Catalyst Acceleration of Error Compensated Methods Leads to Better Communication Complexity

Xun Qian
Shanghai Artificial Intelligence Lab
Shanghai, China

Hanze Dong
The Hong Kong University of Science and Technology
Hong Kong

Tong Zhang
The Hong Kong University of Science and Technology
Hong Kong

Peter Richtárik
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Abstract

Communication overhead is well known to be a key bottleneck in large scale distributed learning, and a particularly successful class of methods which help to overcome this bottleneck is based on the idea of communication compression. Some of the most practically effective gradient compressors, such as TopK, are biased, which causes convergence issues unless one employs a well designed *error compensation/feedback* mechanism. Error compensation is therefore a fundamental technique in the distributed learning literature. In a recent development, Qian et al (NeurIPS 2021) showed that the error-compensation mechanism can be combined with acceleration/momentum, which is another key and highly successful optimization technique. In particular, they developed the error-compensated loop-less Katyusha (ECLK) method, and proved an accelerated linear rate in the strongly convex case. However, the dependence of their rate on the compressor parameter does not match the best dependence obtainable in the non-accelerated error-compensated methods. Our work addresses this problem. We propose several new accelerated error-compensated methods using the *catalyst acceleration* technique, and obtain results that match the best dependence on the compressor parameter in non-accelerated error-compensated methods up to logarithmic terms.

1 INTRODUCTION

In large scale machine learning optimization problems, the data and training need to be distributed among many machines [Verbraeken et al., 2019]. Also in federated learning [Konečný et al., 2016b,a, McMahan et al., 2017, Li et al., 2019], training occurs on edge devices such as mobile phones and smart home devices, where the data is originally captured. In these applications, the distributed machine learning can be characterized as the following composite finite-sum problem

$$\min_{x\in\mathbb{R}^d} P(x) := \left\{ \frac{1}{n}\sum_{\tau=1}^{n} f^{(\tau)}(x) + \psi(x) \right\}, \qquad (1)$$

where $\{f^{(\tau)}(x)\}_{\tau=1}^{n}$ are smooth convex functions distributed over n nodes, and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a regularizer, which is a proper closed convex but possibly non-smooth function. On each node τ, $f^{(\tau)}(x) := \frac{1}{m}\sum\limits_{i=1}^{m} f_i^{(\tau)}(x)$ is the average loss over the training data stored on this node and each $f_i^{(\tau)}$ is smooth and convex.

In distributed and especially federated settings, communication is generally much slower than the local training, which makes the communication overhead become a key bottleneck. In order to overcome this bottleneck, several methods were proposed in the literature, such as using large minibatches [Goyal et al., 2017, You et al., 2017], asynchronous learning [Tsitsiklis et al., 1986, Agarwal and Duchi, 2011, Lian et al., 2015, Recht et al., 2011], and gradient compression [Seide et al., 2014, Alistarh et al., 2017, Bernstein et al., 2018, Wen et al., 2017, Mishchenko et al., 2019]. In this work, we focus on the error-compensated method, which is a gradient compression method and is capable to deal with some effective but biased compressors, such as the TopK compressor.

Related Work. The *error compensation/feedback* mechanism was first introduced in 1-bit SGD [Seide et al., 2014]. Then the error-compensated SGD (ECSGD) was proved

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

to have the same convergence rate as vanilla SGD in the strongly convex case [Stich et al., 2018] and non-convex case [Karimireddy et al., 2019, Tang et al., 2019] when P is smooth. ECSGD was further studied in [Stich and Karimireddy, 2020] under weaker assumptions. When P is non-smooth, it was shown that ECSGD converges at the rate of $\mathcal{O}(1/\sqrt{\delta T})$ in [Karimireddy et al., 2019], where T denotes the iteration number and δ is the compressor parameter defined in (2). The non-accelerated linear convergence can be obtained in EC-LSVRG-DIANA [Gorbunov et al., 2020] in the smooth case, and in the error-compensated loop-less SVRG, Quartz, and SDCA [Qian et al., 2021a] in the composite case. In a recent development, the error-compensated loop-less Katyusha was proposed in [Qian et al., 2021c], and the accelerated linear rate was achieved.

Compressor. In error-compensated methods, contraction compressors are generally used. A randomized map $Q : \mathbb{R}^d \to \mathbb{R}^d$ is called a *contraction compressor* if there exists a constant $\delta \in (0, 1]$ such that

$$\mathbb{E}\left[\|x - Q(x)\|^2\right] \leq (1-\delta)\|x\|^2, \qquad \forall x \in \mathbb{R}^d. \quad (2)$$

Some frequently used contraction compressors include TopK [Alistarh et al., 2018] and RandK [Stich et al., 2018]. Let $1 \leq K \leq d$. The TopK compressor is defined as

$$(\mathrm{TopK}(x))_{\pi(i)} = \begin{cases} (x)_{\pi(i)} & \text{if } i \leq K, \\ 0 & \text{otherwise,} \end{cases}$$

where π is a permutation of $\{1, 2, ..., d\}$ such that $(|x|)_{\pi(i)} \geq (|x|)_{\pi(i+1)}$ for $i = 1, ..., d-1$. For TopK and RandK compressors, we have $\delta \geq K/d$ [Stich et al., 2018].

The *unbiased compressor* is also frequently used in compression algorithms, which is defined as a randomized map $\tilde{Q} : \mathbb{R}^d \to \mathbb{R}^d$, where there exists a constant $\omega \geq 0$ such that $\mathbb{E}[\tilde{Q}(x)] = x$, and

$$\mathbb{E}\left[\|\tilde{Q}(x)\|^2\right] \leq (\omega+1)\|x\|^2, \qquad \forall x \in \mathbb{R}^d. \quad (3)$$

Some frequently used unbiased compressors include random dithering [Alistarh et al., 2017], random sparsification [Stich et al., 2018], and natural compression [Horváth et al., 2019b]. For any $\tilde{Q}$ satisfying (3), $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor satisfying (2) with $\delta = 1/(\omega+1)$[Beznosikov et al., 2020]. Furthermore, unbiased compressors and contraction compressors can be composed to generate new contraction compressors [Qian et al., 2021a].

1.1 Motivation

Communication Complexity of ECLK. There are two contraction compressors Q and Q_1 in ECLK [Qian et al., 2021c] with parameter δ and δ_1 respectively. We first claim that when Q and Q_1 in ECLK are the same type of contraction compressor, but with possibly different compressor parameters (for example, Q and Q_1 are both TopK, but with different values of K), we could always choose the same compressor parameters for Q and Q_1 such that the total communication complexity is less than before or remains the same order as before.

First, from the iteration complexity results for ECLK, it is easy to verify that the iteration complexity will decrease as δ or δ_1 increases. Without less of generality, we assume the communication cost of $Q(x)$ is higher than that of $Q_1(x)$. Since Q and Q_1 are the same type of compressor, we will have $\delta_1 \leq \delta$. Then we can change Q_1 to be Q. In this way, the total communication cost of $Q(x)$ and $Q_1(x)$ at each iteration is at most twice as before, but δ_1 will increase to δ, which implies that the iteration complexity will decrease and the communication complexity is at most twice as before. Thus, for simplicity, we consider $Q = Q_1$ for ECLK.

Dependence on δ for the Iteration Complexity of ECLK. We introduce the following assumption for Problem (1).

Assumption 1.1 *$\frac{1}{n}\sum_{\tau=1}^{n} f^{(\tau)}$ is L_f-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, $f_i^{(\tau)}$ is L-smooth, and ψ is λ-strongly convex.*

Under Assumption 1.1, from Theorem 3.8 in [Qian et al., 2021c], the iteration complexity is

$$\mathcal{O}\left(\left(\frac{1}{\delta} + \frac{1}{\delta_1} + \frac{1}{p} + \sqrt{\frac{L_f}{\lambda}} + \sqrt{\frac{\mathcal{L}_2}{\lambda p}}\right)\log\frac{1}{\epsilon}\right),$$

where

$$\mathcal{L}_2 = \frac{6L}{n} + \frac{112(1-\delta)\bar{L}}{3\delta^2} + \frac{28(1-\delta)L}{3\delta} + \frac{224(1-\delta)\bar{L}p}{\delta^2\delta_1}\left(1 + \frac{2p}{\delta_1}\right)$$

and $p \in (0, 1]$ is the update frequency of the check point. Considering $\delta_1 = \delta$, it is easy to see that the iteration complexity of ECLK is at least $\mathcal{O}\left(\frac{\sqrt{1-\delta}}{\delta\sqrt{\delta}}\sqrt{\frac{\bar{L}}{\lambda}}\log\frac{1}{\epsilon}\right)$. Hence, when $1 - \delta = \Theta(1)$, the dependence on δ of the communication complexity of ECLK would be $1/\delta^{\frac{3}{2}}$, which is worse than EC-LSVRG in the smooth case and EC-SDCA in the composite case [Qian et al., 2021a], where the dependence on δ is $1/\delta$ only. This leads to the following question:

> *Can we design provably accelerated gradient-type methods that work with contractive compressors and the dependence on the compressor parameter δ is $1/\delta$.*

Let us first recall the results for the error-compensated non-accelerated methods. In the composite case, the dependence on the compressor parameter δ of EC-SDCA is better than that of EC-LSVRG [Qian et al., 2021a]. Noticing that L-SVRG [Hofmann et al., 2015, Kovalev et al., 2019] is a primal method and SDCA [Shalev-Shwartz and Zhang, 2012] is a primal-dual method, the better dependence on δ of EC-SDCA than EC-LSVRG indicates that primal-dual methods may be more suitable for the error feedback mechanism. Therefore, it is natural to apply error feedback to

Table 1: Communication Complexity Results for Different Error-Compensated Algorithms (r_Q represents the communication cost of the compressed vector $Q(x)$ for $x \in \mathbb{R}^d$. For simplicity, we choose $Q = Q_1$, and assume $L_f \geq \lambda$, ${}^{R^2}\!/\!{}_{\gamma} \geq \lambda$, where R is defined in Algorithm 4, hence the term ${}^1\!/\!{}_\delta$ is omitted.)

Algorithm	Communication complexity when $\delta \leq {}^1\!/\!{}_m$	Communication complexity under Assumption 2.3 when $\delta \leq {}^1\!/\!{}_m$
EC-LSVRG Smooth Case [Qian et al., 2021a]	$\mathcal{O}\left(\frac{r_Q}{\delta}\frac{\sqrt{L_f\bar{L}}}{\lambda}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{r_Q}{\delta}\frac{L_f}{\lambda}\log\frac{1}{\epsilon}\right)$
EC-SDCA [Qian et al., 2021a]	$\mathcal{O}\left(\frac{r_Q}{\delta}\frac{R\bar{R}}{\lambda\gamma}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{r_Q}{\delta}\frac{R^2}{\lambda\gamma}\log\frac{1}{\epsilon}\right)$
ECLK [Qian et al., 2021c]	$\mathcal{O}\left(\frac{r_Q}{\delta\sqrt{\delta}}\sqrt{\frac{\bar{L}}{\lambda}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{r_Q}{\delta\sqrt{\delta}}\sqrt{\frac{L_f}{\lambda}}\log\frac{1}{\epsilon}\right)$
ECSPDC **This work**	$\mathcal{O}\left(\frac{r_Q}{\delta^2\sqrt{m}}\sqrt{\frac{\bar{R}^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{r_Q}{\delta^2\sqrt{m}}\sqrt{\frac{R^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$
EC-LSVRG + Catalyst Smooth Case **This work**	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{\bar{L}}{\lambda}}\log\frac{1}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{L_f}{\lambda}}\log\frac{1}{\epsilon}\right)$
EC-SDCA + Catalyst **This work**	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{\bar{R}^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{R^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$

Table 2: Communication Complexity Results for EC-LSVRG + Catalyst in the Smooth Case (r_Q represents the communication cost of the compressed vector $Q(x)$ for $x \in \mathbb{R}^d$. The common Assumptions 1.1 and 2.1 are omitted. $A_1 := \delta L_f + {}^{\delta L}\!/\!{}_n + \sqrt{1-\delta}(\sqrt{L_f\bar{L}} + \sqrt{\delta L_f L})$ and $A_2 := \delta L_f + {}^{\delta L}\!/\!{}_n + \sqrt{1-\delta}L_f$.)

Assumptions	Communication complexity
$A_1 \geq \lambda$ $\kappa = A_1 - \lambda$	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\sqrt{\lambda}}\left(\sqrt{\frac{L_f}{\delta}} + \sqrt{\frac{L}{\delta n}} + \frac{(\sqrt{1-\delta}(\sqrt{L_f\bar{L}}+\sqrt{\delta L_f L}))^{\frac{1}{2}}}{\delta} + \frac{\sqrt{1-\delta}(\sqrt{\bar{L}}+\sqrt{\delta L})}{\delta}\right)\log\frac{1}{\epsilon}\right)$
$A_1 < \lambda$ $\kappa = 0$	$\tilde{\mathcal{O}}\left(r_Q\left(\frac{1}{\delta} + \frac{L_f}{\lambda} + \frac{L}{n\lambda} + \frac{\sqrt{1-\delta}(\sqrt{L_f\bar{L}}+\sqrt{\delta L_f L})}{\delta\lambda} + \frac{\sqrt{1-\delta}(\sqrt{\bar{L}}+\sqrt{\delta L})}{\delta\sqrt{\lambda}}\right)\log\frac{1}{\epsilon}\right)$
Assumption 2.3 $A_2 \geq \lambda$ $\kappa = A_2 - \lambda$	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\sqrt{\lambda}}\left(\sqrt{\frac{L_f}{\delta}} + \sqrt{\frac{L}{\delta n}} + \frac{(1-\delta)^{\frac{1}{4}}\sqrt{L_f}}{\delta}\right)\log\frac{1}{\epsilon}\right)$
Assumption 2.3 $A_2 < \lambda$ $\kappa = 0$	$\tilde{\mathcal{O}}\left(r_Q\left(\frac{1}{\delta} + \frac{L_f}{\lambda} + \frac{L}{n\lambda} + \frac{\sqrt{1-\delta}L_f}{\delta\lambda}\right)\log\frac{1}{\epsilon}\right)$

SPDC [Zhang and Xiao, 2017], which is an accelerated primal-dual algorithm, and expect better dependence on the compressor parameter than ECLK. We first propose error-compensated SPDC (Algorithm 4), but unfortunately, we show that the dependence on δ of error-compensated SPDC is at least ${}^1\!/\!{}_{\delta^{\frac{3}{2}}}$. This fact makes us to consider the indirect accelerated methods.

In this work, we give a confirmed answer to the above question by applying Catalyst [Lin et al., 2015], which is a generic method for accelerating first-order algorithms in the sense of Nesterov, to non-accelerated error-compensated methods, where the dependence on δ of the communication complexity could be $\tilde{\mathcal{O}}\left({}^1\!/\!{}_\delta\right)$. Here $\tilde{\mathcal{O}}$ hides some logarithmic terms.

More specific, we use EC-LSVRG and EC-SDCA [Qian et al., 2021a] in Section 3 and Section 4, respectively, to solve the subproblem in Catalyst [Lin et al., 2015]. A key point in Catalyst is the initialization of the algorithm which solves the subproblem. In both Section 3 and Section 4, we first use some naive initialization ways for EC-LSVRG and EC-SDCA. However, then the expected total communication cost will involve an additional term ${}^{U_d}\!/\!{}_{r_Q}$ (defined in Section 3.1), which may be much larger than ${}^1\!/\!{}_\delta$. While the additional term ${}^{U_d}\!/\!{}_{r_Q}$ in the expected total communication cost is actually caused by communicating uncompressed vectors, to avoid this additional term, we propose new initialization ways where only compressed vectors are communicated.

1.2 Contributions

1, First, we propose the error-compensated SPDC (ECSPDC), which is a combination of the error feedback machanism and SPDC [Zhang and Xiao, 2017], and achieve the accelerated linear convergence rate. In the special case where $\delta = 1$, ECSPDC is also an extension of SPDC in the sense that $A_{i\tau}$ in problem (4) is a matrix rather than a vector, and the convergence rate is actually better than SPDC. Specifically, their convergence result does not achieve linear

Table 3: Communication Complexity Results for EC-SDCA + Catalyst (r_Q represents the communication cost of the compressed vector $Q(x)$ for $x \in \mathbb{R}^d$. The common Assumptions 2.1 and 2.2 are omitted. $A_3 := (\frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R\bar{R}}{\delta\gamma} + \frac{\sqrt{1-\delta}RR_m}{\sqrt{\delta}\gamma})/(\frac{1}{\delta}+m)$ and $A_4 := (\frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R^2}{\delta\gamma})/(\frac{1}{\delta}+m)$, where $R, \bar{R}, R_m$ are defined in Algorithm 4.)

Assumptions	Communication complexity
$A_3 \geq \lambda$ $\kappa = A_3 - \lambda$	$\tilde{\mathcal{O}}\left(\frac{r_Q \log\frac{1}{\epsilon}}{\sqrt{\lambda}}\left(\sqrt{\frac{1+\delta m}{\delta}}\sqrt{\frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R\bar{R}}{\delta\gamma} + \frac{\sqrt{1-\delta}RR_m}{\sqrt{\delta}\gamma}} + \sqrt{\frac{(1-\delta)(\bar{R}^2+\delta R_m^2)}{\delta^2\gamma}}\right)\right)$
$A_3 < \lambda$ $\kappa = 0$	$\tilde{\mathcal{O}}\left(r_Q\left(\frac{1}{\delta} + m + \frac{R_m^2}{\lambda n\gamma} + \frac{R^2}{\lambda\gamma} + \frac{\sqrt{1-\delta}R\bar{R}}{\delta\lambda\gamma} + \frac{\sqrt{1-\delta}RR_m}{\sqrt{\delta}\lambda\gamma} + \sqrt{\frac{(1-\delta)(\bar{R}^2+\delta R_m^2)}{\delta^2\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right)$
Assumption 2.3 $A_4 \geq \lambda$ $\kappa = A_4 - \lambda$	$\tilde{\mathcal{O}}\left(\frac{r_Q}{\sqrt{\lambda}}\left(\sqrt{\frac{1}{\delta}+m}\sqrt{\frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R^2}{\delta\gamma}}\right)\log\frac{1}{\epsilon}\right)$
Assumption 2.3 $A_4 < \lambda$ $\kappa = 0$	$\tilde{\mathcal{O}}\left(r_Q\left(\frac{1}{\delta} + m + \frac{R_m^2}{n\lambda\gamma} + \frac{R^2}{\lambda\gamma} + \frac{\sqrt{1-\delta}R^2}{\delta\lambda\gamma}\right)\log\frac{1}{\epsilon}\right)$

speed up with respect to the number of nodes, while ours can obtain linear speed up when the number of nodes is in a certain range.

2, We apply Catalyst [Lin et al., 2015] to EC-LSVRG in the smooth case and EC-SDCA in the composite case [Qian et al., 2021a], respectively. The accelerated linear convergence rates are obtained for both cases, and the dependence on δ of the communication complexities is $\tilde{\mathcal{O}}(1/\delta)$, which matches the best dependence on the compressor parameter in non-accelerated error-compensated methods up to logarithmic terms. The communication complexities of them are summarized in Table 2 and Table 3, and the comparison of the communication complexity results of different error-compensated algorithms when $\delta \leq 1/m$ are summarized in Table 1.

2 ERROR COMPENSATED SPDC

For primal-dual methods, the following problem is usually studied:

$$\min_{x\in\mathbb{R}^d} P(x) := \frac{1}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\phi_{i\tau}(A_{i\tau}^\top x) + g(x), \qquad (4)$$

where $N = mn$ and $A_{i\tau} \in \mathbb{R}^{d\times t}$. Problem (4) is actually equivalent to Problem (1). First, by choosing $f_i^{(\tau)}(x) = \phi_{i\tau}(A_{i\tau}^\top x)$ and $\psi = g$, Problem (4) is a special case of Problem (1). On the other hand, by choosing $A_{i\tau}$ to be the identity matrix, $\phi_{i\tau} = f_i^{(\tau)}$, and $g = \psi$, Problem (4) becomes Problem (1). For simplicity, we assume $L_f = R^2/\gamma$, $\bar{L} = \bar{R}^2/\gamma$, and $L = R_m^2/\gamma$, where R^2, $\bar{R}^2$, and R_m^2 are defined in Algorithm 4. To save space, we only list the assumptions and main results here. The rest can be found in the Appendix.

Assumption 2.1 *The two compressors Q and Q_1 are contraction compressors with parameters δ and δ_1, respectively.*

Assumption 2.2 *Each $\phi_{i\tau} : \mathbb{R}^t \to \mathbb{R}$ is convex and $1/\gamma$-smooth. The regularizer $g : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex.*

Sometimes, we will use the following assumption on the contraction compressor to get better results.

Assumption 2.3 $\mathbb{E}[Q(x)] = \delta x$ *and* $\mathbb{E}[Q_1(x)] = \delta_1 x$.

Under Assumption 2.1 and Assumption 2.2 , the iteration complexity of ECSPDC is

$$\mathcal{O}\left(\left(\frac{1}{\delta} + \frac{1}{\delta_1} + m + \mathcal{R}_2\sqrt{\frac{m}{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right),$$

where

$$\mathcal{R}_2^2 = 2R^2 + \frac{2R_m^2}{n} + \frac{3(1-\delta)}{4}\left(\frac{14\bar{R}^2}{\delta^2} + \frac{7R_m^2}{2\delta} + \frac{84(1-\delta_1)\bar{R}^2}{\delta^2\delta_1^2 m^2} + \frac{42R_m^2}{\delta^2\delta_1 m^2}\right).$$

If Assumption 2.3 is further invoked, the iteration complexity is improved to $\mathcal{O}\left(\left(\frac{1}{\delta} + \frac{1}{\delta_1} + m + \mathcal{R}_3\sqrt{\frac{m}{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right)$, where

$$\mathcal{R}_3^2 = 2R^2 + \frac{2R_m^2}{n} + \frac{21(1-\delta)}{4}\left(\frac{2R^2}{\delta^2} + \frac{11R_m^2}{2\delta n} + \frac{12(1-\delta)\bar{R}^2}{\delta^2 n}\frac{12R^2}{5\delta^2\delta_1^2 m^2} + \frac{228R_m^2}{5\delta^2\delta_1 m^2 n} + \frac{432(1-\delta_1)\bar{R}^2}{5\delta^2\delta_1^2 m^2 n}\right).$$

Comparison to SPDC. If there is no compression in EC-SPDC, i.e., $\delta = \delta_1 = 1$, the iteration complexity becomes

$$\mathcal{O}\left(\left(\frac{N}{n} + \frac{(\sqrt{n}R+R_m)}{n}\sqrt{\frac{N}{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right),$$

which is better than that of SPDC obtained in [Zhang and Xiao, 2017]: $\mathcal{O}\left(\left(\frac{N}{n} + R_m\sqrt{\frac{N}{n\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right)$. Moreover, our result achieves linear speed up with repect to n when $n \leq R_m^2/R^2$.

Dependence on δ. Consider $Q = Q_1$ in ECSPDC. When $1/m \leq \delta$, the iteration complexity is at least

$\mathcal{O}\left(\frac{\bar{R}}{\delta}\sqrt{\frac{m}{\lambda\gamma}}\log\frac{1}{\epsilon}\right) \geq \mathcal{O}\left(\frac{1}{\delta\sqrt{\delta}}\frac{\bar{R}}{\sqrt{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$. When $\delta \leq {}^{1}/_{m}$, we have $\delta \leq {}^{\bar{R}^2}/_{R_m^2}$. Then the iteration complexity becomes

$$\mathcal{O}\left(\left(\frac{1}{\delta} + \frac{1}{\delta^2\sqrt{m}}\frac{\bar{R}}{\sqrt{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right) \geq \mathcal{O}\left(\frac{1}{\delta\sqrt{\delta}}\frac{\bar{R}}{\sqrt{\lambda\gamma}}\log\frac{1}{\epsilon}\right).$$

Hence, the dependence of ECSPDC on δ is at least ${}^{1}/_{\delta^{\frac{3}{2}}}$.

3 EC-LSVRG + CATALYST IN THE SMOOTH CASE

EC-LSVRG [Qian et al., 2021a] is a combination of L-SVRG algorithm [Hofmann et al., 2015, Kovalev et al., 2019, Qian et al., 2021b] and error feedback technique [Seide et al., 2014], and the iteration complexity has the better dependence on the compressor parameter in the smooth case than that in the non-smooth case. In this section, we apply Catalyst [Lin et al., 2015] to EC-LSVRG in the smooth case. First, we restate the Catalyst algorithm and convergence result as follows.

Algorithm 1 Catalyst

```
**Parameters:** $\kappa \geq 0$, $\alpha_0$, sequence $\{\epsilon_k\}_{k\geq 0}$
**Initialization:** $y^0 = x^0 \in \mathbb{R}^d$; $q = \lambda/(\lambda+\kappa)$
**for** $k = 1, 2, 3, ...$ **do**
    Find an approximate solution of the following problem
        $x^k \approx \arg\min_{x\in\mathbb{R}^d}\left\{G_k(x) := P(x) + \frac{\kappa}{2}\|x - y^{k-1}\|^2\right\}$
        such that $G_k(x^k) - G_k^* \leq \epsilon_k$
    Compute $\alpha_k \in (0,1)$ from equation $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$
    Compute
        $y^k = x^k + \beta_k(x^k - x^{k-1})$ with $\beta_k = \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2+\alpha_k}$
**end for**
```

Theorem 3.1 *[Lin et al., 2015] Choose* $\alpha_0 = \sqrt{q}$ *with* $q = {}^{\lambda}/_{(\lambda+\kappa)}$ *and*

$$\epsilon_k = \frac{2}{9}(P(x^0) - P^*)(1-\rho_0)^k \text{ with } \rho_0 < \sqrt{q}.$$

Then, Algorithm 1 generates iterates $\{x^k\}_{k\geq 0}$ *such that*

$$P(x^k) - P^* \leq C(1-\rho_0)^{k+1}(P(x^0) - P^*). \tag{5}$$

with $C = \frac{8}{(\sqrt{q}-\rho_0)^2}$.

In Catalyst (Algorithm 1), G_k^* represents the minimum of G_k. In Theorem 3.1, P^* is the minimum of P, and as discussed in [Lin et al., 2015], the term $P(x^0) - P^*$ in ϵ_k can be replaced by its upper bound, which only affects the corresponding constant in (5).

We use EC-LSVRG to solve the subproblem in Catalyst for the smooth case where ψ is smooth in Problem (1). The main challenge is proposing suitable initial conditions for the subproblem and estimate the corresponding expected inner iteration number.

To save space, we restate EC-LSVRG (and also EC-SDCA) in the Appendix. It should be noticed that EC-LSVRG in the smooth case is applied to the problem without the regularizer term. Thus, to minimize G_k, we move ψ and the quadratic term $\frac{\kappa}{2}\|x - y^{k-1}\|^2$ to each $f_i^{(\tau)}$. We use subscript (k) and superscript K to denote the variables at the k-th outer iteration and K-th inner iteration (for example, $x_{(k)}^K$, $\bar{x}_{(k)}^K$, $x_{(k)}^*$, $e_{\tau,(k)}^K$, and $h_{\tau,(k)}^K$), respectively.

Next, we consider how to initialize EC-LSVRG to obtain the accelerated convergence rate. In [Lin et al., 2015], the Catalyst acceleration was applied to the first-order methods whose convergence rate has the following form

$$G_k(z_t) - G_k^* \leq A(1-\theta)^t(G_k(z^0) - G_k^*), \tag{6}$$

where A is some constant. If we initialize $h_{\tau,(k)}^0$ by the gradient of $f_i^{(\tau)} + \psi + \frac{\kappa}{2}\|\cdot - y^{k-1}\|^2$ at $x_{(k)}^0$. Then the form of the convergence rate of EC-LSVRG becomes form (6), and we can get the following lemma.

Lemma 3.2 *Under Assumptions 1.1, 2.1 and the premise of Theorem 3.1, let us run EC-LSVRG (Algorithm 2) to minimize* G_k *and output* $x^k := \bar{x}_{(k)}^{T_k}$, *where* $T_k := \inf\{K \geq 1, G_k(\bar{x}_{(k)}^K) - G_k^* \leq \epsilon_k\}$. *For the initialization of EC-LSVRG at the* k*-th outer iteration, we choose* $p = \Theta(\delta_1)$, $x_{(k)}^0 = x^{k-1}$, $e_{\tau,(k)}^0 = 0$ *and* $h_{\tau,(k)}^0 = \nabla f^{(\tau)}(x_{(k)}^0) + \nabla\psi(x_{(k)}^0) + \kappa(x_{(k)}^0 - y^{k-1})$. *Then*

$$\mathbb{E}[T_k] \leq \tilde{\mathcal{O}}\left(\frac{1}{\delta} + \frac{1}{\delta_1} + \frac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(\bar{L}+\lambda+\kappa)}}{\delta(\lambda+\kappa)}\right.$$
$$\left. + \frac{L_f}{\lambda+\kappa} + \frac{L}{n(\lambda+\kappa)} + \frac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(L+\lambda+\kappa)}}{\sqrt{\delta}(\lambda+\kappa)}\right),$$

where the notation $\tilde{\mathcal{O}}$ *hides some universal constants and some logarithmic dependencies in* δ, δ_1, λ, κ, L_f, *and* N.

Remark 3.3 *1, It is easy to verify that an optimal choice of* p *in EC-LSVRG is* $\Theta(\delta_1)$. *Hence, we choose* $p = \Theta(\delta_1)$ *in Lemma 3.2 (and also in Lemma 3.4) for simplicity.*

2, As discussed in [Lin et al., 2015], the stopping criteria in the inner loop can be checked by calculating some upper bound of $G_k(\bar{x}_{(k)}^K) - G_k^*$, *such as the duality gap. However, this would cause additional computation and also communication cost. Hence, we can actually view the inner iteration number as a parameter and use Lemma 3.2 as the guidance.*

If we further invoke Assumption 2.3, we can get the following lemma. Since the proof is similar to that of Lemma 3.2, we omit it.

Lemma 3.4 *Under Assumptions 1.1, 2.1, 2.3, and the premise of Theorem 3.1, let us run EC-LSVRG to minimize G_k. Choose the output x^k, T_k, and the initialization of EC-LSVRG at the k-th outer iteration be the same as that in Lemma 3.2. Then*

$$\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\left(\tfrac{1}{\delta} + \tfrac{1}{\delta_1} + \tfrac{L_f}{\lambda+\kappa} + \tfrac{L}{n(\lambda+\kappa)} + \tfrac{\sqrt{1-\delta}(L_f+\lambda+\kappa)}{\delta(\lambda+\kappa)}\right).$$

3.1 Communication Complexity

In this subsection, we discuss the total communication cost by using EC-LSVRG + Catalyst. Same as the claim in the discussion of the communication complexity of ECLK, for simplicity, we choose $Q = Q_1$ in EC-LSVRG.

Denote the communication cost of an vector in $\mathbb{R}^d$ as U_d and the communication cost of the compressed vector in $\mathbb{R}^d$ by using the compressor Q as r_Q. From Theorem 3.1, to achieve $P(x^k) - P^* \le \epsilon$, the outer iteration number is $\tilde{\mathcal{O}}\left(\frac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}} \log \frac{1}{\epsilon}\right)$, and from Lemma 3.2, the expected inner iteration number is

$$\begin{aligned}&\tilde{\mathcal{O}}\Bigg(\tfrac{1}{\delta} + \tfrac{L_f+L/n}{\lambda+\kappa} + \tfrac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(\bar{L}+\lambda+\kappa)}}{\delta(\lambda+\kappa)} \\ &\qquad + \tfrac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(L+\lambda+\kappa)}}{\sqrt{\delta}(\lambda+\kappa)}\Bigg) \\ &= \tilde{\mathcal{O}}\left(\tfrac{1}{\delta} + \tfrac{a_1}{\lambda+\kappa} + \tfrac{b_1}{\sqrt{\lambda+\kappa}}\right),\end{aligned}$$

where we denote $a_1 := L_f + \frac{L}{n} + \frac{\sqrt{1-\delta}(\sqrt{L_f\bar{L}}+\sqrt{\delta L_f L})}{\delta}$ and $b_1 := \frac{\sqrt{1-\delta}\left(\sqrt{\bar{L}}+\sqrt{\delta L}\right)}{\delta}$. Noticing that at each outer iteration, we need to communicate the uncompressed vector $h^0_{\tau,(k)}$, the expected total communication cost becomes

$$\begin{aligned}&\tilde{\mathcal{O}}\left(\left(\tfrac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}}\left(\tfrac{1}{\delta} + \tfrac{a_1}{\lambda+\kappa} + \tfrac{b_1}{\sqrt{\lambda+\kappa}}\right) r_Q + \tfrac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}} U_d\right) \log \tfrac{1}{\epsilon}\right) \\ &= \tilde{\mathcal{O}}\left(\tfrac{r_Q}{\sqrt{\lambda}} \log \tfrac{1}{\epsilon}\left(\left(\tfrac{1}{\delta} + \tfrac{U_d}{r_Q}\right)\sqrt{\lambda+\kappa} + \tfrac{a_1}{\sqrt{\lambda+\kappa}} + b_1\right)\right).\end{aligned}$$

Optimal κ. Since $\kappa \ge 0$ in Catalyst, it is easy to get the optimal κ for minimizing the expected total communication cost. Let $\lambda_1 := a_1 / \left(\frac{1}{\delta} + \frac{U_d}{r_Q}\right)$. If $\lambda \le \lambda_1$, then the optimal κ is $\lambda_1 - \lambda$. If $\lambda > \lambda_1$, then the optimal κ is 0. Or equivalently, the optimal $\kappa = \max\{\lambda_1, \lambda\} - \lambda$.

Similarly, under the additional Assumption 2.3, from Theorem 3.1 and Lemma 3.4, the expected total communication cost is

$$\tilde{\mathcal{O}}\left(\tfrac{r_Q}{\sqrt{\lambda}} \log \tfrac{1}{\epsilon}\left(\left(\tfrac{1}{\delta} + \tfrac{U_d}{r_Q}\right)\sqrt{\lambda+\kappa} + \tfrac{a_2}{\sqrt{\lambda+\kappa}}\right)\right),$$

where $a_2 := L_f + \frac{L}{n} + \frac{\sqrt{1-\delta}L_f}{\delta}$. Let $\lambda_2 := a_2 / \left(\frac{1}{\delta} + \frac{U_d}{r_Q}\right)$. Then the optimal $\kappa = \max\{\lambda_2, \lambda\} - \lambda$.

For TopK, if we use 64 bits for each element in $\mathbb{R}^d$, $\frac{U_d}{r_Q} = \frac{64d}{(64+\log d)K} = \Theta\left(\frac{d}{K \log d}\right)$. Even though the theoretical δ for TopK is ${K}/{d}$, the actual value could be much larger than ${K}/{d}$ in practice. Then ${U_d}/{r_Q}$ may not be able to be bounded by $\mathcal{O}({1}/{\delta})$, and thus the communication complexity may be even worse than ECLK and ECSPDC. Next, we consider how to avoid the additional term ${U_d}/{r_Q}$ in the expected total communication cost.

3.2 Remove the Dependence on ${U_d}/{r_Q}$

Due to the communication of uncompressed vectors at each outer iteration of the stratergies in Lemmas 3.2 and 3.4, the expected total communication complexities depend on ${U_d}/{r_Q}$, which may be much larger than ${1}/{\delta}$. In this subsection, we show that we can actually remove the dependence on ${U_d}/{r_Q}$ by communicating the compressed vector only. The initialization procedures and estimations of the expected inner iteration number are states in the following two lemmas.

Lemma 3.5 *Under Assumptions 1.1, 2.1, and the premise of Theorem 3.1, let us run EC-LSVRG to minimize G_k and output $x^k := x^{T_k}_{(k)}$, $h^{T_k}_{\tau,(k)}$, and $e^{T_k}_{\tau,(k)}$, where $T_k := \inf\{K \ge 1, \Phi^K_{3,(k)} + G_k(x^K_{(k)}) - G^*_k \le \epsilon_k\}$. For the initialization of EC-LSVRG at the k-th outer iteration, we choose $p = \Theta(\delta_1)$, $x^0_{(k)} = x^{k-1}$, $e^0_{\tau,(k)} = 0$ or $e^{T_{k-1}}_{\tau,(k-1)}$, and $h^0_{\tau,(k)} = h^{T_{k-1}}_{\tau,(k-1)}$ or $h^{T_{k-1}}_{\tau,(k-1)} + \kappa(y^{k-2} - y^{k-1})$ ($y^{-1} = y^0$). Then*

$$\begin{aligned}\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\Bigg(&\tfrac{1}{\delta} + \tfrac{1}{\delta_1} + \tfrac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(\bar{L}+\lambda+\kappa)}}{\delta(\lambda+\kappa)} \\ &+ \tfrac{L_f}{\lambda+\kappa} + \tfrac{L}{n(\lambda+\kappa)} + \tfrac{\sqrt{(1-\delta)(L_f+\lambda+\kappa)(L+\lambda+\kappa)}}{\sqrt{\delta}(\lambda+\kappa)}\Bigg),\end{aligned}$$

where the notation $\tilde{\mathcal{O}}$ hides some universal constants and some logarithmic dependencies in δ, δ_1, λ, κ, L_f, and N.

Lemma 3.6 *Under Assumptions 1.1, 2.1, 2.3, and the premise of Theorem 3.1, let us run EC-LSVRG to minimize G_k and output $x^k := x^{T_k}_{(k)}$, $h^{T_k}_{\tau,(k)}$, and $e^{T_k}_{\tau,(k)}$, where $T_k := \inf\{K \ge 1, \Phi^K_{4,(k)} + G_k(x^K_{(k)}) - G^*_k \le \epsilon_k\}$. Choose the initialization of EC-LSVRG at the k-th outer iteration be the same as that in Lemma 3.5. Then $\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\left(\frac{1}{\delta} + \frac{1}{\delta_1} + \frac{L_f}{\lambda+\kappa} + \frac{L}{n(\lambda+\kappa)} + \frac{\sqrt{(1-\delta)}(L_f+\lambda+\kappa)}{\delta(\lambda+\kappa)}\right)$.*

Communication Complexity. Same as the analysis in Section 3.1, the expected total communication cost of EC-LSVRG + Catalyst with the output and initialization precedures in Lemmas 3.5 and 3.6 can be obtained by simply replacing ${U_d}/{r_Q}$ with 0. It is evident that the communication complexity depends on ${1}/{\delta}$ only up to logarithmic terms. In particular, if $1 - \delta = \Theta(1)$, $\delta \le \min\{{\bar{L}}/{L}, {n^2 L_f}/{L}\}$ and $L_f \ge \lambda$, then an optimal κ is $\sqrt{L_f\bar{L}} - \lambda$, and the corresponding communication complexity is $\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{\bar{L}}{\lambda}} \log \frac{1}{\epsilon}\right)$. If Assumption 2.3 is further invoked, when $1-\delta=\Theta(1)$, $\delta \le {}^{nL_f}/_{L}$, and $L_f \ge \lambda$, an optimal κ is $L_f-\lambda$, and the corresponding communication complexity is $\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{L_f}{\lambda}}\log\frac{1}{\epsilon}\right)$.

4 EC-SDCA + CATALYST

In this section, we consider Problem (4). Let $\xi := \frac{1}{\lambda}g$. Then ξ is 1-strongly convex if g is λ-strongly convex. We apply the catalyst to problem (4), and for the subproblem, we use the error-compensated SDCA (Algorithm 3) in [Qian et al., 2021a] to solve it. At the k-th outer iteration, we use EC-SDCA to minimize $G_k(x) := P(x) + \frac{\kappa}{2}\|x-y^{k-1}\|^2$, and we also use subscript (k) and superscript K to denote the variables at the k-th outer iteration and K-th inner iteration (for instance, $x^K_{(k)}$, $\alpha^K_{(k)}$, $e^K_{\tau,(k)}$, $e^K_{(k)}$, and $u^K_{(k)}$).

To apply EC-SDCA at the k-th outer iteration in Algorithm 1, we need to initialize $\alpha^0_{i\tau,(k)}$. It is natural to use the values of $\alpha_{i\tau}$ in the last inner loop to initialize $\alpha^0_{i\tau,(k)}$, and this is indeed the case in [Shalev-Shwartz and Zhang, 2014], where the accelerated SDCA was studied. Then in order to initialize $u^0_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha^0_{i\tau,(k)}$, the uncompressed vector $A_{i\tau}\alpha^0_{i\tau,(k)}$ need to be communicated. We state the initialization procedures formally and estimate the expected inner iteration number in the next two lemmas.

Lemma 4.1 *Assume $\delta < 1$. Under Assumptions 2.1, 2.2, and the premise of Theorem 3.1, let us run EC-SDCA (Algorithm 3) to minimize G_k and output $(x^k,\alpha^k) := (x^{T_k+1}_{(k)}, \alpha^{T_k}_{(k)})$, where $T_k := \inf\{K \ge 1, \sqrt{4n+\delta mn}\Psi^K_{3,(k)} + 2(G_k(x^{K+1}_{(k)}) - G^*_k) \le \epsilon_k\}$. For the initialization of EC-SDCA at the k-th iteration, we choose $\alpha^0_{(k)} = \alpha^{k-1}$ ($\alpha^0=0$) and $e^0_{\tau,(k)}=0$. Then*

$$\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\left(\frac{1}{\delta} + m + \frac{a_3}{\lambda+\kappa} + \frac{b_3}{\sqrt{\lambda+\kappa}}\right),$$

where $a_3 := \frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R\bar{R}}{\delta\gamma} + \frac{\sqrt{1-\delta}RR_m}{\sqrt{\delta}\gamma}$, $b_3 := \frac{1}{\delta}\sqrt{\frac{(1-\delta)(\bar{R}^2+\delta R_m^2)}{\gamma}}$ and the notation $\tilde{\mathcal{O}}$ hides some universal constants and some logarithmic dependencies in δ, λ, κ, R, and N.

Remark 4.2 *In EC-SDCA, ${}^{R^2}/_{\gamma} \ge \lambda+\kappa$ is assumed. However, by adding the term $\frac{b_3}{\sqrt{\lambda+\kappa}}\log\frac{1}{\epsilon}$ to the iteration complexity, the assumption ${}^{R^2}/_{\gamma} \ge \lambda+\kappa$ is no longer needed, which can be seen easily from the proof of Theorem 3.3 in [Qian et al., 2021a].*

If we further invoke Assumption 2.3 on the compressors in EC-SDCA, we can get the following better result. The proof is similar to that of Lemma 4.1, thus we omit it.

Lemma 4.3 *Assume $\delta < 1$. Under Assumptions 2.1, 2.2, 2.3, and the premise of Theorem 3.1, let us run EC-SDCA to minimize G_k and output $(x^k,\alpha^k) := (x^{T_k+1}_{(k)}, \alpha^{T_k}_{(k)})$, where $T_k := \inf\{K \ge 1, 3\sqrt{2+\delta m}\Psi^K_{4,(k)} + 2(G_k(x^{K+1}_{(k)}) - G^*_k) \le \epsilon_k\}$. For the initialization of EC-SDCA at the k-th iteration, we choose $\alpha^0_{(k)} = \alpha^{k-1}$ ($\alpha^0=0$) and $e^0_{\tau,(k)}=0$. Then $\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\left(\frac{1}{\delta}+m+\frac{a_4}{\lambda+\kappa}\right)$, where $a_4 := \frac{R_m^2}{n\gamma} + \frac{R^2}{\gamma} + \frac{\sqrt{1-\delta}R^2}{\delta\gamma}$.*

4.1 Communication Complexity

In this subsection, we discuss the total communication cost by using EC-SDCA + Catalyst. From Theorem 3.1, to get $P(x^k)-P^* \le \epsilon$, the outer iteration number is $\tilde{\mathcal{O}}\left(\frac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}}\log\frac{1}{\epsilon}\right)$, and from Lemma 4.1, the expected inner iteration number is $\tilde{\mathcal{O}}\left(\frac{1}{\delta}+m+\frac{a_3}{\lambda+\kappa}+\frac{b_3}{\sqrt{\lambda+\kappa}}\right)$. Noticing that at each outer iteration, we need to communicate the uncompressed vector to initialize $u^0_{(k)}$, the expected total communication cost is

$$\begin{aligned}&\tilde{\mathcal{O}}\left(\left(\frac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}}\left(\frac{1}{\delta}+m+\frac{a_3}{\lambda+\kappa}+\frac{b_3}{\sqrt{\lambda+\kappa}}\right)r_Q\right.\right.\\&\qquad\left.\left.+\frac{\sqrt{\lambda+\kappa}}{\sqrt{\lambda}}U_d\right)\log\frac{1}{\epsilon}\right)\\&=\tilde{\mathcal{O}}\left(\frac{r_Q}{\sqrt{\lambda}}\log\frac{1}{\epsilon}\left(\left(\frac{1+\delta m}{\delta}+\frac{U_d}{r_Q}\right)\sqrt{\lambda+\kappa}+\frac{a_3}{\sqrt{\lambda+\kappa}}+b_3\right)\right).\end{aligned}$$

Optimal κ. Since $\lambda+\kappa \ge \lambda$, it is easy to obtain the optimal κ for minimizing the expected total communication cost. Let $\lambda_3 := a_3/(\frac{1}{\delta}+m+\frac{U_d}{r_Q})$. Then the optimal κ is $\max\{\lambda,\lambda_3\}-\lambda$.

Similarly, under the additional Assumption 2.3, from Theorem 3.1 and Lemma 4.3, the expected total communication cost is

$$\tilde{\mathcal{O}}\left(\frac{r_Q}{\sqrt{\lambda}}\log\frac{1}{\epsilon}\left(\left(\frac{1}{\delta}+m+\frac{U_d}{r_Q}\right)\sqrt{\lambda+\kappa}+\frac{a_4}{\sqrt{\lambda+\kappa}}\right)\right).$$

Let $\lambda_4 := a_4/(\frac{1}{\delta}+m+\frac{U_d}{r_Q})$. Then the optimal κ is $\max\{\lambda,\lambda_4\}-\lambda$.

The term ${}^{U_d}/_{r_Q}$ also shows up in the expected total communication cost of EC-SDCA + Catalyst. As we analyzed in Section 3.1, the presence of ${}^{U_d}/_{r_Q}$ may make the communication complexity worse than ECLK and ECSPDC. In next subsection, we try to remove the dependence on ${}^{U_d}/_{r_Q}$.

4.2 Remove the Dependence on ${}^{U_d}/_{r_Q}$

As we can see from the analysis of the communication complexity, the term ${}^{U_d}/_{r_Q}$ shows up because of the communication of uncompressed vectors. Hence, in order to remove the dependence on ${}^{U_d}/_{r_Q}$, we need to find initialization procedures that do not need the communication of uncompressed vectors. Fortunately, by investigating the proofs of EC-SDCA, we find out that the relation $u^0_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha^0_{i\tau,(k)}$ in the

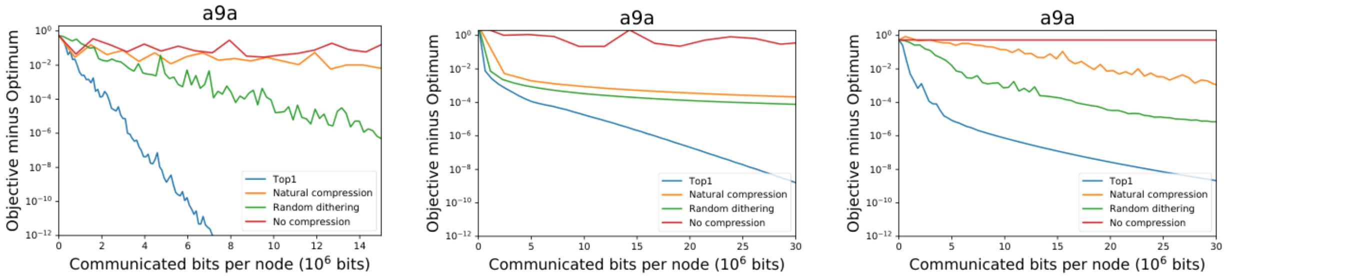

Figure 1: The Communication Complexity Performance of ECSDCA-Catalyst, ECLSVRG-Catalyst, and ECSPDC Used with Compressors: Top1 VS Random Dithering VS Natural Compression VS No Compression on `a9a` Data Set

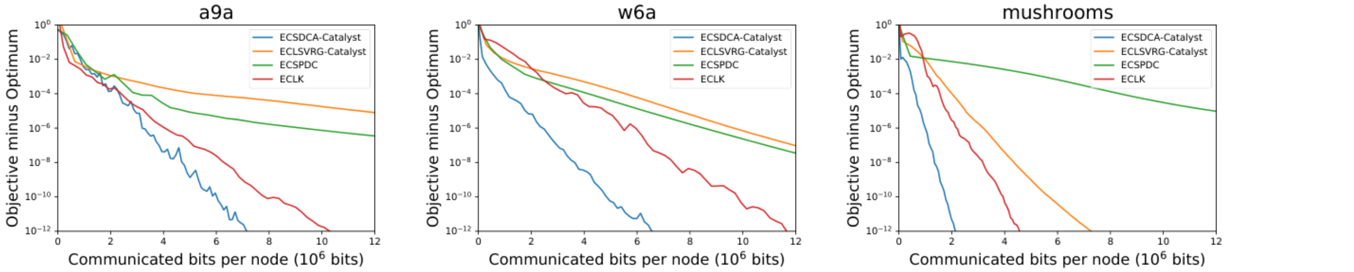

Figure 2: The Communication Complexity Performance of ECSDCA-Catalyst VS ECLSVRG-Catalyst VS ECSPDC VS ECLK for Top1 Compressor on `a9a`, `w6a`, and `mushrooms` Data Sets

initialization is not necessary, and the relation $\tilde{u}^K_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^{n}\sum_{i=1}^{m} A_{i\tau}\alpha^K_{i\tau,(k)}$ is actually essential in the proofs, and need to be maintained. This leads to the initialization procedures in the next two lemmas, and the communication of uncompressed vectors is actually not needed for the initialization at each outer iteration.

Lemma 4.4 *Assume $\delta < 1$. Under Assumptions 2.1, 2.2, and the premise of Theorem 3.1, let us run EC-SDCA to minimize G_k and output $x^k := x^{T_k+1}_{(k)}$, $\alpha^k := \alpha^{T_k}_{(k)}$, $u^{T_k}_{(k)}$, and $e^{T_k}_{\tau,(k)}$, where $T_k := \inf\{K \geq 1, \sqrt{4n+\delta mn}\Psi^K_{3,(k)} + 2(G_k(x^{K+1}_{(k)}) - G^*_k) \leq \epsilon_k\}$. For the initialization of EC-SDCA at the k-th iteration, we choose $\alpha^0_{(k)} = \alpha^{k-1}$ ($\alpha^0 = 0$), $u^0_{(k)} = u^{T_{k-1}}_{(k-1)}$ ($u^0_{(1)} = 0$), and $e^0_{\tau,(k)} = e^{T_{k-1}}_{\tau,(k-1)}$ ($e^0_{\tau,(1)} = 0$). Then $\mathbb{E}[T_k] \leq \tilde{\mathcal{O}}\left(\frac{1}{\delta} + m + \frac{a_3}{\lambda+\kappa} + \frac{b_3}{\sqrt{\lambda+\kappa}}\right)$.*

Lemma 4.5 *Assume $\delta < 1$. Under Assumptions 2.1, 2.2, 2.3, and the premise of Theorem 3.1, let us run EC-SDCA to minimize G_k and output $x^k := x^{T_k+1}_{(k)}$, $\alpha^k := \alpha^{T_k}_{(k)}$, $u^{T_k}_{(k)}$, and $e^{T_k}_{\tau,(k)}$, where $T_k := \inf\{K \geq 1, 3\sqrt{2+\delta m}\Psi^K_{4,(k)} + 2(G_k(x^{K+1}_{(k)}) - G^*_k) \leq \epsilon_k\}$. Choose the initialization of EC-SDCA at the k-th iteration be the same as that in Lemma 4.4. Then $\mathbb{E}[T_k] \leq \tilde{\mathcal{O}}\left(\frac{1}{\delta} + m + \frac{a_4}{\lambda+\kappa}\right)$.*

Communication Complexity. Same as the analysis in Section 4.1, the expected total communication cost of EC-SDCA + Catalyst with the output and initialization precedures in Lemmas 4.4 and 4.5 can be obtained by simply replacing ${}^{U_d}/_{r_Q}$ with 0, and only depends on ${}^1/_\delta$ up tp logarithmic terms. In particular, if $\delta \leq {}^1/_m$ and ${}^{R\bar{R}}/_\gamma \geq \lambda$, then an optimal κ is ${}^{R\bar{R}}/_\gamma - \lambda$, and the corresponding communication complexity is $\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{\bar{R}^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$. If Assumption 2.3 is further invoked, when $\delta \leq {}^1/_m$ and ${}^{R^2}/_\gamma \geq \lambda$ an optimal κ is ${}^{R^2}/_\gamma - \lambda$, and the corresponding communication complexity is $\tilde{\mathcal{O}}\left(\frac{r_Q}{\delta}\sqrt{\frac{R^2}{\lambda\gamma}}\log\frac{1}{\epsilon}\right)$.

5 EXPERIMENTS

In this section, we implement our algorithms on the real world binary logistic regression tasks:

$$x \mapsto \log\left(1 + \exp(-y_i A_i^\top x)\right) + \frac{\lambda}{2}\|x\|^2,$$

where A_i, y_i are training sample pairs. We use the data sets: `a9a`, `w6a`, `phishing`, and `mushrooms` from LIBSVM

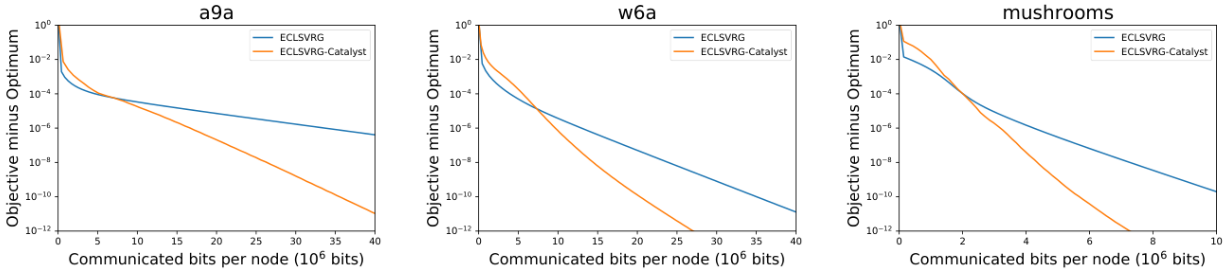

Figure 3: The Communication Complexity Performance of ECLSVRG VS ECLSVRG-Catalyst for Top1 Compressor on `a9a`, `w6a`, and `mushrooms` Data Sets

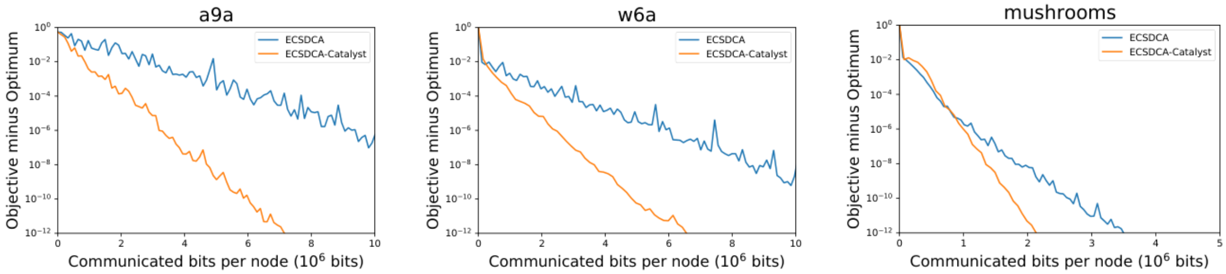

Figure 4: The Communication Complexity Performance of EC-SDCA VS ECSDCA-Catalyst for Top1 Compressor on `a9a`, `w6a`, and `mushrooms` Data Sets

Library [Chang and Lin, 2011]. More experiments can be found in the Appendix.

Compressors. In the experiments, we use Top1 and some contraction compressors transformed by unbiased ones such as random dithering ($s = \sqrt{d}$) and natural compression.

Parameters. We set $\lambda = 1 \times 10^{-5}$ and $n = 20$. For all experiments, we use grid search to obtain the learning rate $\{10^{-t}, t = 0, 1, 2 \cdots\}$. For ECSPDC, we use bisect method to obtain the argmax operator, θ is chosen by Theorem C.7. For ECLSVRG and ECLK, we set $Q = Q_1$ and $p = \delta$. For Catalyst, we choose κ by grid search $\{10^t \lambda : t \in \mathbb{Z}\}$. For the stopping criteria of the inner loop, a heuristic strategy was proposed for Catalyst in [Lin et al., 2015], where the inner loop is constrained to perform at most mn iterations. We employ this strategy similarly and the inner loop size is searched from $\{kd : k = 1, 2, 5, 10, 100\}$, where d is the dimension of data.

5.1 Effectiveness of TopK Compressor

First, we demonstrate the effectiveness of TopK compressor compared with random dithering, natural compression, and no compression. Figure 1 shows that compression can improve the performance with respect to the communication complexity in general, and TopK is specifically effective.

5.2 Comparison of Different Accelerated Error Compensated Algorithms

We compare Catalyst-based error-compensated algorithms and ECSPDC with ECLK, and also use the Top1 compressor. Figure 2 shows that the performance of ECSDCA-Catalyst is the best for our tested data sets, which indicates the potential of the Catalyst-based error-compensated algorithm.

5.3 Improvements from Catalyst Acceleration

In this subsection, we compare Catalyst-based error-compensated algorithms with their baselines, namely, ECSDCA and ECLSVRG, where Top1 compressor is used. Figures 3 and 4 show that Catalyst acceleration can indeed boost the speed of both ECSDCA and ECLSVRG with respect to the communication complexity significantly, which matches our theory.

Acknowledgements

The work of Xun Qian and Peter Richtárik was supported by the Extreme Computing Research Center at KAUST, and the work of Peter Richtárik was also partially supported by the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence.

References

A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720, 2017.

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 (NeurIPS)*, pages 5973–5983, 2018.

J. Bernstein, Y. X. Wang, K. Azizzadenesheli, and A. Anandkumar. Signsgd: Compressed optimisation for nonconvex problems. *The 35th International Conference on Machine Learning*, pages 560–569, 2018.

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

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. In *Neural Information Processing Systems (NeurIPS)*, 2020.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv: 1706.2677*, 2017.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019a.

Samuel Horváth, Chen-Yu Ho, L'udovít Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019b.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

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

Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016b.

D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. *arXiv: 1901.08689*, 2019.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *arXiv preprint arXiv:1506.02186*, 2015.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

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

Xun Qian, Hanze Dong, Peter Richtárik, and Tong Zhang. Error compensated loopless SVRG, Quartz, and SDCA for distributed optimization. *arXiv preprint arXiv:2109.10049*, 2021a.

Xun Qian, Zheng Qu, and Peter Richtárik. L-svrg and l-katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22:1–49, 2021b.

Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021c. URL `https://openreview.net/forum?id=dSqtddFibt2`.

B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

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. *Fifteenth Annual*

Conference of the International Speech Communication Association, 2014.

S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *arXiv: 1211.2717*, 2012.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International conference on machine learning*, pages 64–72. PMLR, 2014.

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

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21:1–36, 2020.

H. Tang, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6155–6165, 2019.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *Automatic Control, IEEE Transactions on*, 31(9):803–812, 1986.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *ACM Computing Surveys*, 2019.

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

Y. You, I. Gitman, and B. Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv: 1708.03888*, 2017.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18:1–42, 2017.

Contents

Appendix

A EXTRA EXPERIMENTS

A.1 Effectiveness of TopK Compressor

We demonstrate the effectiveness of TopK compressor compared with random dithering, natural compression, and no compression. Figures 5, 6, and 7 show that compression can improve the performance with respect to the communication complexity in general, and TopK is specifically effective.

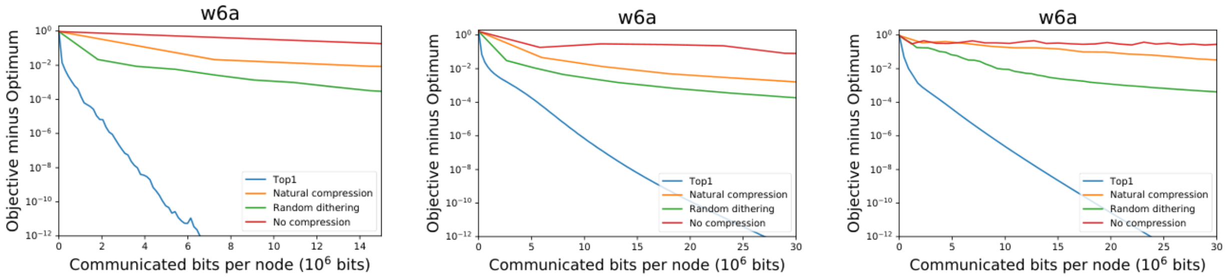

Figure 5: The Communication Complexity Performance of ECSDCA-Catalyst, ECLSVRG-Catalyst, and ECSPDC Used with Compressors: Top1 VS Random Dithering VS Natural Compression VS No Compression on `w6a` Data Set

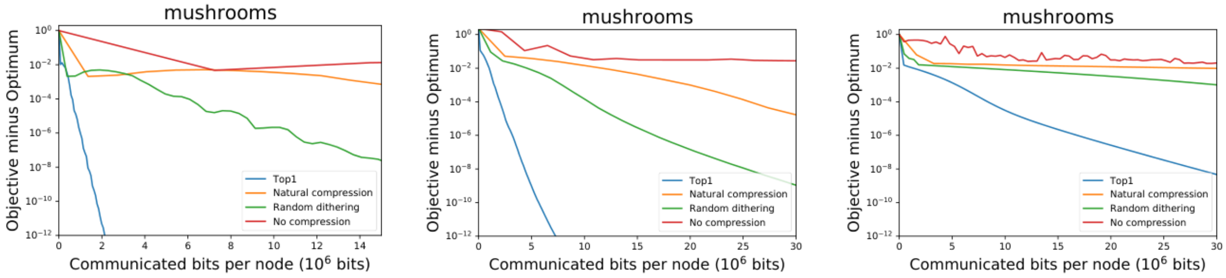

Figure 6: The Communication Complexity Performance of ECSDCA-Catalyst, ECLSVRG-Catalyst, and ECSPDC Used with Compressors: Top1 VS Random Dithering VS Natural Compression VS No Compression on `mushrooms` Data Set

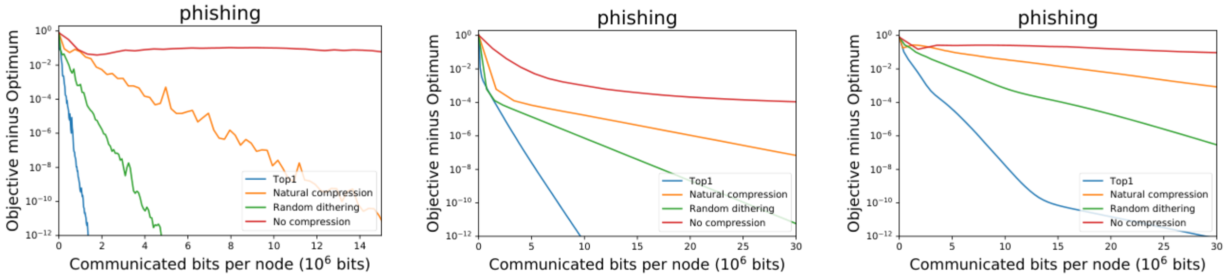

Figure 7: The Communication Complexity Performance of ECSDCA-Catalyst, ECLSVRG-Catalyst, and ECSPDC Used with Compressors: Top1 VS Random Dithering VS Natural Compression VS No Compression on `phishing` Data Set.

B EC-LSVRG AND EC-SDCA ALGORITHMS

In this section, we restate the two algorithms: EC-LSVRG and EC-SDCA in [Qian et al., 2021a].

Algorithm 2 Error compensated loopless SVRG (EC-LSVRG)

```
Parameters: stepsize $\eta > 0$; probability $p \in (0, 1]$
Initialization: $x^0 = w^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$; $h_\tau^0 \in \mathbb{R}^d$; $h^0 = \frac{1}{n}\sum_{\tau=1}^n h_\tau^0$
for $k = 0, 1, 2, \ldots$ do
    for $\tau = 1, \ldots, n$ do
        Sample $i_k^\tau$ uniformly and independently in $[m]$ on each node
        $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k) + \nabla f^{(\tau)}(w^k) - h_\tau^k$
        $y_\tau^k = Q(\eta g_\tau^k + e_\tau^k), \quad e_\tau^{k+1} = e_\tau^k + \eta g_\tau^k - y_\tau^k$
        $z_\tau^k = Q_1(\nabla f^{(\tau)}(w^k) - h_\tau^k), \quad h_\tau^{k+1} = h_\tau^k + z_\tau^k$
        $u_\tau^{k+1} = 0$ for $\tau = 2, \ldots, n$
        $u_1^{k+1} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$
        Send $y_\tau^k$, $z_\tau^k$, and $u_\tau^{k+1}$ to the other nodes
        Receive $y_\tau^k$, $z_\tau^k$, and $u_\tau^{k+1}$ from the other nodes
        $y^k = \frac{1}{n}\sum_{\tau=1}^n y_\tau^k, \quad z^k = \frac{1}{n}\sum_{\tau=1}^n z_\tau^k$
        $u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1}$
        $x^{k+0.5} = x^k - (y^k + \eta h^k)$
        $x^{k+1} = \mathrm{prox}_{\eta\psi}(x^{k+0.5})$
        $w^{k+1} = \begin{cases} x^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$
        $h^{k+1} = h^k + z^k$
    end for
end for
```

Algorithm 3 Error compensated SDCA (EC-SDCA)

```
**Parameters:** $\theta > 0$; $R_m := \max_{i,\tau} \|A_{i\tau}\|$; $\bar{R}^2 := \max_{\tau\in[n]}\{\frac{1}{m}\lambda_{\max}(\sum_{i=1}^m A_{i\tau}A_{i\tau}^\top)\}$; $R^2 := \frac{1}{N}\lambda_{\max}(\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}A_{i\tau}^\top)$; $p_{i\tau} = \frac{1}{m} \in \mathbb{R}$ for $i \in [m]$ and $\tau \in [n]$; positive constants $v_{i\tau} = R_m^2 + nR^2 \in \mathbb{R}$ for $i \in [m]$ and $\tau \in [n]$
**Initialization:** $\alpha^0 \in \mathbb{R}^{tN}$; $x^0 \in \mathbb{R}^d$; $u^0 = \frac{1}{\lambda N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha_{i\tau}^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$ for $\tau \in [n]$
**for** $k = 0, 1, 2, \ldots$ **do**
    **for** $\tau = 1, \ldots, n$ **do**
        $x^{k+1} = \nabla g^*(u^k)$
        $\alpha_{i\tau}^{k+1} = \alpha_{i\tau}^k$ for $i \in [m]$
        Sample $i_k^\tau$ uniformly and independently in $[m]$ on each node
        $\Delta\alpha_{i_k^\tau\tau}^{k+1} = -\theta p_{i_k^\tau\tau}^{-1}\alpha_{i_k^\tau\tau}^k - \theta p_{i_k^\tau\tau}^{-1}\nabla\phi_{i_k^\tau\tau}(A_{i_k^\tau\tau}^\top x^{k+1})$
        $\alpha_{i_k^\tau\tau}^{k+1} = \alpha_{i_k^\tau\tau}^k + \Delta\alpha_{i_k^\tau\tau}^{k+1}$
        $y_\tau^k = Q\left(\frac{1}{\lambda m}A_{i_k^\tau\tau}\Delta\alpha_{i_k^\tau\tau}^{k+1} + e_\tau^k\right)$
        $e_\tau^{k+1} = e_\tau^k + \frac{1}{\lambda m}A_{i_k^\tau\tau}\Delta\alpha_{i_k^\tau\tau}^{k+1} - y_\tau^k$
        Send $y_\tau^k$ to the other nodes
        Receive $y_\tau^k$ from the other nodes
        $u^{k+1} = u^k + \frac{1}{n}\sum_{\tau=1}^n y_\tau^k$
    **end for**
**end for**
```

C ERROR COMPENSATED SPDC

In problem (4), we can replace each $\phi_{i\tau}(A_{i\tau}^\top x)$ by convex conjugation, i.e.,

$$\phi_{i\tau}(A_{i\tau}^\top x) = \sup_{y \in \mathbb{R}^t} \{\langle y, A_{i\tau}^\top x\rangle - \phi_{i\tau}^*(y)\},$$

where $\phi_{i\tau}^*$ is the conjugate function of $\phi_{i\tau}$. This leads to the following convex-concave saddle point problem

$$\min_{x\in\mathbb{R}^d} \max_{Y\in\mathbb{R}^{tN}} f(x,Y) := g(x) + \tfrac{1}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}(\langle y_{i\tau}, A_{i\tau}^\top x\rangle - \phi_{i\tau}^*(y_{i\tau}))),$$

where $Y = (y_{11}^\top, ..., y_{m1}^\top, ..., y_{n1}^\top, ..., y_{mn}^\top)^\top \in \mathbb{R}^{tN}$ and $y_{i\tau} \in \mathbb{R}^t$.

Algorithm 4 Error Compensated SPDC (ECSPDC)

```
Parameters: stepsize parameters $\sigma > 0$; $\eta > 0$; $\theta \in (0,1)$ ; $R_m := \max_{i,\tau}\|A_{i\tau}\|$; $\bar{R}^2 := \max_{\tau\in[n]}\{\frac{1}{m}\lambda_{\max}(\sum_{i=1}^m A_{i\tau}A_{i\tau}^\top)\}$; $R^2 := \frac{1}{N}\lambda_{\max}(\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}A_{i\tau}^\top)$
Initialization: $x^0 = z^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$; $y_{i\tau}^0 \in \mathbb{R}^t$; $u_\tau^0 = \frac{1}{m}\sum_{i=1}^m A_{i\tau}y_{i\tau}^0$; $h_\tau^0 \in \mathbb{R}^d$; $h^0 = \frac{1}{n}\sum_{\tau=1}^n h_\tau^0$
for $k = 0, 1, 2, ...$ do
    for $\tau = 1, ..., n$ do in parallel
        Sample $i_k^\tau$ uniformly and independently in $[m]$ on each node
        $y_{i\tau}^{k+1} = \begin{cases} \arg\max_{y\in\mathbb{R}^t}\left\{\langle y, A_{i\tau}^\top z^k\rangle - \phi_{i\tau}^*(y) - \frac{1}{2\sigma}\|y - y_{i\tau}^k\|^2\right\} & \text{if } i = i_k^\tau \\ y_{i\tau}^k & \text{if } i \neq i_k^\tau \end{cases}$
        $\Delta_\tau^k = Q(A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k) + u_\tau^k - h_\tau^k + e_\tau^k)$
        $u_\tau^{k+1} = u_\tau^k + \frac{1}{m}A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k)$,    $h_\tau^{k+1} = h_\tau^k + Q_1(u_\tau^k - h_\tau^k)$
        $e_\tau^{k+1} = e_\tau^k + A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k) + u_\tau^k - h_\tau^k - \Delta_\tau^k$
        Send $\Delta_\tau^k$ and $Q_1(u_\tau^k - h_\tau^k)$ to the other nodes
        Receive $\Delta_\tau^k$ and $Q_1(u_\tau^k - h_\tau^k)$ from the other nodes
        $\Delta^k = \frac{1}{n}\sum_{\tau=1}^n \Delta_\tau^k$
        $x^{k+1} = \arg\min_{x\in\mathbb{R}^d}\left\{g(x) + \langle h^k + \Delta^k, x\rangle + \frac{\|x - x^k\|^2}{2\eta}\right\}$
        $h^{k+1} = h^k + \frac{1}{n}\sum_{\tau=1}^n Q_1(u_\tau^k - h_\tau^k)$,    $z^{k+1} = x^{k+1} + \theta(x^{k+1} - x^k)$
    end for
end for
```

Description of error-compensated SPDC (Algorithm 4). In distributed SPDC, the search direction at the k-th iteration is

$$\frac{1}{n}\sum_{\tau=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m} A_{i\tau}y_{i\tau}^k + A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k)\right),$$

where i_k^τ is sampled uniformly and independently in $[m] := \{1, 2, ..., m\}$ on each node. When $y_{i\tau}$ goes to the optimal solution, the term $y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k$ will go to zero, while another term $\frac{1}{m}\sum_{i=1}^m A_{i\tau}y_{i\tau}$ may not. Then in the presence of the compression error, the linear convergence rate could not be achieved by compressing this search direction directly. Hence, like ECLK, we introduce a vector h_τ^k to learn $u_\tau^k = \frac{1}{m}\sum_{i=1}^m A_{i\tau}y_{i\tau}$ iteratively. This learning scheme was first proposed in DIANA [Horváth et al., 2019a] with the unbiased compressor. More precisely, we perform the following update on each node

$$h_\tau^{k+1} = h_\tau^k + Q_1(u_\tau^k - h_\tau^k),$$

where Q_1 is a contraction compressor. Now we apply the compression and error feedback mechanism to

$$u_\tau^k - h_\tau^k + A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k), \tag{7}$$

and add $h^k := \frac{1}{n}\sum_{\tau=1}^n h_\tau^k$ back after aggregation. We use e_τ^k to denote the compression error on each node, and add it to (7) before compression. After compression, e_τ^k is updated by the compression error at the current step:

$$e_\tau^{k+1} = e_\tau^k + u_\tau^k - h_\tau^k + A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k) - Q(e_\tau^k + u_\tau^k - h_\tau^k + A_{i_k^\tau\tau}(y_{i_k^\tau\tau}^{k+1} - y_{i_k^\tau\tau}^k)),$$

where Q is also a contraction compressor. The rest steps are the same as SPDC [Zhang and Xiao, 2017]. Next we introduce some useful variables.

Let $e^k := \frac{1}{n}\sum_{\tau=1}^n e_\tau^k$ and $u^k := \frac{1}{n}\sum_{\tau=1}^n u_\tau^k$ for $k \geq 0$. Define $\tilde{x}^k = x^k - \eta e^k$ for $k \geq 0$. We denote the optimal solution of the above saddle point problem as (x^*, Y^*), where

$$Y^* = ((y_{11}^*)^\top, ..., (y_{m1}^*)^\top, ..., (y_{n1}^*)^\top, ..., (y_{mn}^*)^\top)^\top.$$

Now we are ready to construct some Lyapunov functions. For $k \geq 0$, define

$$\begin{aligned}\Phi_2^k := &\left(\tfrac{1}{2\eta} + \tfrac{\lambda}{4}\right)\|\tilde{x}^k - x^*\|^2 + \left(\tfrac{1}{4\sigma} + \tfrac{\gamma}{2}\right)\tfrac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 + \tfrac{3(\eta+\lambda\eta^2)}{\delta n}\sum_{\tau=1}^n \|e_\tau^k\|^2 \\ &+ f(x^k, Y^*) - f(x^*, Y^*) + m\left(f(x^*, Y^*) - f(x^*, Y^k)\right) + \tfrac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2,\end{aligned}$$

where $Y^k = ((y_{11}^k)^\top, ..., (y_{m1}^k)^\top, ..., (y_{n1}^k)^\top, ..., (y_{mn}^k)^\top)^\top$, and

$$\begin{aligned}\Psi_2^k := &\left(\tfrac{1}{2\eta} + \tfrac{\lambda}{4}\right)\|\tilde{x}^k - x^*\|^2 + \left(\tfrac{1}{4\sigma} + \tfrac{\gamma}{2}\right)\tfrac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 + f(x^k, Y^*) - f(x^*, Y^*) \\ &+ m\left(f(x^*, Y^*) - f(x^*, Y^k)\right) + \tfrac{3(\eta+\lambda\eta^2)}{\delta}\|e^k\|^2 + \tfrac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\sum_{\tau=1}^n \|e_\tau^k\|^2 \\ &+ \tfrac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\|h^k - u^k\|^2 + \tfrac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2.\end{aligned}$$

First, we introduce Lemma C.1, which is useful in the analysis of samplings in distributed systems.

Lemma C.1 *Let $S = \{(i^\tau, \tau) |\ i^\tau$ is chosen from $[m]$ uniformly and independently for all $\tau \in [n]\}$. For any given $w_{i\tau} \in \mathbb{R}^t$ for $i \in [m]$ and $\tau \in [n]$, we have*

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^n A_{i^\tau\tau} w_{i^\tau\tau}\right\|^2\right] \leq \left(\frac{nR^2 + R_m^2}{m}\right)\sum_{\tau=1}^n\sum_{i=1}^m \|w_{i\tau}\|^2.$$

The following two lemmas show the evolution of the error terms $\sum_{\tau=1}^n \|e_\tau^k\|^2$ and $\|e^k\|^2$. The proofs are similar to that of Lemmas 3.4 and B.4 in [Qian et al., 2021c], hence we omit them.

Lemma C.2 *We have*

$$\begin{aligned}\frac{1}{n}\sum_{\tau=1}^n \mathbb{E}_k[\|e_\tau^{k+1}\|^2] \leq \left(1 - \frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2 + \frac{4(1-\delta)}{\delta n}\sum_{\tau=1}^n \|u_\tau^k - h_\tau^k\|^2 \\ + \frac{(1-\delta)}{mn}\left(\frac{4\bar{R}^2}{\delta} + R_m^2\right)\sum_{\tau=1}^n\sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.\end{aligned}$$

Lemma C.3 *Under Assumption 2.3, we have*

$$\begin{aligned}\mathbb{E}_k\|e^{k+1}\|^2 \leq \left(1 - \frac{\delta}{2}\right)\|e^k\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \|e_\tau^k\|^2 + \frac{4(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \|u_\tau^k - h_\tau^k\|^2 \\ + \frac{4(1-\delta)}{\delta}\|u^k - h^k\|^2 + \frac{(1-\delta)}{mn}\left(\frac{4R^2}{\delta} + \frac{5R_m^2}{n}\right)\sum_{\tau=1}^n\sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.\end{aligned}$$

We analyze the evolution of $\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2$ and $\|h^k - u^k\|^2$ in the next two lemmas.

Lemma C.4 *We have*

$$\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}_k[\|h_\tau^{k+1}-u_\tau^{k+1}\|^2] \leq \left(1-\frac{\delta_1}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\|h_\tau^k-u_\tau^k\|^2 + \frac{1}{m^3 n}\left(\frac{2(1-\delta_1)\bar{R}^2}{\delta_1}+R_m^2\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k-y_{i\tau}^k\|^2.$$

Lemma C.5 *Under Assumption 2.3, we have*

$$\mathbb{E}_k\|h^{k+1}-u^{k+1}\|^2 \leq (1-\delta_1)\|h^k-u^k\|^2+\frac{\delta_1}{n^2}\sum_{\tau=1}^{n}\|h_\tau^k-u_\tau^k\|^2 + \frac{1}{m^3 n}\left(\frac{R^2}{\delta_1}+\frac{R_m^2}{n}\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k-y_{i\tau}^k\|^2.$$

The dual problem of problem (4) is

$$\max_{Y\in\mathbb{R}^{tN}} D(Y) := \min_{x\in\mathbb{R}^d} f(x,Y) = -\tfrac{1}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\phi_{i\tau}^*(y_{i\tau}) - g^*\left(-\tfrac{1}{N}AY\right), \tag{8}$$

where g^* is the conjugate function of g and

$$A=[A_{11},...,A_{m1},...,A_{n1},...,A_{mn}]\in\mathbb{R}^{d\times tN}. \tag{9}$$

Recall that $R^2=\frac{1}{N}\lambda_{\max}(\sum_{\tau=1}^{n}\sum_{i=1}^{m}A_{i\tau}A_{i\tau}^\top)=\frac{1}{N}\|A\|^2$. We have the following lemma.

Lemma C.6 *[Lemma 3 in Zhang and Xiao, 2017] Let Assumption 2.2 hold. Then for any point $(x,Y)\in\mathrm{dom}(f(x,Y))$, we have*

$$P(x)\leq f(x,Y^*)+\tfrac{R^2}{2\gamma}\|x-x^*\|^2, \quad \text{and} \quad D(Y)\geq f(x^*,Y)-\tfrac{R^2}{2\lambda N}\|Y-Y^*\|^2.$$

Theorem C.7 *Let Assumption 2.1 and Assumption 2.2 hold. Set $\sigma=\frac{1}{2\mathcal{R}_1}\sqrt{\frac{m\lambda}{\gamma}}$, $\eta=\frac{1}{2\mathcal{R}_1}\sqrt{\frac{\gamma}{m\lambda}}$, and $\theta=1-\min\left\{\frac{1}{m+4\mathcal{R}_1\sqrt{m/(\lambda\gamma)}},\frac{\delta}{6},\frac{\delta_1}{6}\right\}$, where $\mathcal{R}_1>0$ will be chosen later. (i) Let $\mathcal{R}_1^2=\mathcal{R}_2^2:=2R^2+\frac{2R_m^2}{n}+\frac{3(1-\delta)}{4}\left(\frac{14\bar{R}^2}{\delta^2}+\frac{7R_m^2}{2\delta}+\frac{84(1-\delta_1)\bar{R}^2}{\delta^2\delta_1^2m^2}+\frac{42R_m^2}{\delta^2\delta_1m^2}\right)$. Assume $\frac{\mathcal{R}_2^2}{\lambda\gamma}\geq 1$. Then*

$$\mathbb{E}[\Phi_2^k]\leq\epsilon\left(\Phi_2^0+\tfrac{1}{4\sigma n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^0-y_{i\tau}^*\|^2\right),$$

as long as $k\geq\mathcal{O}\left(\left(\frac{1}{\delta}+\frac{1}{\delta_1}+m+\mathcal{R}_2\sqrt{\frac{m}{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right)$. In particular, if $\frac{1}{m}\leq\mathcal{O}(\delta_1)$, then the iteration complexity becomes

$$k\geq\mathcal{O}\left(\left(\left(R+\tfrac{R_m}{\sqrt{n}}+\tfrac{\sqrt{(1-\delta)}\bar{R}}{\delta}+\tfrac{\sqrt{(1-\delta)}R_m}{\sqrt{\delta}}\right)\sqrt{\tfrac{m}{\lambda\gamma}}\tfrac{1}{\delta}+m\right)\log\tfrac{1}{\epsilon}\right).$$

(ii)Let $\mathcal{R}_1^2=\mathcal{R}_3^2$, where

$$\mathcal{R}_3^2:=2R^2+\tfrac{2R_m^2}{n}+\tfrac{21(1-\delta)}{4}\left(\tfrac{2R^2}{\delta^2}+\tfrac{11R_m^2}{2\delta n}+\tfrac{12(1-\delta)\bar{R}^2}{\delta^2 n}\tfrac{12R^2}{5\delta^2\delta_1^2m^2}+\tfrac{228R_m^2}{5\delta^2\delta_1m^2n}+\tfrac{432(1-\delta_1)\bar{R}^2}{5\delta^2\delta_1^2m^2n}\right).$$

Let Assumption 2.3 hold and assume $\frac{\mathcal{R}_3^2}{\lambda\gamma}\geq 1$. Then $\mathbb{E}[\Psi_2^k]\leq\epsilon\left(\Psi_2^0+\frac{1}{4\sigma n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^0-y_{i\tau}^*\|^2\right)$ as long as $k\geq\mathcal{O}\left(\left(\frac{1}{\delta}+\frac{1}{\delta_1}+m+\mathcal{R}_3\sqrt{\frac{m}{\lambda\gamma}}\right)\log\frac{1}{\epsilon}\right)$. If $\frac{1}{m}\leq\mathcal{O}(\delta_1)$, then the iteration complexity becomes

$$k\geq\mathcal{O}\left(\left(\left(R+\tfrac{R_m}{\sqrt{n}}+\tfrac{\sqrt{(1-\delta)}R}{\delta}+\tfrac{\sqrt{(1-\delta)}R_m}{\sqrt{\delta n}}\right)\sqrt{\tfrac{m}{\lambda\gamma}}+\tfrac{1}{\delta}+m\right)\log\tfrac{1}{\epsilon}\right).$$

From Lemma C.6, same as Corollary 4 in [Zhang and Xiao, 2017], we can bound the primal-dual gap in the following theorem.

Theorem C.8 *Let Assumption 2.2 hold. Then we have*

$$P(x^k) - D(Y^k) \leq \left(1 + \tfrac{R^2}{\lambda\gamma}\right) \left(f(x^k, Y^*) - f(x^*, Y^*) + m\left(f(x^*, Y^*) - f(x^*, Y^k)\right)\right).$$

Since $f(x^k, Y^*) - f(x^*, Y^*) + m(f(x^*, Y^*) - f(x^*, Y^k))$ is bounded by Φ_2^k or Ψ_2^k, the iteration complexity of the primal-dual gap can be deduced easily from Theorem C.7. Hence, we omit it.

D PROOFS OF LEMMA C.1, LEMMA C.4, LEMMA C.5, AND THEOREM C.7

D.1 Proof of Lemma C.1

Let $W = (w_{11}^\top, ..., w_{m1}^\top, w_{12}^\top, ..., w_{m2}^\top, ..., w_{n1}^\top, ..., w_{mn}^\top)^\top \in \mathbb{R}^{tN}$. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\tau=1}^n A_{i^\tau\tau} w_{i^\tau\tau}\right\|^2\right] \\
&= \mathbb{E}\left[\sum_{\tau_1\neq\tau_2} \langle A_{i^{\tau_1}\tau_1} w_{i^{\tau_1}\tau_1}, A_{i^{\tau_2}\tau_2} w_{i^{\tau_2}\tau_2}\rangle\right] + \mathbb{E}\left[\sum_{\tau=1}^n \|A_{i^\tau\tau} w_{i^\tau\tau}\|^2\right] \\
&= \sum_{\tau_1\neq\tau_2} \left\langle \frac{1}{m}\sum_{i=1}^m A_{i\tau_1} w_{i\tau_1}, \frac{1}{m}\sum_{j=1}^m A_{j\tau_2} w_{j\tau_2} \right\rangle + \frac{1}{m}\sum_{\tau=1}^n\sum_{i=1}^m \|A_{i\tau} w_{i\tau}\|^2 \\
&= \frac{1}{m^2}\sum_{\tau_1\neq\tau_2}\sum_{i,j=1}^m \langle A_{i\tau_1} w_{i\tau_1}, A_{j\tau_2} w_{j\tau_2}\rangle + \frac{1}{m}\sum_{\tau=1}^n\sum_{i=1}^m \|A_{i\tau} w_{i\tau}\|^2 \\
&= \frac{1}{m^2}\sum_{\tau_1,\tau_2=1}^n\sum_{i,j=1}^m \langle A_{i\tau_1} w_{i\tau_1}, A_{j\tau_2} w_{j\tau_2}\rangle - \frac{1}{m^2}\sum_{\tau=1}^n\sum_{i,j=1}^m \langle A_{i\tau} w_{i\tau}, A_{j\tau} w_{j\tau}\rangle + \frac{1}{m}\sum_{\tau=1}^n\sum_{i=1}^m \|A_{i\tau} w_{i\tau}\|^2 \\
&= \frac{1}{m^2}\|AW\|^2 - \frac{1}{m^2}\sum_{\tau=1}^n\left\|\sum_{i=1}^m A_{i\tau} w_{i\tau}\right\|^2 + \frac{1}{m}\sum_{\tau=1}^n\sum_{i=1}^m \|A_{i\tau} w_{i\tau}\|^2 \\
&\leq \frac{1}{m^2}\|AW\|^2 + \frac{1}{m}\sum_{\tau=1}^n\sum_{i=1}^m \|A_{i\tau} w_{i\tau}\|^2 \\
&\leq \frac{NR^2}{m^2}\sum_{\tau=1}^n\sum_{i=1}^m \|w_{i\tau}\|^2 + \frac{1}{m}R_m^2\sum_{\tau=1}^n\sum_{i=1}^m \|w_{i\tau}\|^2 \\
&= \left(\frac{nR^2}{m} + \frac{1}{m}R_m^2\right)\sum_{\tau=1}^n\sum_{i=1}^m \|w_{i\tau}\|^2,
\end{aligned}$$

where in the second equality, we use the fact that i^{τ_1} is indpendent of i^{τ_2} for $\tau_1 \neq \tau_2$, and in the last inequality, we use $\|A_{i\tau}\| \leq \max_{i,\tau}\|A_{i\tau}\| = R_m$ and

$$\frac{1}{N}\lambda_{\max}(AA^\top) = \frac{1}{N}\lambda_{\max}(\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}A_{i\tau}^\top) = R^2.$$

D.2 Proof of Lemma C.4

First, we have

$$\begin{aligned}
&\mathbb{E}_k\|h_\tau^{k+1} - u_\tau^{k+1}\|^2 \\
&= \mathbb{E}_k\left\|h_\tau^{k+1} - u_\tau^k - \mathbb{E}_k[u_\tau^{k+1} - u_\tau^k] + \mathbb{E}_k[u_\tau^{k+1} - u_\tau^k] - (u_\tau^{k+1} - u_\tau^k)\right\|^2 \\
&= \mathbb{E}_k\left\|h_\tau^{k+1} - u_\tau^k - \mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 + \mathbb{E}_k\left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k] - (u_\tau^{k+1} - u_\tau^k)\right\|^2 \\
&= \mathbb{E}_k\left\|h_\tau^{k+1} - u_\tau^k - \mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 + \mathbb{E}_k\|u_\tau^{k+1} - u_\tau^k\|^2 - \left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 \\
&\leq (1+\beta)\mathbb{E}_k\|h_\tau^{k+1} - u_\tau^k\|^2 + \left(1 + \frac{1}{\beta} - 1\right)\left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 + \mathbb{E}_k\|u_\tau^{k+1} - u_\tau^k\|^2 \qquad (10) \\
&\leq (1-\delta_1)(1+\beta)\|h_\tau^k - u_\tau^k\|^2 + \frac{1}{\beta}\left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 + \mathbb{E}_k\|u_\tau^{k+1} - u_\tau^k\|^2 \\
&= \left(1 - \frac{\delta_1}{2}\right)\|h_\tau^k - u_\tau^k\|^2 + \frac{2(1-\delta_1)}{\delta_1}\left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 + \mathbb{E}_k\|u_\tau^{k+1} - u_\tau^k\|^2,
\end{aligned}$$

where we use Young's inequality for any $\beta > 0$ in the first inequality, in the second inequality we use the contraction property of Q_1, in the last equality we choose $\beta = \frac{\delta_1}{2(1-\delta_1)}$ when $\delta_1 < 1$. When $\delta_1 = 1$, it is easy to see that the above inequality also holds.

Since $u_\tau^{k+1} - u_\tau^k = \frac{1}{m} A_{i_k^\tau \tau}(y_{i_k^\tau \tau}^{k+1} - y_{i_k^\tau \tau}^k)$, we have

$$\begin{aligned}\mathbb{E}_k \|u_\tau^{k+1} - u_\tau^k\|^2 &\leq \frac{R_m^2}{m^2} \mathbb{E}_k \left\| y_{i_k^\tau \tau}^{k+1} - y_{i_k^\tau \tau}^k \right\|^2 \\ &= \frac{R_m^2}{m^3} \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.\end{aligned}$$

From $\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k] = \frac{1}{m^2} \sum_{i=1}^m A_{i\tau}(\tilde{y}_{i\tau}^k - y_{i\tau}^k)$, we can get

$$\begin{aligned}\left\|\mathbb{E}_k[u_\tau^{k+1} - u_\tau^k]\right\|^2 &= \frac{1}{m^4} \left\| \sum_{i=1}^m A_{i\tau}(\tilde{y}_{i\tau}^k - y_{i\tau}^k) \right\|^2 \\ &\leq \frac{1}{m^4} \left\|[A_{1\tau}, ..., A_{m\tau}]\right\|^2 \cdot \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2 \\ &= \frac{1}{m^4} \lambda_{\max} \left(\sum_{i=1}^m A_{i\tau} A_{i\tau}^\top \right) \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2 \\ &\leq \frac{\bar{R}^2}{m^3} \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.\end{aligned}$$

Combining the above three inequalities, we arrive at

$$\mathbb{E}_k \|h_\tau^{k+1} - u_\tau^{k+1}\|^2 \leq \left(1 - \frac{\delta_1}{2}\right) \|h_\tau^k - u_\tau^k\|^2 + \frac{1}{m^3} \left(\frac{2(1-\delta_1)\bar{R}^2}{\delta_1} + R_m^2 \right) \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.$$

Summing up the above inequality from $\tau = 1$ to n and dividing both sides of the resulting inequality by n , we can get the result.

D.3 Proof of Lemma C.5

First, same as (10), we can obtain

$$\mathbb{E}_k \|h^{k+1} - u^{k+1}\|^2 \leq (1+\beta) \|h^{k+1} - u^k\|^2 + \frac{1}{\beta} \left\|\mathbb{E}_k[u^{k+1} - u^k]\right\|^2 + \mathbb{E}_k \|u^{k+1} - u^k\|^2,$$

for any $\beta > 0$.

Under Assumption 2.3, same as the analysis of $\mathbb{E}_k \|h^{k+1} - \nabla f(w^k)\|^2$ in Lemma B.5 of [Qian et al., 2021c], we can get

$$\mathbb{E}_k \|h^{k+1} - u^k\|^2 \leq (1-\delta_1)^2 \|h^k - u^k\|^2 + \frac{(1-\delta_1)\delta_1}{n^2} \sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2.$$

Combining the above two inequalities yields that

$$\begin{aligned}\mathbb{E}_k \|h^{k+1} - u^{k+1}\|^2 &\leq (1+\beta)(1-\delta_1)^2 \|h^k - u^k\|^2 + \frac{(1+\beta)(1-\delta_1)\delta_1}{n^2} \sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2 \\ &\quad + \frac{1}{\beta} \left\|\mathbb{E}_k[u^{k+1} - u^k]\right\|^2 + \mathbb{E}_k \|u^{k+1} - u^k\|^2 \\ &= (1-\delta_1) \|h^k - u^k\|^2 + \frac{\delta_1}{n^2} \sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2 \\ &\quad + \frac{(1-\delta_1)}{\delta_1} \left\|\mathbb{E}_k[u^{k+1} - u^k]\right\|^2 + \mathbb{E}_k \|u^{k+1} - u^k\|^2,\end{aligned}$$

where we choose $\beta = \frac{\delta_1}{1-\delta_1}$ when $\delta_1 < 1$. When $\delta_1 = 1$, $h^{k+1} = u^k$, thus the above inequality also holds.

Since $u_\tau^{k+1} - u_\tau^k = \frac{1}{m} A_{i_k^\tau \tau}(y_{i_k^\tau \tau}^{k+1} - y_{i_k^\tau \tau}^k)$, we have

$$\begin{aligned}
\mathbb{E}_k \|u^{k+1} - u^k\|^2 &= \mathbb{E}_k \left\| \frac{1}{mn} \sum_{\tau=1}^n A_{i_k^\tau \tau}(y_{i_k^\tau \tau}^{k+1} - y_{i_k^\tau \tau}^k) \right\|^2 \\
&= \mathbb{E}_k \left\| \frac{1}{mn} \sum_{\tau=1}^n A_{i_k^\tau \tau}(\tilde{y}_{i_k^\tau \tau}^{k} - y_{i_k^\tau \tau}^k) \right\|^2 \\
&\overset{Lemma\ C.1}{\leq} \frac{nR^2 + R_m^2}{m^3 n^2} \sum_{\tau=1}^n \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.
\end{aligned}$$

From $\mathbb{E}_k[u^{k+1} - u^k] = \frac{1}{m^2 n} \sum_{\tau=1}^n \sum_{i=1}^m A_{i\tau}(\tilde{y}_{i\tau}^k - y_{i\tau}^k)$, we can obtain

$$\begin{aligned}
\left\| \mathbb{E}_k[u^{k+1} - u^k] \right\|^2 &\leq \frac{1}{m^4 n^2} \|A\|^2 \sum_{\tau=1}^n \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2 \\
&= \frac{R^2}{m^3 n} \sum_{\tau=1}^n \sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2,
\end{aligned}$$

where A is defined in (9). Combining the above three inequalities, we can get the result.

D.4 Proof of Theorem C.7

(i) Let $\tilde{y}_{i\tau}^k := \arg\max_{y \in \mathbb{R}^t} \left\{ \langle y, A_{i\tau}^\top z^k \rangle - \phi_{i\tau}^*(y) - \frac{1}{2\sigma}\|y - y_{i\tau}^k\|^2 \right\}$. Then similar to (47) in [Zhang and Xiao, 2017], we can get

$$\begin{aligned}
&\left(\frac{1}{2\sigma} + \frac{(m-1)\gamma}{2m}\right)\frac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 \\
&\geq \left(\frac{1}{2\sigma} + \frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \mathbb{E}_k\|y_{i\tau}^{k+1} - y_{i\tau}^*\|^2 + \frac{1}{2\sigma n}\sum_{\tau=1}^n\sum_{i=1}^m \mathbb{E}_k\|y_{i\tau}^{k+1} - y_{i\tau}^k\|^2 \\
&\quad + \frac{1}{n}\mathbb{E}_k\left[\sum_{\tau=1}^n (\phi_{i_k^\tau \tau}^*(y_{i_k^\tau \tau}^{k+1}) - \phi_{i_k^\tau \tau}^*(y_{i_k^\tau \tau}^k))\right] \\
&\quad + \frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m (\phi_{i\tau}^*(y_{i\tau}^k) - \phi_{i\tau}^*(y_{i\tau}^*)) - \mathbb{E}_k\langle u^k - u^* + m(u^{k+1} - u^k), z^k\rangle, \qquad (11)
\end{aligned}$$

where $u^k = \frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}y_{i\tau}^k$ and $u^* = \frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}y_{i\tau}^*$.

From the update rule of x^{k+1} and optimality condition, we have

$$\partial g(x^{k+1}) + h^k + \Delta^k + \frac{1}{\eta}(x^{k+1} - x^k) = 0.$$

For $h^k + \Delta^k$, we have

$$\begin{aligned}
h^k + \Delta^k &= \frac{1}{n}\sum_{\tau=1}^n (h_\tau^k + \Delta_\tau^k) \\
&= \frac{1}{n}\sum_{\tau=1}^n (h_\tau^k + e_\tau^k + A_{i_k^\tau \tau}(y_{i_k^\tau \tau}^{k+1} - y_{i_k^\tau \tau}^k) + u_\tau^k - h_\tau^k - e_\tau^{k+1}) \\
&= \frac{1}{n}\sum_{\tau=1}^n (e_\tau^k - e_\tau^{k+1} + m(u_\tau^{k+1} - u_\tau^k)) \\
&= e^k - e^{k+1} + u^k + m(u^{k+1} - u^k).
\end{aligned}$$

By using the above two equalities, we can obtain

$$\begin{aligned}
\tilde{x}^{k+1} &= x^{k+1} - \eta e^{k+1} \\
&= x^k - \eta(\partial g(x^{k+1}) + h^k + \Delta^k) - \eta e^{k+1} \\
&= x^k - \eta(\partial g(x^{k+1}) + e^k - e^{k+1} + u^k + m(u^{k+1} - u^k) - e^{k+1}) \\
&= \tilde{x}^k - \eta(\partial g(x^{k+1}) + u^k + m(u^{k+1} - u^k)).
\end{aligned}$$

Then similar to Lemma B.3 in [Qian et al., 2021c], we can get

$$\begin{aligned}
\langle u^k + m(u^{k+1} - u^k), x^* - x^{k+1}\rangle \geq &\left(\frac{1}{2\eta} + \frac{\lambda}{4}\right)\|\tilde{x}^{k+1} - x^*\|^2 - \frac{1}{2\eta}\|\tilde{x}^k - x^*\|^2 - \frac{\eta}{2}\|e^k\|^2 \\
&- \left(\frac{\eta}{2} + \frac{\lambda\eta^2}{2}\right)\|e^{k+1}\|^2 + \frac{1}{4\eta}\|x^{k+1} - x^k\|^2 + g(x^{k+1}) - g(x^*).
\end{aligned}$$

Rearranging terms and taking conditional expectation, we arrive at

$$\begin{aligned}
\frac{1}{2\eta}\|\tilde{x}^k - x^*\|^2 \geq &\left(\frac{1}{2\eta} + \frac{\lambda}{4}\right)\mathbb{E}_k\|\tilde{x}^{k+1} - x^*\|^2 + \frac{1}{4\eta}\mathbb{E}_k\|x^{k+1} - x^k\|^2 + \mathbb{E}_k[g(x^{k+1}) - g(x^*)] \\
&+ \mathbb{E}_k\langle u^k + m(u^{k+1} - u^k), x^{k+1} - x^*\rangle - \frac{\eta}{2}\|e^k\|^2 - \left(\frac{\eta}{2} + \frac{\lambda\eta^2}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2. \qquad (12)
\end{aligned}$$

Similar to (51) in [Zhang and Xiao, 2017], we can get

$$
\begin{aligned}
& f(x^{k+1}, Y^*) - f(x^*, Y^*) + m\left(f(x^*, Y^*) - f(x^*, Y^{k+1})\right) - (m-1)\left(f(x^*, Y^*) - f(x^*, Y^k)\right) \\
= {} & \frac{1}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\left(\phi^*_{i\tau}(y^k_{i\tau}) - \phi^*_{i\tau}(y^*_{i\tau})\right) + \frac{1}{n}\sum_{\tau=1}^{n}\left(\phi^*_{i^\tau_k\tau}(y^{k+1}_{i^\tau_k\tau}) - \phi^*_{i^\tau_k\tau}(y^k_{i^\tau_k\tau})\right) + g(x^{k+1}) - g(x^*) \\
& + \langle u^*, x^{k+1}\rangle - \langle u^k, x^*\rangle + m\langle u^k - u^{k+1}, x^*\rangle.
\end{aligned}
$$

Combining (11), (12), and the above equality after taking conditional expectation, we can obtain

$$
\begin{aligned}
& \frac{1}{2\eta}\|\tilde{x}^k - x^*\|^2 + \left(\frac{1}{2\sigma} + \frac{(m-1)\gamma}{2m}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y^k_{i\tau} - y^*_{i\tau}\|^2 + (m-1)\left(f(x^*, Y^*) - f(x^*, Y^k)\right) \\
\geq {} & \left(\frac{1}{2\eta} + \frac{\lambda}{4}\right)\mathbb{E}_k\|\tilde{x}^{k+1} - x^*\|^2 + \left(\frac{1}{2\sigma} + \frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\mathbb{E}_k\|y^{k+1}_{i\tau} - y^*_{i\tau}\|^2 + \frac{1}{4\eta}\mathbb{E}_k\|x^{k+1} - x^k\|^2 \\
& + \frac{1}{2\sigma mn}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}^k_{i\tau} - y^k_{i\tau}\|^2 + \mathbb{E}_k\left[f(x^{k+1}, Y^*) - f(x^*, Y^*) + m\left(f(x^*, Y^*) - f(x^*, Y^{k+1})\right)\right] \\
& + \mathbb{E}_k\langle u^k - u^* + m(u^{k+1} - u^k), x^{k+1} - z^k\rangle - \frac{\eta}{2}\|e^k\|^2 - \left(\frac{\eta}{2} + \frac{\lambda\eta^2}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2, \qquad (13)
\end{aligned}
$$

where we also use the fact that $\sum_{\tau=1}^{n}\sum_{i=1}^{m}\mathbb{E}_k\|y^{k+1}_{i\tau} - y^k_{i\tau}\|^2 = \frac{1}{m}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}^k_{i\tau} - y^k_{i\tau}\|^2$.

Next we estimate the term $\mathbb{E}_k\langle u^k - u^* + m(u^{k+1} - u^k), x^{k+1} - z^k\rangle$. We have

$$
\begin{aligned}
& \langle u^k - u^* + m(u^{k+1} - u^k), x^{k+1} - z^k\rangle \\
= {} & \frac{1}{N}\langle AY^k - AY^* + mA(Y^{k+1} - Y^k), x^{k+1} - x^k - \theta(x^k - x^{k-1})\rangle \\
= {} & \frac{1}{N}\langle AY^{k+1} - AY^*, x^{k+1} - x^k\rangle - \frac{\theta}{N}\langle AY^k - AY^*, x^k - x^{k-1}\rangle \\
& + \frac{m-1}{N}\langle AY^{k+1} - AY^k, x^{k+1} - x^k\rangle + \frac{m\theta}{N}\langle AY^{k+1} - AY^k, x^k - x^{k-1}\rangle, \qquad (14)
\end{aligned}
$$

where we define $x^{-1} := x^0$ to guarantee $z^0 = x^0$. By Cauchy-Schwarz inequality, we have

$$
|\langle AY^{k+1} - AY^k, x^{k+1} - x^k\rangle| \leq \frac{n}{8\eta}\|x^{k+1} - x^k\|^2 + \frac{2\eta}{n}\|A(Y^{k+1} - Y^k)\|^2.
$$

By Lemma C.1, we further have

$$
\begin{aligned}
\mathbb{E}_k\|A(Y^{k+1} - Y^k)\|^2 &= \mathbb{E}_k\left\|\sum_{\tau=1}^{n}A_{i^\tau_k\tau}(y^{k+1}_{i^\tau_k\tau} - y^k_{i^\tau_k\tau})\right\|^2 \\
&= \mathbb{E}_k\left\|\sum_{\tau=1}^{n}A_{i^\tau_k\tau}(\tilde{y}^k_{i^\tau_k\tau} - y^k_{i^\tau_k\tau})\right\|^2 \\
&\overset{Lemma\ C.1}{\leq} \frac{nR^2 + R_m^2}{m}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}^k_{i\tau} - y^k_{i\tau}\|^2.
\end{aligned}
$$

Thus, we arrive at

$$
\mathbb{E}_k\langle AY^{k+1} - AY^k, x^{k+1} - x^k\rangle \geq -\frac{n}{8\eta}\mathbb{E}_k\|x^{k+1} - x^k\|^2 - \frac{2\eta(nR^2 + R_m^2)}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}^k_{i\tau} - y^k_{i\tau}\|^2.
$$

Simiarly, we can get

$$
\mathbb{E}_k\langle AY^{k+1} - AY^k, x^k - x^{k-1}\rangle \geq -\frac{n}{8\eta}\|x^k - x^{k-1}\|^2 - \frac{2\eta(nR^2 + R_m^2)}{N}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}^k_{i\tau} - y^k_{i\tau}\|^2.
$$

Combining the above two inequalities with (13) and (14), we have

$$\begin{aligned}
&\frac{1}{2\eta}\|\tilde{x}^k - x^*\|^2 + \left(\frac{1}{2\sigma} + \frac{(m-1)\gamma}{2m}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^k - y_{i\tau}^*\|^2 + (m-1)\left(f(x^*,Y^*) - f(x^*,Y^k)\right) \\
&\quad + \theta\left(f(x^k,Y^*) - f(x^*,Y^*)\right) + \frac{\theta}{8\eta}\|x^k - x^{k-1}\|^2 + \frac{\theta}{N}\langle AY^k - AY^*, x^k - x^{k-1}\rangle \\
&\geq \left(\frac{1}{2\eta} + \frac{\lambda}{4}\right)\mathbb{E}_k\|\tilde{x}^{k+1} - x^*\|^2 + \left(\frac{1}{2\sigma} + \frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\mathbb{E}_k\|y_{i\tau}^{k+1} - y_{i\tau}^*\|^2 + \frac{1}{8\eta}\mathbb{E}_k\|x^{k+1} - x^k\|^2 \\
&\quad + \mathbb{E}_k\left[f(x^{k+1},Y^*) - f(x^*,Y^*) + m\left(f(x^*,Y^*) - f(x^*,Y^{k+1})\right)\right] + \frac{1}{N}\mathbb{E}_k\langle AY^{k+1} - AY^*, x^{k+1} - x^k\rangle \\
&\quad + \frac{1}{2N}\left(\frac{1}{\sigma} - \frac{8\eta(nR^2 + R_m^2)}{n}\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2 - \frac{\eta}{2}\|e^k\|^2 - \left(\frac{\eta}{2} + \frac{\lambda\eta^2}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2,
\end{aligned} \tag{15}$$

where we add the nonnegative term $\theta\left(f(x^k,Y^*) - f(x^*,Y^*)\right)$ to the left-hand side of the above inequality.

From Lemma C.2, we have

$$\begin{aligned}
&\frac{3(\eta+\lambda\eta^2)}{\delta n}\sum_{\tau=1}^{n}\mathbb{E}_k\|e_\tau^{k+1}\|^2 + \frac{\eta}{2n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 + \frac{\eta+\lambda\eta^2}{2n}\sum_{\tau=1}^{n}\mathbb{E}_k\|e_\tau^{k+1}\|^2 \\
&\leq \frac{\eta+\lambda\eta^2}{n}\left(\left(\frac{3}{\delta} + \frac{1}{2}\right)\left(1 - \frac{\delta}{2}\right) + \frac{1}{2}\right)\sum_{\tau=1}^{n}\|e_\tau^k\|^2 + \frac{4(1-\delta)(\eta+\lambda\eta^2)}{\delta n}\left(\frac{3}{\delta} + \frac{1}{2}\right)\sum_{\tau=1}^{n}\|h_\tau^k - u_\tau^k\|^2 \\
&\quad + \frac{(1-\delta)(\eta+\lambda\eta^2)}{mn}\left(\frac{3}{\delta} + \frac{1}{2}\right)\left(\frac{4\bar{R}^2}{\delta} + R_m^2\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2 \\
&\leq \frac{3(\eta+\lambda\eta^2)}{\delta n}\left(1 - \frac{\delta}{6}\right)\sum_{\tau=1}^{n}\|e_\tau^k\|^2 + \frac{14(1-\delta)(\eta+\lambda\eta^2)}{\delta^2 n}\sum_{\tau=1}^{n}\|h_\tau^k - u_\tau^k\|^2 \\
&\quad + \frac{(1-\delta)(\eta+\lambda\eta^2)}{mn}\left(\frac{14\bar{R}^2}{\delta^2} + \frac{7R_m^2}{2\delta}\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.
\end{aligned} \tag{16}$$

From Lemma C.4, we have

$$\begin{aligned}
&\frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\sum_{\tau=1}^{n}\mathbb{E}_k\|h_\tau^{k+1} - u_\tau^{k+1}\|^2 + \frac{14(1-\delta)(\eta+\lambda\eta^2)}{\delta^2 n}\sum_{\tau=1}^{n}\|h_\tau^k - u_\tau^k\|^2 \\
&\leq \frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\left(1 - \frac{\delta_1}{2} + \frac{\delta_1}{3}\right)\sum_{\tau=1}^{n}\|h_\tau^k - u_\tau^k\|^2 \\
&\quad + \frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 m^3 n}\left(\frac{2(1-\delta_1)\bar{R}^2}{\delta_1} + R_m^2\right)\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.
\end{aligned} \tag{17}$$

From (15), (16), (17), and the fact that $\|e^k\|^2 \le \frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2$, we arrive at

$$\begin{aligned}
&\left(\frac{1}{2\eta}+\frac{\lambda}{4}\right)\mathbb{E}_k\|\tilde{x}^{k+1}-x^*\|^2 + \left(\frac{1}{2\sigma}+\frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \mathbb{E}_k\|y_{i\tau}^{k+1}-y_{i\tau}^*\|^2 + \frac{1}{8\eta}\mathbb{E}_k\|x^{k+1}-x^k\|^2 \\
&\quad + \mathbb{E}_k\left[f(x^{k+1},Y^*)-f(x^*,Y^*)+m\left(f(x^*,Y^*)-f(x^*,Y^{k+1})\right)\right] + \frac{1}{N}\mathbb{E}_k\langle AY^{k+1}-AY^*, x^{k+1}-x^k\rangle \\
&\quad + \frac{3(\eta+\lambda\eta^2)}{\delta n}\sum_{\tau=1}^n \mathbb{E}_k\|e_\tau^{k+1}\|^2 + \frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\sum_{\tau=1}^n \mathbb{E}_k\|h_\tau^{k+1}-u_\tau^{k+1}\|^2 \\
\le{}& \frac{1}{2\eta}\|\tilde{x}^k-x^*\|^2 + \left(\frac{1}{2\sigma}+\frac{(m-1)\gamma}{2m}\right)\frac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k-y_{i\tau}^*\|^2 + (m-1)\left(f(x^*,Y^*)-f(x^*,Y^k)\right) \\
&+ \theta\left(f(x^k,Y^*)-f(x^*,Y^*)\right) + \frac{\theta}{8\eta}\|x^k-x^{k-1}\|^2 + \frac{\theta}{N}\langle AY^k-AY^*, x^k-x^{k-1}\rangle \\
&+ \frac{3(\eta+\lambda\eta^2)}{\delta n}\left(1-\frac{\delta}{6}\right)\sum_{\tau=1}^n\|e_\tau^k\|^2 + \frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\left(1-\frac{\delta_1}{6}\right)\sum_{\tau=1}^n\|h_\tau^k-u_\tau^k\|^2 \\
&- \frac{1}{mn}\left(\frac{1}{2\sigma}-\frac{4\eta(nR^2+R_m^2)}{n}-(1-\delta)(\eta+\lambda\eta^2)\left(\frac{14\bar{R}^2}{\delta^2}+\frac{7R_m^2}{2\delta}+\frac{84(1-\delta_1)\bar{R}^2}{\delta^2\delta_1^2 m^2}+\frac{42R_m^2}{\delta^2\delta_1 m^2}\right)\right) \\
&\cdot\sum_{\tau=1}^n\sum_{i=1}^m\|\tilde{y}_{i\tau}^k-y_{i\tau}^k\|^2. \qquad (18)
\end{aligned}$$

Define

$$\begin{aligned}
\Phi_1^k :={}& \left(\frac{1}{2\eta}+\frac{\lambda}{4}\right)\|\tilde{x}^k-x^*\|^2 + \left(\frac{1}{2\sigma}+\frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^n\sum_{i=1}^m\|y_{i\tau}^k-y_{i\tau}^*\|^2 + \frac{1}{8\eta}\|x^k-x^{k-1}\|^2 \\
&+ f(x^k,Y^*)-f(x^*,Y^*)+m\left(f(x^*,Y^*)-f(x^*,Y^k)\right) + \frac{1}{N}\langle AY^k-AY^*, x^k-x^{k-1}\rangle \\
&+ \frac{3(\eta+\lambda\eta^2)}{\delta n}\sum_{\tau=1}^n\|e_\tau^k\|^2 + \frac{42(1-\delta)(\eta+\lambda\eta^2)}{\delta^2\delta_1 n}\sum_{\tau=1}^n\|h_\tau^k-u_\tau^k\|^2,
\end{aligned}$$

for $k \ge 0$, where $x^{-1}=x^0$. Assume $\frac{\mathcal{R}_2^2}{\lambda\gamma} \ge 1$. Then $\lambda\eta = \frac{1}{2\mathcal{R}_2}\sqrt{\frac{\lambda\gamma}{m}} \le \frac{1}{2}$, and thus

$$\begin{aligned}
&\frac{4\eta(nR^2+R_m^2)}{n} + (1-\delta)(\eta+\lambda\eta^2)\left(\frac{14\bar{R}^2}{\delta^2}+\frac{7R_m^2}{2\delta}+\frac{84(1-\delta_1)\bar{R}^2}{\delta^2\delta_1^2 m^2}+\frac{42R_m^2}{\delta^2\delta_1 m^2}\right) \\
\le{}& \frac{4\eta(nR^2+R_m^2)}{n} + \frac{3(1-\delta)\eta}{2}\left(\frac{14\bar{R}^2}{\delta^2}+\frac{7R_m^2}{2\delta}+\frac{84(1-\delta_1)\bar{R}^2}{\delta^2\delta_1^2 m^2}+\frac{42R_m^2}{\delta^2\delta_1 m^2}\right) \\
={}& 2\eta\mathcal{R}_2^2 = \frac{1}{2\sigma}.
\end{aligned}$$

From (18), the above inequality, and the definition of Φ_1^k, we can get

$$\mathbb{E}_k[\Phi_1^{k+1}] \le \theta\Phi_1^k,$$

where we use $\left(1-\frac{\delta}{6}\right) \le \theta$, $\left(1-\frac{\delta_1}{6}\right) \le \theta$,

$$\frac{m-1}{m} = 1-\frac{1}{m} \le \theta, \quad \frac{1}{2\eta}\Big/\left(\frac{1}{2\eta}+\frac{\lambda}{4}\right) = 1-\frac{1}{1+4\mathcal{R}_2\sqrt{m/(\lambda\gamma)}} \le \theta,$$

and

$$\left(\frac{1}{2\sigma}+\frac{(m-1)\gamma}{2m}\right)\Big/\left(\frac{1}{2\sigma}+\frac{\gamma}{2}\right) = 1-\frac{1}{m+m/(\gamma\sigma)} = 1-\frac{1}{m+2\mathcal{R}_2\sqrt{m/(\lambda\gamma)}} \le \theta.$$

By the tower property, we further have $\mathbb{E}[\Phi_1^{k+1}] \le \theta\mathbb{E}[\Phi_1^k]$. Apply this relation recursively, we can obtain

$$\mathbb{E}[\Phi_1^k] \le \theta^k\Phi_1^0. \qquad (19)$$

From the definition of Φ_2^k, we know that

$$\Phi_1^k = \Phi_2^k + \frac{1}{4\sigma n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 + \frac{1}{8\eta}\|x^k - x^{k-1}\|^2 + \frac{1}{N}\langle AY^k - AY^*, x^k - x^{k-1}\rangle.$$

From Young's inequality, we have

$$\begin{aligned}
\frac{1}{N}\left|\langle AY^k - AY^*, x^k - x^{k-1}\rangle\right| &\leq \frac{\|x^k - x^{k-1}\|^2}{8\eta} + \frac{\|A\|^2\|Y^k - Y^*\|^2}{N^2/(2\eta)} \\
&= \frac{\|x^k - x^{k-1}\|^2}{8\eta} + \frac{2\eta R^2}{N}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 \\
&\leq \frac{\|x^k - x^{k-1}\|^2}{8\eta} + \frac{\eta \mathcal{R}_2^2}{n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2 \\
&= \frac{\|x^k - x^{k-1}\|^2}{8\eta} + \frac{1}{4\sigma n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^k - y_{i\tau}^*\|^2, \qquad (20)
\end{aligned}$$

which indicates that $\Phi_2^k \leq \Phi_1^k$ for $k \geq 0$. Therefore, from (19) we have

$$\mathbb{E}[\Phi_2^k] \leq \theta^k \Phi_1^0 = \theta^k \left(\Phi_2^0 + \frac{1}{4\sigma n}\sum_{\tau=1}^n\sum_{i=1}^m \|y_{i\tau}^0 - y_{i\tau}^*\|^2\right).$$

Finally, from $\frac{R_m^2}{m} \leq \bar{R}^2$, we can get the results.

(ii) Under Assumption 2.3, from Lemma C.2 and Lemma C.3, we have

$$\begin{aligned}
&\frac{3(\eta+\lambda\eta^2)}{\delta}\mathbb{E}_k\|e^{k+1}\|^2 + \frac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\sum_{\tau=1}^n \mathbb{E}_k\|e_\tau^{k+1}\|^2 + \frac{\eta}{2}\|e^k\|^2 + \frac{\eta+\lambda\eta^2}{2}\mathbb{E}_k\|e^{k+1}\|^2 \\
&\leq \frac{3(\eta+\lambda\eta^2)}{\delta}\left(1-\frac{\delta}{6}\right)\|e^k\|^2 + \frac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\left(1-\frac{\delta}{6}\right)\sum_{\tau=1}^n \|e_\tau^k\|^2 \\
&\quad + \frac{14(1-\delta)(\eta+\lambda\eta^2)}{\delta^2}\|h^k - u^k\|^2 + \frac{84(1-\delta)(\eta+\lambda\eta^2)}{\delta^2 n^2}\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2 \\
&\quad + \frac{7(1-\delta)(\eta+\lambda\eta^2)}{\delta m n}\left(\frac{2R^2}{\delta} + \frac{11R_m^2}{2n} + \frac{12(1-\delta)\bar{R}^2}{\delta n}\right)\sum_{\tau=1}^n\sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.
\end{aligned}$$

From Lemma C.4 and Lemma C.5, we have

$$\begin{aligned}
&\frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\mathbb{E}_k\|h^{k+1} - u^{k+1}\|^2 + \frac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\sum_{\tau=1}^n \mathbb{E}_k\|h_\tau^{k+1} - u_\tau^{k+1}\|^2 \\
&\quad + \frac{14(1-\delta)(\eta+\lambda\eta^2)}{\delta^2}\|h^k - u^k\|^2 + \frac{84(1-\delta)(\eta+\lambda\eta^2)}{\delta^2 n^2}\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2 \\
&\leq \frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\left(1-\frac{\delta_1}{6}\right)\|h^k - u^k\|^2 + \frac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\left(1-\frac{\delta_1}{6}\right)\sum_{\tau=1}^n \|h_\tau^k - u_\tau^k\|^2 \\
&\quad + \frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 m^3 n}\left(\frac{R^2}{\delta_1} + \frac{19R_m^2}{n} + \frac{36(1-\delta_1)\bar{R}^2}{\delta_1 n}\right)\sum_{\tau=1}^n\sum_{i=1}^m \|\tilde{y}_{i\tau}^k - y_{i\tau}^k\|^2.
\end{aligned}$$

Combining the above two inequalities and (15), we arrive at

$$\begin{aligned}
&\left(\frac{1}{2\eta}+\frac{\lambda}{4}\right)\mathbb{E}_k\|\tilde{x}^{k+1}-x^*\|^2+\left(\frac{1}{2\sigma}+\frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\mathbb{E}_k\|y_{i\tau}^{k+1}-y_{i\tau}^*\|^2+\frac{1}{8\eta}\mathbb{E}_k\|x^{k+1}-x^k\|^2\\
&\quad+\mathbb{E}_k\left[f(x^{k+1},Y^*)-f(x^*,Y^*)+m\left(f(x^*,Y^*)-f(x^*,Y^{k+1})\right)\right]\\
&\quad+\frac{3(\eta+\lambda\eta^2)}{\delta}\mathbb{E}_k\|e^{k+1}\|^2+\frac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\sum_{\tau=1}^{n}\mathbb{E}_k\|e_\tau^{k+1}\|^2+\frac{1}{N}\mathbb{E}_k\langle AY^{k+1}-AY^*,x^{k+1}-x^k\rangle\\
&\quad+\frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\mathbb{E}_k\|h^{k+1}-u^{k+1}\|^2+\frac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\sum_{\tau=1}^{n}\mathbb{E}_k\|h_\tau^{k+1}-u_\tau^{k+1}\|^2\\
&\le\frac{1}{2\eta}\|\tilde{x}^k-x^*\|^2+\left(\frac{1}{2\sigma}+\frac{(m-1)\gamma}{2m}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^k-y_{i\tau}^*\|^2+(m-1)\left(f(x^*,Y^*)-f(x^*,Y^k)\right)\\
&\quad+\theta\left(f(x^k,Y^*)-f(x^*,Y^*)\right)+\frac{\theta}{8\eta}\|x^k-x^{k-1}\|^2+\frac{\theta}{N}\langle AY^k-AY^*,x^k-x^{k-1}\rangle\\
&\quad+\frac{3(\eta+\lambda\eta^2)}{\delta}\left(1-\frac{\delta}{6}\right)\|e^k\|^2+\frac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\left(1-\frac{\delta}{6}\right)\sum_{\tau=1}^{n}\|e_\tau^k\|^2\\
&\quad+\frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\left(1-\frac{\delta_1}{6}\right)\|h^k-u^k\|^2+\frac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\left(1-\frac{\delta_1}{6}\right)\sum_{\tau=1}^{n}\|h_\tau^k-u_\tau^k\|^2\\
&\quad-\frac{1}{mn}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|\tilde{y}_{i\tau}^k-y_{i\tau}^k\|^2\cdot\left(\frac{1}{2\sigma}-\frac{4\eta(nR^2+R_m^2)}{n}-7(1-\delta)(\eta+\lambda\eta^2)\right.\\
&\quad\left.\cdot\left(\frac{2R^2}{\delta^2}+\frac{11R_m^2}{2\delta n}+\frac{12(1-\delta)\bar{R}^2}{\delta^2 n}+\frac{12R^2}{5\delta^2\delta_1^2m^2}+\frac{228R_m^2}{5\delta^2\delta_1 m^2 n}+\frac{432(1-\delta_1)\bar{R}^2}{5\delta^2\delta_1^2m^2n}\right)\right). \qquad (21)
\end{aligned}$$

Define

$$\begin{aligned}
\Psi_1^k:=&\left(\frac{1}{2\eta}+\frac{\lambda}{4}\right)\|\tilde{x}^k-x^*\|^2+\left(\frac{1}{2\sigma}+\frac{\gamma}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^k-y_{i\tau}^*\|^2+\frac{1}{8\eta}\|x^k-x^{k-1}\|^2\\
&+f(x^k,Y^*)-f(x^*,Y^*)+m\left(f(x^*,Y^*)-f(x^*,Y^k)\right)+\frac{1}{N}\mathbb{E}_k\langle AY^k-AY^*,x^k-x^{k-1}\rangle\\
&+\frac{3(\eta+\lambda\eta^2)}{\delta}\|e^k\|^2+\frac{21(1-\delta)(\eta+\lambda\eta^2)}{\delta n^2}\sum_{\tau=1}^{n}\|e_\tau^k\|^2\\
&+\frac{84(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1}\|h^k-u^k\|^2+\frac{1512(1-\delta)(\eta+\lambda\eta^2)}{5\delta^2\delta_1 n^2}\sum_{\tau=1}^{n}\|h_\tau^k-u_\tau^k\|^2,
\end{aligned}$$

for $k\ge 0$, where $x^{-1}=x^0$.

Assume $\frac{\mathcal{R}_3^2}{\lambda\gamma}\ge 1$. Then $\lambda\eta=\frac{1}{2\mathcal{R}_3}\sqrt{\frac{\lambda\gamma}{m}}\le\frac{1}{2}$, and thus

$$\begin{aligned}
&\frac{4\eta(nR^2+R_m^2)}{n}+7(1-\delta)(\eta+\lambda\eta^2)\\
&\cdot\left(\frac{2R^2}{\delta^2}+\frac{11R_m^2}{2\delta n}+\frac{12(1-\delta)\bar{R}^2}{\delta^2 n}+\frac{12R^2}{5\delta^2\delta_1^2m^2}+\frac{228R_m^2}{5\delta^2\delta_1 m^2 n}+\frac{432(1-\delta_1)\bar{R}^2}{5\delta^2\delta_1^2m^2n}\right)\\
&\le 2\eta\mathcal{R}_3^2=\frac{1}{2\sigma}.
\end{aligned}$$

From (21), the above inequality, and the definition of Ψ_1^k, we can get

$$\mathbb{E}_k[\Psi_1^{k+1}]\le\theta\Psi_1^k.$$

By the tower property, we further have $\mathbb{E}[\Psi_1^{k+1}]\le\theta\mathbb{E}[\Psi_1^k]$. Apply this relation recursively, we can obtain

$$\mathbb{E}[\Psi_1^k]\le\theta^k\Psi_1^0. \qquad (22)$$

From the definition of Ψ_2^k, we know

$$\Psi_1^k = \Psi_2^k + \frac{1}{4\sigma n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^k - y_{i\tau}^*\|^2 + \frac{1}{8\eta}\|x^k - x^{k-1}\|^2 + \frac{1}{N}\langle AY^k - AY^*, x^k - x^{k-1}\rangle.$$

From (20), we have $\Psi_2^k \leq \Psi_1^k$ for $k \geq 0$. Thus, from (22) we can obtain

$$\mathbb{E}[\Psi_2^k] \leq \theta^k \Psi_1^0 = \theta^k\left(\Psi_2^0 + \frac{1}{4\sigma n}\sum_{\tau=1}^{n}\sum_{i=1}^{m}\|y_{i\tau}^0 - y_{i\tau}^*\|^2\right).$$

At the end, from $\frac{R_m^2}{m} \leq \bar{R}^2$ and $\frac{\bar{R}^2}{n} \leq R^2$, we can get the results.

E PROOFS FOR EC-LSVRG + CATALYST

E.1 Proof of Lemma 3.2

First, from Theorem 2.10 in [Qian et al., 2021a] and the initialization rules of $h^0_{\tau,(k)}$ and $e^0_{\tau,(k)}$, we have

$$\mathbb{E}[G_k(\bar{x}^K_{(k)}) - G_k^*] \leq \frac{9(\lambda+\kappa)\|x^0_{(k)} - x^*_{(k)}\|^2 + 2(G_k(x^0_{(k)}) - G^*_k)}{1-(1-\tilde{\theta})^{K+1}}(1-\tilde{\theta})^K,$$

where we denote $\tilde{\theta} := \min\{\frac{(\lambda+\kappa)\eta}{2}, \frac{\delta}{4}, \frac{\delta_1}{4}, \frac{p}{4}\}$. Since G_k is $(\lambda+\kappa)$-strongly convex, we have

$$G_k(x) - G^*_k \geq \frac{\lambda+\kappa}{2}\|x - x^*_{(k)}\|^2,$$

for any $x \in \mathbb{R}^d$, which indicates that

$$\mathbb{E}[G_k(\bar{x}^K_{(k)}) - G_k^*] \leq \frac{18(G_k(x^0_{(k)}) - G^*_k)}{1-(1-\tilde{\theta})^{K+1}}(1-\tilde{\theta})^K.$$

Noticing that $\ln(1-a) + a \leq 0$ for any $a \in (0,1)$, we have $(1-\tilde{\theta})^{\frac{1}{\tilde{\theta}}} \leq \frac{1}{e} < 0.37$. Now we first let $K \geq \frac{1}{\tilde{\theta}}$, then we have $(1-\tilde{\theta})^{K+1} \leq 0.37$, which yields

$$\mathbb{E}[G_k(\bar{x}^K_{(k)}) - G_k^*] \leq 30(G_k(x^0_{(k)}) - G^*_k)(1-\tilde{\theta})^K.$$

Then similar to the proof of Lemma C.1 in [Lin et al., 2015], but choosing T_0 as

$$T_0 = \max\left\{\frac{1}{\tilde{\theta}}, \frac{1}{\tilde{\theta}}\log\left(\frac{1}{1-e^{-\tilde{\theta}}}\frac{30(G_k(x^0_{(k)}) - G^*_k)}{\epsilon_k}\right)\right\}$$

instead, we can obtain

$$\mathbb{E}[T_k] \leq \max\left\{\frac{1}{\tilde{\theta}}, \frac{1}{\tilde{\theta}}\log\left(\frac{60(G_k(x^0_{(k)}) - G^*_k)}{\tilde{\theta}\epsilon_k}\right)\right\} + 1.$$

Within the above inequality, similar to the proof of Proposition 3.2 in [Lin et al., 2015], we can get $\mathbb{E}[T_k] \leq \tilde{\mathcal{O}}(1/\tilde{\theta})$, where the notation $\tilde{\mathcal{O}}$ hides some constants and some logorithmic dependencies in λ, κ, and $\tilde{\theta}$. At last, by the stepsize rule in Theorem 2.10 in [Qian et al., 2021a], we can obtain the result.

E.2 Proof of Lemma 3.5

From the inequality above (27) in the proof of Theorem 2.10 in [Qian et al., 2021a], we have

$$\mathbb{E}[G_k(x^K_{(k)}) - G^*_k] \leq \frac{18}{\eta}\mathbb{E}[\Phi^K_{3,(k)}] \leq \frac{18}{\eta}(1-\tilde{\theta}_1)^K\Phi^0_{3,(k)},$$

where we denote $\tilde{\theta}_1 := \min\left\{\frac{(\lambda+\kappa)\eta}{2}, \frac{\delta}{4}, \frac{\delta_1}{4}, \frac{p}{4}\right\}$. Thus, we can get

$$\mathbb{E}[\Phi^K_{3,(k)} + G_k(x^K_{(k)}) - G^*_k] \leq \left(1+\frac{18}{\eta}\right)(1-\tilde{\theta}_1)^K\Phi^0_{3,(k)}.$$

Then from Lemma C.1 in [Lin et al., 2015], we know

$$\mathbb{E}[T_k] \leq \tilde{\mathcal{O}}\left(\frac{1}{\tilde{\theta}_1}\log\left(\frac{(1+18/\eta)\Phi^0_{3,(k)}}{\epsilon_k}\right)\right) = \tilde{\mathcal{O}}\left(\frac{1}{\tilde{\theta}_1}\log\left(\frac{\Phi^0_{3,(k)}}{\epsilon_k}\right)\right).$$

Next we will show that $\log\left(\frac{\Phi^0_{3,(k)}}{\epsilon_k}\right) \leq \tilde{\mathcal{O}}(1)$, which concludes the proof. From the definition of $\Phi^K_{3,(k)}$ and the initialization rule at each outer iteration, we have

$$\begin{aligned}\Phi^0_{3,(k)} &= \|x^{k-1} - e^0_{(k)} - x^*_{(k)}\|^2 + \tfrac{12(L_f+\kappa)\eta}{n\delta}\sum_{\tau=1}^n \|e^0_{\tau,(k))}\|^2 + \tilde{\eta}(G_k(x^{k-1}) - G^*_k) \\ &\quad + \tfrac{192(1-\delta)(L_f+\kappa)\eta^3}{\delta^2\delta_1 n}\sum_{\tau=1}^n \|h^0_{\tau,(k)} - \nabla f^\tau(x^{k-1}) - \nabla\psi(x^{k-1}) - \kappa(x^{k-1} - y^{k-1})\|^2,\end{aligned}$$

where we denote $\tilde{\eta} = \frac{4}{3p}\left(\frac{48(1-\delta)(L_f+\kappa)\eta^3}{\delta}\left(\frac{4(\bar{L}+\kappa)}{\delta} + L + \kappa + \frac{16(\bar{L}+\kappa)p}{\delta\delta_1}\left(1 + \frac{2p}{\delta_1}\right)\right) + \frac{4(L+\kappa)\eta^2}{n}\right)$. We estimate each term in the above equality respectively. Since G_k is $(\lambda+\kappa)$-strongly convex, we have

$$\|x^{k-1} - e^0_{(k)} - x^*_{(k)}\|^2 \leq 2\|x^{k-1} - x^*_{(k)}\|^2 + 2\|e^0_{(k)}\|^2 \leq \frac{4}{\lambda+\kappa}(G_k(x^{k-1}) - G^*_k) + 2\|e^0_{(k)}\|^2.$$

For the third term, define $G^{(\tau)}_k(x) := f^{(\tau)}(x) + \psi(x) + \frac{\kappa}{2}\|x - y^{k-1}\|^2$ for simplicity. If $h^0_{\tau,(k)} = h^{T_{k-1}}_{\tau,(k-1)}$, then we have

$$\begin{aligned}&\frac{1-\delta}{n}\sum_{\tau=1}^n \|h^0_{\tau,(k)} - \nabla f^\tau(x^{k-1}) - \nabla\psi(x^{k-1}) - \kappa(x^{k-1} - y^{k-1})\|^2 \\ &= \frac{1-\delta}{n}\sum_{\tau=1}^n \|h^{T_{k-1}}_{\tau,(k-1)} - \nabla f^\tau(x^{k-1}) - \nabla\psi(x^{k-1}) - \kappa(x^{k-1} - y^{k-1})\|^2 \\ &= \quad \frac{1-\delta}{n}\sum_{\tau=1}^n \|h^{T_{k-1}}_{\tau,(k-1)} - \nabla G^\tau_k(x^{k-1})\|^2 \\ &\leq \frac{3(1-\delta)}{n}\sum_{\tau=1}^n \Big(\|h^{T_{k-1}}_{\tau,(k-1)} - \nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)})\|^2 \\ &\quad + \|\nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)}) - \nabla G^{(\tau)}_{k-1}(x^{k-1})\|^2 + \|\nabla G^{(\tau)}_{k-1}(x^{k-1}) - \nabla G^{(\tau)}_k(x^{k-1})\|^2\Big).\end{aligned}$$

Since $\Phi^{T_{k-1}}_{3,(k-1)} + G_{k-1}(x^{T_{k-1}}_{(k-1)}) - G^*_{k-1} \leq \epsilon_{k-1}$, we have

$$\frac{3(1-\delta)}{n}\sum_{\tau=1}^n \|h^{T_{k-1}}_{\tau,(k-1)} - \nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)})\|^2 \leq \frac{\delta^2\delta_1}{192(L_f+\kappa)\eta^3}\Phi^{T_{k-1}}_{3,(k-1)} \leq \frac{\delta^2\delta_1}{192(L_f+\kappa)\eta^3}\epsilon_{k-1}.$$

From the smoothness of $G^{(\tau)}_{k-1}$, we have

$$\begin{aligned}&\frac{1}{n}\sum_{\tau=1}^n \|\nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)}) - \nabla G^{(\tau)}_{k-1}(x^{k-1})\|^2 \\ &\leq \frac{2}{n}\sum_{\tau=1}^n \|\nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)}) - \nabla G^{(\tau)}_{k-1}(x^*_{(k-1)})\|^2 + \frac{2}{n}\sum_{\tau=1}^n \|\nabla G^{(\tau)}_{k-1}(x^{k-1}) - \nabla G^{(\tau)}_{k-1}(x^*_{(k-1)})\|^2 \\ &\leq 4(\bar{L}+\kappa)\left(G_{k-1}(w^{T_{k-1}}_{(k-1)}) - G^*_{k-1} + G_{k-1}(x^{k-1}) - G^*_{k-1}\right) \\ &\leq \frac{4(\bar{L}+\kappa)}{\tilde{\eta}}\Phi^{T_{k-1}}_{3,(k-1)} + 4(\bar{L}+\kappa)(G_{k-1}(x^{k-1}) - G^*_{k-1}) \\ &\leq 4(\bar{L}+\kappa)\left(\frac{1}{\tilde{\eta}} + 1\right)\epsilon_{k-1}.\end{aligned}$$

For $\frac{1}{n}\sum_{\tau=1}^n \|\nabla G^{(\tau)}_{k-1}(x^{k-1}) - \nabla G^{(\tau)}_k(x^{k-1})\|^2$, we have

$$\frac{1}{n}\sum_{\tau=1}^n \|\nabla G^{(\tau)}_{k-1}(x^{k-1}) - \nabla G^{(\tau)}_k(x^{k-1})\|^2 = \kappa^2\|y^{k-1} - y^{k-2}\|^2.$$

For the third term, if $h^0_{\tau,(k)} = h^{T_{k-1}}_{\tau,(k-1)} + \kappa(y^{k-2} - y^{k-1})$, then we have

$$\begin{aligned}
& \frac{1-\delta}{n}\sum_{\tau=1}^{n}\|h^0_{\tau,(k)} - \nabla f^{\tau}(x^{k-1}) - \nabla\psi(x^{k-1}) - \kappa(x^{k-1} - y^{k-1})\|^2 \\
= & \frac{1-\delta}{n}\sum_{\tau=1}^{n}\|h^{T_{k-1}}_{\tau,(k-1)} - \nabla f^{\tau}(x^{k-1}) - \nabla\psi(x^{k-1}) - \kappa(x^{k-1} - y^{k-2})\|^2 \\
= & \quad \frac{1-\delta}{n}\sum_{\tau=1}^{n}\|h^{T_{k-1}}_{\tau,(k-1)} - \nabla G^{\tau}_{k-1}(x^{k-1})\|^2 \\
\leq & \frac{3(1-\delta)}{n}\sum_{\tau=1}^{n}\left(\|h^{T_{k-1}}_{\tau,(k-1)} - \nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)})\|^2 + \|\nabla G^{(\tau)}_{k-1}(w^{T_{k-1}}_{(k-1)}) - \nabla G^{(\tau)}_{k-1}(x^{k-1})\|^2\right).
\end{aligned}$$

If $e^0_{\tau,(k)} = 0$, then $\|e^0_{(k)}\|^2 = \|e^0_{\tau,(k)}\|^2 = 0$. If $e^0_{\tau,(k)} = e^{T_{k-1}}_{\tau,(k-1)}$, then

$$\|e^0_{\tau,(k)}\|^2 \leq \frac{1}{n}\sum_{\tau=1}^{n}\|e^0_{\tau,(k)}\|^2 \leq \frac{\delta}{12(L_f+\kappa)\eta}\Phi^{T_{k-1}}_{3,(k-1)} \leq \frac{\delta}{12(L_f+\kappa)\eta}\epsilon_{k-1}.$$

Moreover, for any $a \geq 1$ and $b \geq 1$, we have $\log(a+b) \leq \log(2\max\{a,b\}) \leq \log(a) + \log(b) + 1$. Using the above estimations, we conclude that

$$\log\left(\frac{\Phi^0_{3,(k)}}{\epsilon_k}\right) \leq \tilde{\mathcal{O}}\left(\frac{G_k(x^{k-1}) - G_k^*}{\epsilon_k}\right) + \tilde{\mathcal{O}}\left(\frac{\|y^{k-1} - y^{k-2}\|^2}{\epsilon_k}\right) + \mathcal{O}(1)$$

Finally, from Lemmas B.1 and B.2 in [Lin et al., 2015], and similar to the proof of Proposition 3.2 in [Lin et al., 2015], we can get $\log\left(\frac{\Phi^0_{3,(k)}}{\epsilon_k}\right) \leq \tilde{\mathcal{O}}(1)$.

F PROOFS FOR EC-SDCA + CATALYST

F.1 Proof of Lemma 4.1

In the proof, we borrow the methodology of [Shalev-Shwartz and Zhang, 2014] for an accelerated proximal stochastic dual coordinate ascent method. First, from Theorem 3.5 in [Qian et al., 2021a], we have

$$\mathbb{E}[\Psi_{3,(k)}^K] \leq \left(1 - \min\left\{\theta, \frac{\delta}{4}\right\}\right)^K \epsilon_{D,(k)}^0,$$

for $K \geq 1$, and θ is chosen as in (9) in [Qian et al., 2021a]. Moreover, from the proof of Theorem 3.5 in [Qian et al., 2021a], it is easy to verify that for $K \geq 1$

$$\mathbb{E}[\epsilon_{P,(k)}^K] = \mathbb{E}[G_k(x_{(k)}^K) - G_k^*] \leq \frac{1}{\theta}\left(1 - \min\left\{\theta, \frac{\delta}{4}\right\}\right)^K \epsilon_{D,(k)}^0.$$

Combining the above two inequalities, we get

$$\mathbb{E}[\sqrt{4n+\delta mn}\Psi_{3,(k)}^K + 2(G_k(x_{(k)}^{K+1}) - G_k^*)] \leq \left(\sqrt{4n+\delta mn} + \frac{2}{\theta}\right)\left(1 - \min\left\{\theta, \frac{\delta}{4}\right\}\right)^K \epsilon_{D,(k)}^0, \tag{23}$$

for $K \geq 1$. Then from (23) and similar to Lemma C.1 in [Lin et al., 2015], we know $\mathbb{E}[T_1] \leq \tilde{\mathcal{O}}\left(\frac{1}{\delta} + m + \frac{\mathcal{M}_1}{\lambda+\kappa}\right)$. Next we consider $k > 1$ case.

Define $\tilde{\lambda} = \lambda + \kappa$, $\tilde{\xi}(x) = \frac{\lambda}{\tilde{\lambda}}\xi(x) + \frac{\kappa}{2\tilde{\lambda}}\|x\|^2$, and $\xi_k(x) = \frac{\lambda}{\tilde{\lambda}}\xi(x) + \frac{\kappa}{2\tilde{\lambda}}\|x - y^{k-1}\|^2$. Denote the objective function of the dual problem of minimizing G_k as $D_k(\alpha)$ and $v(\alpha) = \frac{1}{\tilde{\lambda}N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha_{i\tau}$. Then $G_k(x) = \frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m \phi_{i\tau}(A_{i\tau}^\top x) + \tilde{\lambda}\xi_k(x)$ and $D_k(\alpha) = -\frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m \phi_{i\tau}^*(-\alpha_{i\tau}) - \tilde{\lambda}\xi_k^*(v(\alpha))$. Noticing that

$$\xi_k^*(u) = \max_x \langle x, u + \frac{\kappa}{\tilde{\lambda}}y^{k-1}\rangle - \tilde{\xi}(x) - \frac{\kappa}{2\tilde{\lambda}}\|y^{k-1}\|^2 = \tilde{\xi}^*\left(u + \frac{\kappa}{\tilde{\lambda}}y^{k-1}\right) - \frac{\kappa}{2\tilde{\lambda}}\|y^{k-1}\|^2, \tag{24}$$

we can also write $D_k(\alpha) = -\frac{1}{N}\sum_{\tau=1}^n\sum_{i=1}^m \phi_{i\tau}^*(-\alpha_{i\tau}) - \tilde{\lambda}\tilde{\xi}^*\left(v(\alpha) + \frac{\kappa}{\tilde{\lambda}}y^{k-1}\right) + \frac{\kappa}{2}\|y^{k-1}\|^2$. Then we have

$$\begin{aligned}
& - D_k(\alpha^{k-1}) + D_{k-1}(\alpha^{k-1}) \\
&= \tilde{\lambda}\tilde{\xi}^*\left(v(\alpha^{k-1}) + \frac{\kappa}{\tilde{\lambda}}y^{k-1}\right) - \tilde{\lambda}\tilde{\xi}^*\left(v(\alpha^{k-1}) + \frac{\kappa}{\tilde{\lambda}}y^{k-2}\right) + \frac{\kappa}{2}\|y^{k-2}\|^2 - \frac{\kappa}{2}\|y^{k-1}\|^2 \\
&\leq \kappa\langle\nabla\tilde{\xi}^*\left(v(\alpha^{k-1}) + \frac{\kappa}{\tilde{\lambda}}y^{k-2}\right), y^{k-1} - y^{k-2}\rangle + \frac{\kappa^2}{2\tilde{\lambda}}\|y^{k-1} - y^{k-2}\|^2 + \frac{\kappa}{2}\|y^{k-2}\|^2 - \frac{\kappa}{2}\|y^{k-1}\|^2 \\
&= \kappa\langle\nabla\xi_{k-1}^*\left(v(\alpha^{k-1})\right), y^{k-1} - y^{k-2}\rangle + \frac{\kappa^2}{2\tilde{\lambda}}\|y^{k-1} - y^{k-2}\|^2 + \frac{\kappa}{2}\|y^{k-2}\|^2 - \frac{\kappa}{2}\|y^{k-1}\|^2,
\end{aligned} \tag{25}$$

where we use $\tilde{\xi}^*$ is 1-smooth in the first inequality and (24) in the last equality. From (33) in [Qian et al., 2021a], we know $\tilde{u}_{(k-1)}^{T_{k-1}} = v(\alpha_{(k-1)}^{T_{k-1}}) = v(\alpha^{k-1}) = u_{(k-1)}^{T_{k-1}} + e_{(k-1)}^{T_{k-1}}$. Moreover, from the update of EC-SDCA, we know the output at the k-th outer iteration $x^{k-1} = x_{(k-1)}^{T_{k-1}+1} = \nabla\xi_{k-1}^*(u_{(k-1)}^{T_{k-1}})$. Thus, for $\langle\nabla\xi_{k-1}^*\left(v(\alpha^{k-1})\right), y^{k-1} - y^{k-2}\rangle$, we have

$$\begin{aligned}
& \langle\nabla\xi_{k-1}^*\left(v(\alpha^{k-1})\right), y^{k-1} - y^{k-2}\rangle - \langle x^{k-1}, y^{k-1} - y^{k-2}\rangle \\
&= \langle\nabla\xi_{k-1}^*\left(\tilde{u}_{(k-1)}^{T_{k-1}}\right) - \nabla\tilde{\xi}_{k-1}^*(u_{(k-1)}^{T_{k-1}}), y^{k-1} - y^{k-2}\rangle \\
&\leq \frac{1}{2}\|\nabla\xi_{k-1}^*\left(\tilde{u}_{(k-1)}^{T_{k-1}}\right) - \nabla\xi_{k-1}^*(u_{(k-1)}^{T_{k-1}})\|^2 + \frac{1}{2}\|y^{k-1} - y^{k-2}\|^2 \\
&\leq \frac{1}{2}\|e_{(k-1)}^{T_{k-1}}\|^2 + \frac{1}{2}\|y^{k-1} - y^{k-2}\|^2,
\end{aligned}$$

where we use ξ_{k-1}^* is 1-smooth in the last inequality. From (25), the above inequality, and $\tilde{\lambda} > \kappa$, we arrive at

$$\begin{aligned}
& - D_k(\alpha^{k-1}) + D_{k-1}(\alpha^{k-1}) \\
&\leq \kappa\langle x^{k-1}, y^{k-1} - y^{k-2}\rangle + \frac{\kappa}{2}\|e_{(k-1)}^{T_{k-1}}\|^2 + \kappa\|y^{k-1} - y^{k-2}\|^2 + \frac{\kappa}{2}\|y^{k-2}\|^2 - \frac{\kappa}{2}\|y^{k-1}\|^2.
\end{aligned}$$

Furthermore, we have

$$G_k(x^{k-1}) = G_{k-1}(x^{k-1}) + \frac{\kappa}{2}\|y^{k-1}\|^2 - \frac{\kappa}{2}\|y^{k-2}\|^2 + \kappa\langle x^{k-1}, y^{k-2} - y^{k-1}\rangle.$$

Combining the above two inequalities, we obtain

$$G_k(x^{k-1}) - D_k(\alpha^{k-1}) \le G_{k-1}(x^{k-1}) - D_{k-1}(\alpha^{k-1}) + \frac{\kappa}{2}\|e^{T_{k-1}}_{(k-1)}\|^2 + \kappa\|y^{k-1} - y^{k-2}\|^2.$$

Since $\bar{R}^2 \le nR^2$ and $R_m^2 \le mnR^2$, we have

$$\Psi^K_{3,(k)} \ge G_k^* - D_k(\alpha^K_{(k)}) + \frac{2\rho}{\delta}\|e^K_{(k)}\|^2 \ge G_k^* - D_k(\alpha^K_{(k)}) + \frac{\lambda+\kappa}{\sqrt{2n+\delta mn}}\|e^K_{(k)}\|^2,$$

which implies that

$$G^*_{k-1} - D_{k-1}(\alpha^{k-1}) \le \frac{\sqrt{4n+\delta mn}}{2}\Psi^{T_{k-1}}_{3,(k-1)} \le \frac{1}{2}\epsilon_{k-1},$$

and

$$\frac{\kappa}{2}\|e^{T_{k-1}}_{(k-1)}\|^2 \le \frac{\sqrt{2n+\delta mn}}{2}\cdot\frac{\lambda+\kappa}{\sqrt{2n+\delta mn}}\|e^{T_{k-1}}_{(k-1)}\|^2 \le \frac{\sqrt{2n+\delta mn}}{2}\Psi^{T_{k-1}}_{3,(k-1)} \le \frac{\epsilon_{k-1}}{2}.$$

Moreover, it is easy to see that $G_{k-1}(x^{k-1}) - G^*_{k-1} = G_{k-1}(x^{T_{k-1}+1}_{(k-1)}) - G^*_{k-1} \le \frac{1}{2}\epsilon_{k-1}$. Therefore,

$$\epsilon^0_{D,(k)} \le G_k(x^{k-1}) - D_k(\alpha^{k-1}) \le \frac{3}{2}\epsilon_{k-1} + \kappa\|y^{k-1} - y^{k-2}\|^2.$$

Then similar to the proofs of Proposition 3.2 and Lemma C.1 in [Lin et al., 2015], we can get $\mathbb{E}[T_k] \le \tilde{\mathcal{O}}\left(\frac{1}{\delta} + m + \frac{\mathcal{M}_1}{\lambda+\kappa} + \frac{1}{\delta}\sqrt{\frac{(1-\delta)(\bar{R}^2+\delta R_m^2)}{(\lambda+\kappa)\gamma}}\right)$.

F.2 Proof of Lemma 4.4

From the equality above (33) in the proof of Theorem 3.3 in [Qian et al., 2021a], we know

$$\tilde{u}^{K+1}_{(k)} - \tilde{u}^K_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^{n} A_{i_K^\tau\tau}\Delta\alpha^{K+1}_{i_K^\tau\tau,(k)},$$

where $\tilde{u}^K_{(k)} = u^K_{(k)} + e^K_{(k)}$. Hence, as long as

$$\tilde{u}^0_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^{n}\sum_{i=1}^{m} A_{i\tau}\alpha^0_{i\tau,(k)}, \tag{26}$$

we will have $\tilde{u}^K_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha^K_{i\tau,(k)}$ for all $K \ge 0$. From the initialization rule at each outer iteration, it is easy to see that (26) is satisfied for all $k \ge 1$. Therefore,

$$\tilde{u}^K_{(k)} = \frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^{n}\sum_{i=1}^{m} A_{i\tau}\alpha^K_{i\tau,(k)}, \tag{27}$$

for all $k \ge 1$ and $K \ge 0$. Moreover, from the proofs for EC-SDCA, it is easy to verify that the convergence results still hold as long as (27) is satisfied, and it does not matter whether $u^0_{(k)}$ equal to $\frac{1}{(\lambda+\kappa)N}\sum_{\tau=1}^n\sum_{i=1}^m A_{i\tau}\alpha^0_{i\tau,(k)}$ or not. Then similar to the proof of Lemma 4.1, we have

$$\begin{aligned}
&\mathbb{E}[\sqrt{4n+\delta mn}\Psi^K_{3,(k)} + 2(G_k(x^{K+1}_{(k)}) - G_k^*)] \\
&\le \left(\sqrt{4n+\delta mn} + \frac{2}{\theta}\right)\left(1 - \min\left\{\theta, \frac{\delta}{4}\right\}\right)^K \left(\epsilon^0_{D,(k)} + \frac{2(\rho+\theta(\lambda+\kappa))}{\delta n}\sum_{\tau=1}^{n}\|e^0_{\tau,(k)}\|^2\right),
\end{aligned}$$

and

$$\epsilon_{D,(k)}^{0} \leq \frac{3}{2}\epsilon_{k-1} + \kappa\|y^{k-1} - y^{k-2}\|^2.$$

Moreover, from the initialization of $e_{\tau,(k)}^{0}$, we have

$$\frac{2(\rho + \theta(\lambda + \kappa))}{\delta n}\sum_{\tau=1}^{n}\|e_{\tau,(k)}^{0}\|^2 \leq \Phi_{3,(k-1)}^{T_{k-1}} \leq \frac{1}{2}\epsilon_{k-1}.$$

Then similar to the proofs of Proposition 3.2 and Lemma C.1 in [Lin et al., 2015], we can get the result.