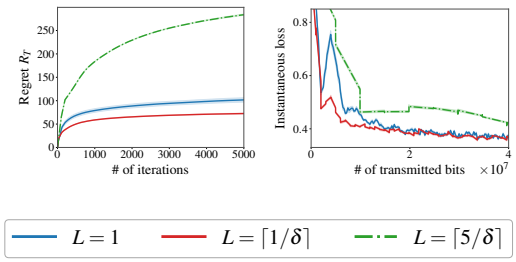



Figure 12: Results for the a9a dataset with different block sizes.

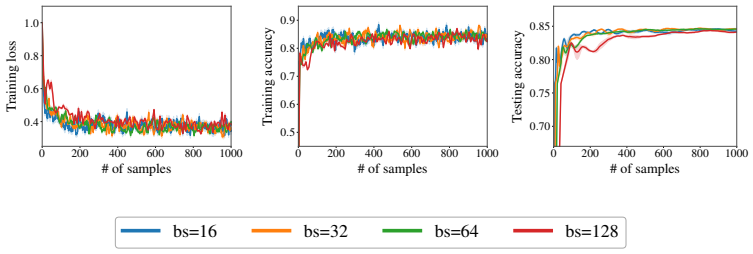



Figure 13: Results for the a9a dataset with $n = 16$ under different batch sizes.