Error Compensated Proximal SGD and RDA

Xun Qian XUN.QIAN@KAUST.EDU.SA
KAUST

Hanze Dong HDONGAJ@UST.HK
Hong Kong University of Science and Technology

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
KAUST

Tong Zhang TONGZHANG@UST.HK
Hong Kong University of Science and Technology

Abstract

Communication cost is a key bottleneck in distributed training of large machine learning models. In order to reduce the amount of communicated data, quantization and error compensation techniques have recently been studied. While the error compensated stochastic gradient descent (SGD) with contraction compressor (e.g., TopK) was proved to have the same convergence rate as vanilla SGD in the smooth case, it is unknown in the regularized case. In this paper, we study the error compensated proximal SGD and error compensated regularized dual averaging (RDA) with contraction compressor for the composite finite-sum optimization problem. Unlike the smooth case, the leading term in the convergence rate of error compensated proximal SGD is dependent on the contraction compressor parameter in the composite case, and the dependency can be improved by introducing a reference point to reduce the compression noise. For error compensated RDA, we can obtain better dependency of compressor parameter in the convergence rate. Extensive numerical experiments are presented to validate the theoretical results.

1. Introduction

In order to train modern large scale machine learning systems, one typically collects a large set of labelled data which is then used to train a supervised statistical model, e.g., logistics regression or a neural network. In many applications, the size of the data set precludes the possibility to solve the problem on a single machine, and the data and the training itself needs be distributed among several machines [28]. In federated learning [13, 14, 16, 18], training occurs on edge devices such as mobile phones and smart home devices, where the data is originally captured.

In this work, we consider distributed machine learning with regularized convex models, which can be posed as the optimization problem

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

where $f(x) := \frac{1}{n}\sum_{\tau} f^{(\tau)}(x)$ is an average of n smooth convex functions $f^{(\tau)} : \mathbb{R}^d \to \mathbb{R}$ distributed over n nodes (devices, computers), and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function representing a possibly nonsmooth regularizer. On each node, $f^{(\tau)}(x)$ is an average of m smooth convex functions

$$f^{(\tau)}(x) = \frac{1}{m} \sum_{i=1}^{m} f_i^{(\tau)}(x),$$

representing the average loss over the training data stored on node τ. For example, in Lasso regression (L_1-regularized least squares) [26], we have $f_i^{(\tau)}(x) = ((z_i^{(\tau)})^\top x - y_i^{(\tau)})^2$, and $\psi(x) = \lambda \|x\|_1$. Here $(z_i^{(\tau)}, y_i^{(\tau)}) \in \mathbb{R}^d \times \mathbb{R}$ represents the i-th training sample on node τ.

We assume that problem (1) has at least one optimal solution x^* and define the variance

$$\sigma_\tau^2 := \frac{1}{m} \sum_{i=1}^{m} \|\nabla f_i^{(\tau)}(x^*) - \nabla f^{(\tau)}(x^*)\|^2,$$

and the average variance $\sigma^2 := \frac{1}{n} \sum_{\tau=1}^{n} \sigma_\tau^2$.

For distributed learning problems of this form, distributed stochastic gradient algorithms are the preferred methods. In these methods, each node computes the stochastic gradient from a minibatch sampled from the locally stored training data, and an aggregated stochastic gradient is subsequently computed by averaging the local stochastic gradients over all machines. In distributed and especially federated settings, communication is generally much slower than local computation, which makes the communication overhead become a key bottleneck. There are several ways to tackle this issue, such as using large mini-batches [8, 32], asynchronous learning [1, 17, 20, 27], and gradient compression [2, 4, 11, 19, 21, 29].

1.1. Gradient compression

The idea of gradient compression, which is the technique investigated in this paper, is to compress the stochastic gradients on each node before sending them over the network for aggregation. Although this approach reduces the communication cost, it introduces errors into the gradients, and thus can slow down convergence. Moreover, for biased compression, the SGD algorithm may not converge. This can be shown by the following simple example:

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} \left(f_1(x) + f_2(x) \right),$$

where $f_1(x) = x_1^2$ and $f_2(x) = 2x_1 + x_2^2$. In SGD, the two stochastic gradients are

$$\begin{pmatrix} 2x_1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 \\ 2x_2 \end{pmatrix}.$$

If we use the initial point $(1, 0.5)^\top$ and run SGD with the Top1 compressor, it is easy to see that the second coordinate x_2 will never be updated.

Two techniques have been invented in the literature to alleviate the slow-down of convergence due to compression: *error compensation* [23, 24, 30] and *variance reduction* [9, 10, 19].

Error compensation. The key idea of error compensation is to remember the accumulated gradient compression error on each node, and use it to compensate the next gradient before this new gradient gets compressed and transmitted. For unbiased compressors, if we assume the accumulated compression error is bounded, the convergence rate of error compensated SGD is the same as vanilla SGD [24]. However, if we only assume bounded stochastic gradient, then in order to guarantee the boundedness of the accumulated quantization error, some decaying factor needs to be involved in general, and error compensated SGD is proved to have some advantages over compressed SGD for convex quadratic problems [30]. On the other hand, for contraction compressors (e.g., TopK compressor [3]), error compensated SGD actually has the same convergence rate as Vanilla SGD [22, 23, 25]. For f is non-smooth, and $\psi = 0$, the error compensated SGD where the stochastic subgradient is used was studied in [12] for single node case, and the convergence rate is of order $O\left(\frac{1}{\sqrt{\delta k}}\right)$.

There are two types of stochastic gradient methods for (1) that are commonly used in practice: proximal SGD [6], and regularized dual averaging (RDA) [31]. In this paper, we investigate the error compensated proximal SGD and error compensated regularized dual averaging (RDA) with contraction compressor for solving (1).

Variance reduction. The second technique for fixing the slow down from compression—one that we do not purse in this paper—is a variance reduction technique tailored to remove the excess variance introduced by unbiased stochastic compression operators. The main idea of these schemes is to maintain an estimate on each node of the gradient of the local function at the optimum, and to use this estimate to perturb the local gradient before compression takes place. The first such method, called SEGA, was proposed in [9], who focused on the single node ($n = 1$) case, and linear compressors, i.e., sketches. However, their theory works for strongly convex objectives and general convex regularizers. In particular, they prove linear convergence to the solution, and their rates can be related to the rates of randomized coordinate descent methods. This work was then extended to the $n > 1$ case, and the DIANA method was developed [19]. The communication complexity of DIANA improves upon the rate of SGD without compression and they show that a certain level of compression, one that depends on problem conditioning and the number of nodes, comes without any increase in the number of communication rounds. In the nonconvex setting the rate is the same as for SGD. This technique has not been explored with biased compression operators. Variance reduction techniques have recently been extended to distributed fixed point methods which compress iterates instead of gradients [5]. A general framework for studying the convergence of SGD, one also unifying classical variance reduced methods for finite sum problems and variance reduced techniques for compression for SGD, was proposed in [7].

1.2. Contributions

(i) For error compensated proximal SGD, under certain conditions, the leading term in the convergence rate is $O(1/\sqrt{n\delta k})$, where δ is the contraction compressor parameter. To obtain ϵ-optimal solution, the iteration complexity is $O(1/(n\delta\epsilon^2))$, which means linear speed up w.r.t. the number of nodes n.

(ii) For error compensated RDA, under certain conditions, the leading term in the convergence rate is $O(1/\sqrt{\sqrt{n\delta}k})$. To obtain ϵ-optimal solution, the iteration complexity is $O(1/(\sqrt{n\delta}\epsilon^2))$, which has better dependency of δ than error compensated proximal SGD, however, in the meantime, only has $\sqrt{n}$ speed up w.r.t the number of nodes n.

2. Gradient Compression Methods

There are mainly two types of compression operators being used in the literature: contraction compressor and unbiased compressor. They are defined as follows.

$Q : \mathbb{R}^d \to \mathbb{R}^d$ is a *contraction compressor* if there is a $0 < \delta \le 1$ such that

$$\mathbb{E}\|x - Q(x)\|^2 \le (1-\delta)\|x\|^2 \tag{2}$$

for all $x \in \mathbb{R}^d$. $\tilde{Q}$ is an *unbiased compressor* if there is $\omega \ge 0$ such that

$$\mathbb{E}[\tilde{Q}(x)] = x \quad \text{and} \quad \mathbb{E}\|\tilde{Q}(x)\|^2 \le (\omega+1)\|x\|^2 \tag{3}$$

for all $x \in \mathbb{R}^d$.

Some frequently used contraction compressors are TopK compressor and RandK compressor [23]. In general, given an arbitrary unbiased compressor, we can obtain a contraction compressor via scaling as follows. For any unbiased compressor $\tilde{Q}$ satisfying (3), $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor satisfying (2) with $\delta = \frac{1}{\omega+1}$. Indeed,

$$\begin{aligned} \mathbb{E}\|\tfrac{1}{\omega+1}\tilde{Q}(x) - x\|^2 &= \tfrac{1}{(\omega+1)^2}\mathbb{E}\|\tilde{Q}(x)\|^2 + \|x\|^2 - \tfrac{2}{\omega+1}\mathbb{E}\langle \tilde{Q}(x), x\rangle \\ &\le \tfrac{1}{\omega+1}\|x\|^2 + \|x\|^2 - \tfrac{2}{\omega+1}\|x\|^2 \;=\; \left(1 - \tfrac{1}{\omega+1}\right)\|x\|^2. \end{aligned}$$

We may use the following assumptions for the contraction compressor in some cases.

Assumption 2.1 $\mathbb{E}[Q(x)] = \delta x$.

It is easy to verify that RandK compressor satisfies Assumption 2.1 with $\delta = \frac{K}{d}$, and $\tilde{Q}/(\omega+1)$, where $\tilde{Q}$ is any unbiased compressor, also satisfies Assumption 2.1 with $\delta = \frac{1}{\omega+1}$.

Assumption 2.2 *For $x_\tau = \frac{\eta}{L_1} g_\tau^k + e_\tau^k \in \mathbb{R}^d$ ($x_\tau = g_\tau^k + e_\tau^k$), $\tau = 1, ..., n$ and $k \ge 0$ in Algorithm 1 (Algorithm 2), there exist $\delta' > 0$ such that $\mathbb{E}[Q(x_\tau)] = Q(x_\tau)$, and $\left\|\sum\limits_{\tau=1}^{n} (Q(x_\tau) - x_\tau)\right\|^2 \le (1-\delta')\left\|\sum\limits_{\tau=1}^{n} x_\tau\right\|^2$.*

For TopK, we have $\mathbb{E}[Q(x)] = Q(x)$ for any $x \in \mathbb{R}^d$. If $Q(x_\tau)$ is close to x_τ, then δ' could be larger than $\frac{K}{d}$. Whenever Assumption 2.2 is needed, if $\delta > \delta'$, we could decrease δ such that $\delta = \min\{\delta, \delta'\}$. In this way, we have the uniform parameter δ for the contraction compressor.

3. Error Compensated Proximal SGD

Proximal SGD is a standard optimization method for large scale machine learning with composite objective functions such as L_1 regularized sparse learning problems [6, 15]. For distributed learning, one needs to compute stochastic gradients on each machine, and then aggregate them over multiple nodes. Gradient compression can be used to reduce the communication cost. In this section, we present an error compensated proximal SGD algorithm, and analyze its convergence in distributed training.

The error compensated proximal SGD algorithm proposed in this paper is given in Algorithm 1. The following is a high-level description of this method. All nodes maintain the same copies of x^k, w^k, y^k, and u^k. On each node, an error vector e_τ^k is maintained, and is updated by the compression error at last iteration:

$$e_\tau^{k+1} = e_\tau^k + \gamma g_\tau^k - Q(\gamma g_\tau^k + e_\tau^k);$$

a scalar u_τ^k is also maintained, and only u_1^k will be updated. The summation of u_τ^k is u^k, and we use u^k to control the update frequency of the reference point w^k. At each iteration, each node calculates a stochastic gradient $\nabla f_{i_k^\tau}^{(\tau)}$ and subtracts $\nabla f^\tau(w^k)$ from it to reduce the noise. Then all nodes compress the sum of γg_τ^k and the error vector e_τ^k, independently of each other, and each sends their compressed vector y_τ^k and u_τ^{k+1} to the other nodes. If $u^k = 1$, each node also sends $\nabla f^{(\tau)}(w^k)$ to the other nodes. After each node accumulates the compressed vectors, the full gradient at the reference point $\gamma \nabla f(w^k)$ needs to be added. The proximal step is taken on each node, where we use the standard proximal operator: $\mathrm{prox}_{\gamma\psi}(x) := \arg\min_y \left\{ \frac{1}{2}\|x-y\|^2 + \gamma\psi(y) \right\}$. The reference point w^k will be updated if $u^{k+1} = 1$. It is easy to see that w^k will be updated with propobility p at each iteration.

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

We introduce some notations to reveal some relations between iteration k and $k+1$ in Algorithm 1. Let $e^k = \frac{1}{n}\sum_{\tau=1}^n e_\tau^k$, $g^k = \frac{1}{n}\sum_{\tau=1}^n g_\tau^k$, and $\tilde{x}^k = x^k - e^k$ for $k \geq 0$. Then

$$e^{k+1} = \frac{1}{n}\sum_{\tau=1}^n \left(e_\tau^k + \gamma g_\tau^k - y_\tau^k\right) = e^k + \gamma g^k - y^k,$$

and

$$\begin{aligned} \tilde{x}^{k+1} &= x^{k+1} - e^{k+1} \\ &= x^{k+0.5} - \gamma\partial\psi(x^{k+1}) - e^{k+1} \\ &= x^k - y^k - \gamma\nabla f(w^k) - \gamma\partial\psi(x^{k+1}) - e^k - \gamma g^k + y^k \\ &= \tilde{x}^k - \gamma(g^k + \nabla f(w^k) + \partial\psi(x^{k+1})) \\ &= \tilde{x}^k - \gamma\left(\frac{1}{n}\sum_{\tau=1}^n \nabla f_{i_k^\tau}^{(\tau)}(x^k) + \partial\psi(x^{k+1})\right). \end{aligned}$$

The above equality plays a significant role in the convergence analysis. In fact, for uncompressed distributed proximal SGD (that is, when $Q(x) = x$ for all x), we have the following iteration:

$$x^{k+1} = x^k - \gamma \left(\frac{1}{n} \sum_{\tau=1}^{n} \nabla f^{(\tau)}_{i_k^\tau}(x^k) + \partial \psi(x^{k+1}) \right).$$

We have the following analogy for the error-compensated compressed distributed proximal SGD of Algorithm 1. For simplicity, we denote $\tilde{g}^k = \frac{1}{n} \sum_{\tau=1}^{n} \nabla f^{(\tau)}_{i_k^\tau}(x^k)$. Then

$$\tilde{x}^{k+1} = \tilde{x}^k - \gamma(\tilde{g}^k + \partial \psi(x^{k+1})),$$

and $\mathbb{E}_k[\tilde{g}^k] = \nabla f(x^k)$, where $\mathbb{E}_k[\cdot]$ denotes the conditional expectation on x^k.

This recursive relationship implies that with error-compensation, the iterate of $\tilde{x}$ is analogous to that of standard proximal SGD. Therefore the analysis of proximal SGD can be adapted to analyze Algorithm 1. In order to develop a convergence theory for our method, we need the following assumption.

Assumption 3.1 *$f_i^{(\tau)}$ is L-smooth for $1 \le i \le m$ and $1 \le \tau \le n$.*

The convergence of Algorithm 1 is given by the following theorem. It shows that the leading term of $O(1/\sqrt{k})$ is reduced to $O(1/\sqrt{nk})$ with n nodes. This implies that under suitable conditions, linear speed up of convergence can be achieved when we increase the number of nodes.

Theorem 1 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Let $\bar{x}^k := \frac{1}{k} \sum_{j=1}^{k} x^j$. (i) If $p = 0$, then there exists a constant stepsize $\gamma \le \frac{\delta^2}{48L}$ such that*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \quad = \quad O\left(\frac{L\|x^0 - x^*\|^2}{\delta^2 k} + \frac{\|x^0 - x^*\|\sqrt{\sigma^2/\delta + L(P(w^0) - P(x^*))/\delta^2}}{\sqrt{k}} \right).$$

If $p > 0$, then there exists a constant stepsize $\gamma \le \frac{\delta^2}{80L}$ such that

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \quad = \quad O\left(\frac{1}{k}\left(\frac{L\|x^0 - x^*\|^2}{\delta^2} + \frac{P(w^0) - P(x^*)}{p} \right) + \frac{\sigma\|x^0 - x^*\|}{\sqrt{\delta k}} \right).$$

(ii) Under Assumption 2.1 or Assumption 2.2, if $p = 0$, then there exists a constant stepsize $\gamma \le \frac{\delta^2}{(64 + 304/n)L}$ such that

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \quad = \quad O\left(\frac{L\|x^0 - x^*\|^2}{\delta^2 k} + \frac{\|x^0 - x^*\|\sqrt{\sigma^2/(n\delta) + L(P(w^0) - P(x^*))/\delta^2}}{\sqrt{k}} \right).$$

If $p > 0$, then there exists a constant stepsize $\gamma \le \frac{\delta^2}{(128 + 592/n)L}$ such that

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \quad = \quad O\left(\frac{1}{k}\left(\frac{L\|x^0 - x^*\|^2}{\delta^2} + \frac{P(w^0) - P(x^*)}{p} \right) + \frac{\sigma\|x^0 - x^*\|}{\sqrt{n\delta k}} \right).$$

The following results, with easier to interpret leading terms, are direct consequences of the above theorem.

Corollary 2 *Under the premise of Theorem 1. Choose the same stepsize as in Theorem 1 in each case.*

(i) If $p = 0$ *and* $k \geq O\left(\frac{1}{\delta^2}\right)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\delta\sqrt{k}}\right).$$

If $p > 0$ *and* $k \geq O\left(\frac{1}{\delta^3} + \frac{\delta}{p^2}\right)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{\delta k}}\right).$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$ *and* $k \geq O\left(\frac{1}{\delta^2}\right)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\delta\sqrt{k}}\right).$$

If $p > 0$ *and* $k \geq O\left(\frac{n}{\delta^3} + \frac{n\delta}{p^2}\right)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{n\delta k}}\right).$$

The above corollary shows that by choosing $p > 0$, the dependency of the contraction compressor parameter δ in the leading term of convergence rate can be improved from $1/\delta$ to $1/\sqrt{\delta}$. Furthermore, under Assumption 2.1 or Assumption 2.2, the convergence rate can be improved to $O\left(1/\sqrt{n\delta k}\right)$ with n nodes. In other words, $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$ as long as $k = O(1/(n\delta\epsilon^2))$, which gives linear speed up w.r.t. the number of nodes n.

References

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

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

[3] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

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

[5] Sélim Chraibi, Ahmed Khaled, Dmitry Kovalev, Adil Salim, Peter Richtárik, and Martin Takáč. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:1912.09925*, 2019.

[6] John C. Duchi, Shai Shalev-shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *In Adam Tauman Kalai and Mehryar Mohri, editors, Proc. of the 23th Annual Conference on Learning Theory (COLT'10*, pages 14–26, 2010.

[7] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[8] 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.

[9] Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. SEGA: variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems 31*, pages 2082–2093, 2018.

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

[11] Samuel Horváth, Chen-Yu Ho, Ľudovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

[12] 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.

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

[14] 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*, 2016.

[15] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] S. U. Stich, J. B. Cordonnier, and M. Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[24] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu. Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, pages 7652–7662, 2018.

[25] H. Tang, X. Lian, T. Zhang, and J. Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. *The 36th International Conference on Machine Learning*, pages 6155–6165, 2019.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. URL http://www.jstor.org/stable/2346178.

[27] 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.

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

[29] 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.

[30] J. Wu, W. Huang, J. Huang, and T. Zhang. Error compensated quantized sgd and its applications to large-scale distributed optimization. *The 35th International Conference on Machine Learning*, pages 5321–5329, 2018.

[31] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

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

Appendix

Appendix A. Error Compensated RDA

RDA is a stochastic optimization method for solving large scale machine learning with composite objective functions, and has advantages over proximal SGD in terms of achieving better sparsity for L_1 regularized sparse learning problems [31]. Due to this desirable property, it has received significant attention. Similar to proximal-SGD, gradient compression can be used to reduce the communication cost in the distributed learning setting.

The proposed algorithm is described in Algorithm 2.

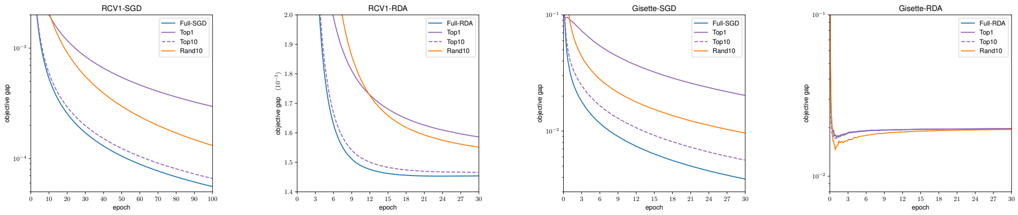


Figure 1: Error Compensated Proximal and Full SGD/RDA.

```
Algorithm 2: Error compensated RDA
Parameters: an auxiliary function $h(x)$ that is strongly onvex on dom $\psi$ and also satisfies
    $$\arg\min_x h(x) \in \arg\min_x \psi(x);$$
a nonegative and nondecreasing sequence $\{\beta_k\}_{k\geq 1}$
Initialization: $x^1 = w^1 = \arg\min_x h(x)$; $\bar{g}^0 = 0 \in \mathbb{R}^d$; $e^1_\tau = 0 \in \mathbb{R}^d$; $u^1 = 1 \in \mathbb{R}$
for $k = 1, 2, ...$ do
    for for $\tau = 1, ..., n$ do
        Sample $i^\tau_k$ uniformly and independently in $[m]$ on each node
        $g^k_\tau = \nabla f^{(\tau)}_{i^\tau_k}(x^k) - \nabla f^{(\tau)}(w^k)$,   $y^k_\tau = Q(g^k_\tau + e^k_\tau)$,   $e^{k+1}_\tau = e^k_\tau + g^k_\tau - y^k_\tau$
        $u^{k+1}_\tau = 0$ for $\tau = 2, ..., n$ ,   $u^{k+1}_1 = \begin{cases} 1 & \text{with propobility } p \\ 0 & \text{with probability } 1-p \end{cases}$
        Send $y^k_\tau$ and $u^{k+1}_\tau$ to the other nodes. Send $\nabla f^{(\tau)}(w^k)$ to the other nodes if $u^k = 1$
        Receive $y^k_\tau$ and $u^{k+1}_\tau$ from the other nodes. Receive $\nabla f^{(\tau)}(w^k)$ from the other nodes if
          $u^k = 1$
    end
    $y^k = \frac{1}{n}\sum_{\tau=1}^n y^k_\tau$,   $u^{k+1} = \sum_{\tau=1}^n u^{k+1}_\tau$,   $\bar{g}^k = \frac{k-1}{k}\bar{g}^{k-1} + \frac{1}{k}(y^k + \nabla f(w^k))$
    $x^{k+1} = \arg\min_x\{\langle \bar{g}^k, x\rangle + \psi(x) + \frac{\beta_k}{k}h(x)\}$,   $w^{k+1} = \begin{cases} x^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$
end
```

At the high level, the way to calculate the search directions and error vectors in Algorithm 2 is the same as that of Algorithm 1. The difference is that x^k is updated by using the average of the gradient estimators in the past.

Same as in the error compensated proximal SGD, we also introduce some notations to show the important connections between iteration $k-1$ and k in Algorithm 2. Let $e^k = \frac{1}{n}\sum_{\tau=1}^n e^k_\tau$, $g^k = \frac{1}{n}\sum_{\tau=1}^n g^k_\tau$, $s^k = k\bar{g}^k$ and $\tilde{s}^k = s^k + e^{k+1}$ for $k \geq 1$. Then we have

$$e^{k+1} = \frac{1}{n}\sum_{\tau=1}^n \left(e^k_\tau + g^k_\tau - y^k_\tau\right) = e^k + g^k - y^k,$$

and

$$\begin{aligned}
\tilde{s}^k &= k\bar{g}^k + e^{k+1} \\
&= (k-1)\bar{g}^{k-1} + y^k + \nabla f(w^k) + e^k + g^k - y^k \\
&= (k-1)\bar{g}^{k-1} + e^k + g^k + \nabla f(w^k) \\
&= (k-1)\bar{g}^{k-1} + e^k + \tfrac{1}{n}\sum_{\tau=1}^{n} \nabla f^{(\tau)}_{i_k^\tau}(x^k) \\
&= \tilde{s}^{k-1} + \tfrac{1}{n}\sum_{\tau=1}^{n} \nabla f^{(\tau)}_{i_k^\tau}(x^k).
\end{aligned}$$

For simplicity, we denote $\tilde{g}^k = \frac{1}{n}\sum_{\tau=1}^{n} \nabla f^{(\tau)}_{i_k^\tau}(x^k)$. Then

$$\tilde{s}^k = \tilde{s}^{k-1} + \tilde{g}^k, \tag{4}$$

and $\mathbb{E}_k[\tilde{g}^k] = \nabla f(x^k)$. Since $\tilde{s}^0 = s^0 = 0$, equality (4) shows that $\tilde{s}^k$ is the summation of the averaging stochastic gradient from iteration 1 to k. From $\tilde{s}^k$, we can define an auxiliary variable $\tilde{x}^{k+1}$ as follows:

$$\tilde{x}^{k+1} = \arg\min_x \left\{ \left\langle \frac{\tilde{s}^k}{k}, x \right\rangle + \psi(x) + \frac{\beta_k}{k} h(x) \right\},$$

for $k \ge 1$, and $\tilde{x}^1 = x^1$.

The above recursive relationship is again analogous to a similar equation for the standard RDA without compression. This means that the iterate of $\tilde{x}$ and $\tilde{s}$ from Algorithm 2 behave similarly as those of the standard RDA. Therefore we can adapt the standard RDA convergence analysis to analyze Algorithm 2.

Next, we will present a convergence analysis of the proposed error compensated RDA method. We need the following assumptions in our theoretical analysis.

Assumption A.1 *$f_i^{(\tau)}$ is L-smooth. h is 1-strongly convex and $h(x^1) = \psi(x^1) = 0$.*

Assumption A.2 *In Algorithm 2, $\|\nabla f^{(\tau)}_{i_k^\tau}(x^k)\|^2 \le G^2$, $\|\nabla f^{(\tau)}(w^k)\|^2 \le G^2$, and $\|\partial h(x^k)\|^2 \le H^2$ for $k \ge 1$. $h(x^*) \le D^2$.*

The convergence of Algorithm 2 is given by the following theorem. It shows that the leading term of $O(1/\sqrt{\delta k})$ can be reduced to $O(1/\sqrt{\sqrt{n\delta} k})$ with n nodes for $n\delta \le 1$ and $p > 0$. It is not surprising that we only get $\sqrt{n}$ speed up with n nodes, since from the convergence analysis of RDA in [31], increasing the minibatch size would not imply linear speed up of the upper bound of the averaging stochastic gradients.

Theorem 3 *Assume the compressor Q in Algorithm 2 is a contraction compressor and Assumptions A.1, A.2 hold. Let $\bar{x}^k := \frac{1}{k}\sum_{j=1}^{k} x^j$.*

(i) If $p = 0$, then for fixed $k \ge O(1/\delta)$, by choosing $\beta_j = 4\sqrt{\frac{k}{\delta}}\frac{\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2/4}}{D}$ for $j \ge 1$, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\tfrac{D}{\sqrt{\delta k}}\sqrt{G^2 + L(P(w^1) - P(x^*)) + \delta\sigma^2} + \left(\tfrac{DG}{\delta\sqrt{\delta k}} + H^2 + \tfrac{G^2}{\delta^2} \right) \tfrac{\ln k}{k} \right).$$

If $p > 0$, then for fixed $k \ge O(1/\delta^{\frac{5}{2}})$, by choosing $\beta_j = \frac{4\sqrt{k}}{\delta^{1/4}}\frac{\sqrt{\sigma^2+24G^2}}{D}$ for $j \ge 1$, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{D\sqrt{\sigma^2+G^2}}{\delta^{1/4}\sqrt{k}} + \frac{LD(P(w^1)-P(x^*))}{k\sqrt{k}\delta^{5/4}p\sqrt{\sigma^2+G^2}} + \left(\frac{DG}{\sqrt{k}\delta^{7/4}} + H^2 + \frac{G^2}{\delta^2} \right) \frac{\ln k}{k} \right).$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$, then for fixed $k \ge O(1/\delta)$, by choosing $\beta_j = 4\sqrt{\frac{k}{\delta}}\frac{\sqrt{G^2+(2+9/n)L(P(w^1)-P(x^*))+3\delta\sigma^2/n}}{D}$ for $j \ge 1$, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\tfrac{D}{\sqrt{\delta k}}\sqrt{G^2 + L(P(w^1) - P(x^*)) + \delta\sigma^2/n} + \left(\tfrac{DG}{\delta\sqrt{\delta k}} + H^2 + \tfrac{G^2}{\delta^2} \right) \tfrac{\ln k}{k} \right).$$

If $p > 0$ and $n\delta \le 1$, then for fixed $k \ge O(n^{\frac{3}{2}}/\delta^{\frac{5}{2}})$, by choosing $\beta_j = \frac{4\sqrt{k}}{(n\delta)^{1/4}}\frac{\sqrt{6\sigma^2+12G^2}}{D}$ for $j \ge 1$, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{D\sqrt{\sigma^2+G^2}}{(n\delta)^{1/4}\sqrt{k}} + \frac{n^{3/4}LD\big(P(w^1)-P(x^*)\big)}{k\sqrt{k}\delta^{5/4}p\sqrt{\sigma^2+G^2}} + \left(\frac{n^{1/4}DG}{\sqrt{k}\delta^{7/4}} + H^2 + \frac{G^2}{\delta^2} \right) \frac{\ln k}{k} \right).$$

To interpret the leading terms easily, we show the direct consequences of the above theorem as follows.

Corollary 4 *Under the premise of Theorem 3. Choose the same stepsize as Theorem 3 in each case. (i) If* $p = 0$ *and* $k \geq O(\frac{1}{\delta^3}(\ln\frac{1}{\delta})^2)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{\delta} k}\right).$$

If $p > 0$ *and* $k \geq O(\frac{1}{\delta p} + \frac{1}{\delta^{3.5}}(\ln\frac{1}{\delta})^2)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{\sqrt{\delta} k}}\right).$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$ *and* $k \geq O(\frac{1}{\delta^3}(\ln\frac{1}{\delta})^2)$*, we have*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{\delta} k}\right).$$

If $p > 0$*,* $n\delta \leq 1$*, and* $k \geq O(\frac{n}{\delta p} + \frac{1}{\delta^{3.5}}(\ln\frac{1}{\delta})^2\sqrt{n}(\ln n)^2)$*, then*

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] = O\left(\frac{1}{\sqrt{\sqrt{n\delta} k}}\right).$$

The above corollary shows that for both cases: $p = 0$ or $p > 0$, the dependence of the convergence rate on the contraction compressor parameter is better than for error compensated proximal SGD. However, to let the leading terms be dominant, the dependence of δ in the lower bound of k is worse than for error compensated proximal SGD.

Appendix B. Communication Cost

In error compensated proximal SGD and error compensated RDA, when w^k is updated, the uncompressed vector need to be transmitted. We denote Δ_1 as the communication cost of the uncompressed vector $x \in \mathbb{R}^d$. Define the compress ratio $r(Q)$ for the contraction compressor Q as

$$r(Q) := \sup_{x \in \mathbb{R}^d} \left\{ \mathbb{E}\left[\frac{\text{communication cost of } Q(x)}{\Delta_1} \right] \right\}.$$

Denote the expected communication cost for k iterations as $\mathcal{T}_k$. Then if we do not count the communication cost of $\nabla f^{(\tau)}(w^k)$, then $\mathcal{T}_k$ is bounded by

$$\mathcal{T}_k \leq (\Delta_1 r(Q) + 1)k, \tag{5}$$

where 1 bit is needed to communicate u_τ^k. When $p > 0$, the expected communication cost at each iteration is bounded by $\Delta_1 r(Q) + 1 + p\Delta_1$, which indicates

$$\mathcal{T}_k \leq \Delta_1 + (\Delta_1 r(Q) + 1 + p\Delta_1)k \leq (\Delta_1 r(Q) + 1)\left(1 + \frac{p + \frac{1}{k}}{r(Q)}\right)k. \tag{6}$$

Hence, when $p \leq O(r(Q))$ and $k \geq O(\frac{1}{r(Q)})$, the expected extra communication cost caused by sending the uncompressed $\nabla f^{(\tau)}(w^k)$ can be controlled. Morevoer, since p is not in the dominant term in the convergence rate, the asymptotic convergence rate will not be affected. Thereofore, for effieiently small ϵ, from Corollary 2 and Corollary 4, we can get $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$ for error compensated proximal SGD if $\mathcal{T}_k = O((\Delta_1 r(Q) + 1)\frac{1}{\delta\epsilon^2})$, and $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$ for error compensated RDA if $\mathcal{T}_k = O((\Delta_1 r(Q) + 1)\frac{1}{\sqrt{\delta}\epsilon^2})$. For proximal SGD, to obtain an ϵ-optimal solution, $\mathcal{T}_k = O\left(\Delta_1 \frac{1}{\epsilon^2}\right)$. Hence, for $\Delta_1 r(Q) \geq O(1)$, if $\frac{r(Q)}{\delta} < 1$ ($\frac{r(Q)}{\sqrt{\delta}} < 1$), then $\mathcal{T}_k$ of the error compensated proximal SGD (RDA) is less than that of proximal SGD. For TopK compressor, $r(Q) = \frac{K(64 + \lceil \log d \rceil)}{64d}$, and in practice δ can be much larger than $\frac{K}{d}$, sometimes even in order $O(1)$.

Table 1: Dataset statistics

DATASET	SPARSE	d	mn	λ_1	λ_2
RCV1	✓	47,236	20,242	10^{-5}	10^{-4}
GISETTE		5,000	6,000	10^{-2}	10^{-2}

Appendix C. Experiments

In this section, we will conduct experiments to validate our theory. In particular, we will (i) investigate the efficiency of our error compensated proximal SGD/RDA methods; (ii) provide a comparison to demonstrate superior convergence properties against a quantization-based method; (3) demonstrate the impact on the parameter p.

Dataset: Our experiment involved two datasets, namely, *RCV1*, *Gisette*. Note that *RCV1* is saved in a sparse format. Dataset information is provided in Table 1. We also provide additional datasets in the Appendix.

Remark: Conventional QSGD [2] does not provide convergence properties for non-smooth proximal case, so we choose DIANA[1] [10, 19] as our competitor, which support the non-smooth case theoretically. Besides, we refer to *k-bit* as the quantization compressor in [2] with level $s = 2^k$.

Table 2: Communication cost (unit: 1×10^3 bytes) per iteration.

DATASET	FULL	$K = 1$	$K = 10$	1-BIT	2-BIT
RCV1	0.89	0.01	0.10	0.32	0.33
GISETTE	40	0.01	0.10	0.23	0.40

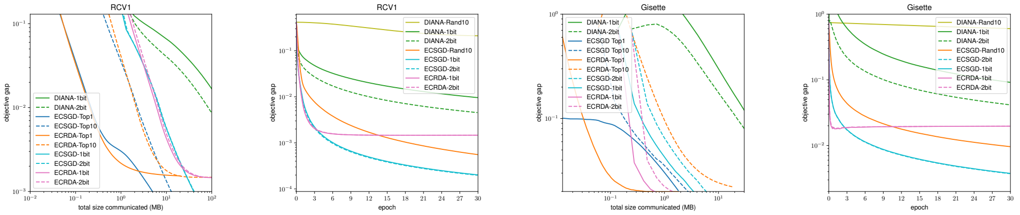


Figure 2: Comparison to Quantization and RandK-DIANA

Settings: We implement our algorithm on L_1-L_2 regularized logistic regression[2] (corresponding regularizer coefficient λ_1, λ_2). All time-dependent experiments in this work were run on a machine with 2 processors (14-core): Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz. System environment includes: Ubuntu 18.04, Python 3.7.3, numpy, sklearn, and mpi4py. Initial stepsize is tuned from $\{10^{-t}, t = -1, 0, \cdots, 4\}$. Step size decay for SGD is based on $\gamma_0/(1 + \lambda_1\gamma_0 t + \lambda_2\gamma_0 t)$. Similarly, β for RDA is chosen by $1 + \gamma_0\sqrt{t}$. For DIANA, we choose $\alpha = 1/(\omega + 1)$. We compute the loss by the weighted average of all x^j as in [22] (This step is mainly to make the plot smoother, and the convergence speed of the diagram is not much different from the real one).

C.1. Comparison to lossless gradient

First, we study the convergence behavior of our algorithm. Figure 1 suggests that TopK strategy is significantly superior to RandK, one of the reasons is that the effective features are concentrated in a few coordinates only. In this case, the TopK algorithm is quite efficient, so that the communication in 10 dimensions can estimate the lossless gradient. However, as shown in Table 2, the communication cost of $K = 10$ is only 1/7 for RCV1 (sparse dataset) and 1/300 for Gisette (non-sparse dataset). This makes the contraction operator very promising in distributed scenarios.

[1] We only discuss quantization and RandK compressors. Other variants of DIANA are not included in this work.

[2] Although our algorithm supports general convex case, we implement L_1-L_2 regularization to make comparison with DIANA. We also have experiments with L_1 regularization in the Appendix.

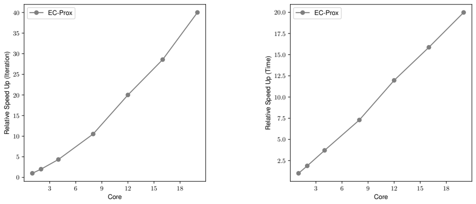


Figure 3: Distributed Error Compensated Proximal SGD (EC-Prox: Top10) on Gisette ($\lambda_1 = \lambda_2 = 10^{-2}$)

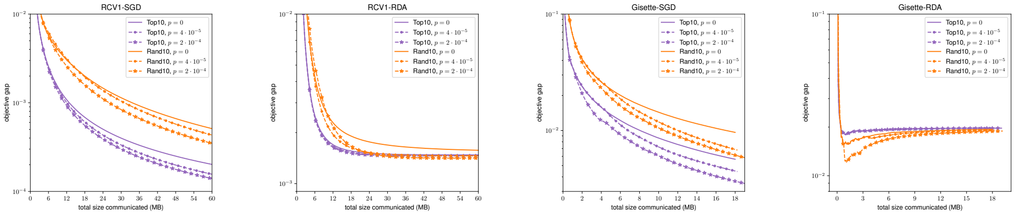


Figure 4: Error Compensated Proximal SGD/RDA comparison with different p.

C.2. Comparison to quantization

Most communication-efficient distributed optimization methods construct unbiased gradient estimates for transmission. One of the most common algorithms is quantization, which reduces the amount of transmitted information by sacrificing transmission precision. But in practice, it is not the most efficient one. In this part, we use random dithering as the quantization operation here ($s = 2^1, 2^2$). Slightly different from our algorithm, DIANA encodes the gradient of both f and L_2 regularization to keep it strongly convex. Figure 2 and Table 2 indicate that our error compensated proximal algorithm based on the TopK operator outperforms the variant of 1-bit quantization-based method in terms of both convergence speed and communication cost. Error compensation also works well in RandK setting.

C.3. Distributed results: linear speed up

In this subsection, we show the performance of the error compensated proximal SGD algorithm with Top10 compressor for different number of nodes. From Figure 3, we can see the linear speed up of the error compensated proximal SGD with respect to the number of nodes.

C.4. Impact of p

In Figure 4, we compare the behaviors under different probabilities p. It shows that although $p = 0$ also has a good performance, there is a certain increase in increasing the probability p.

Appendix D. Lemmas 5, 6, and 7

Lemma 5 *If $\gamma \leq \frac{1}{4L}$, then*

$$\mathbb{E}\|\tilde{x}^{k+1} - x^*\|^2 \leq \mathbb{E}\|\tilde{x}^k - x^*\|^2 + 2\gamma\mathbb{E}(P(x^*) - P(x^{k+1})) + \mathbb{E}\|e^k\|^2 + \mathbb{E}\|e^{k+1}\|^2 + 4\gamma^2\mathbb{E}\|\tilde{g}^k - \nabla f(x^k)\|^2.$$

Lemma 6 *If $f_i^{(\tau)}$ is L-smooth, then we have*

$$\mathbb{E}\|\tilde{g}^k - \nabla f(x^k)\|^2 \leq \frac{4L}{n}\mathbb{E}[P(x^k) - P(x^*)] + \frac{2}{n}\sigma^2. \tag{7}$$

Lemma 7 *(i) If $p = 0$, we have*

$$\sum_{j=0}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \frac{16L\gamma^2}{\delta^2} \sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{4\gamma^2(k+1)}{\delta} \left(\sigma^2 + \frac{4L}{\delta}(P(w^0) - P(x^*)) \right).$$

If $p > 0$, we have

$$\sum_{j=0}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \frac{32L\gamma^2}{\delta^2} \sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{16L\gamma^2}{\delta^2 p} \left(P(w^0) - P(x^*)\right) + \frac{4}{\delta}\sigma^2\gamma^2(k+1).$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$, we have

$$\sum_{j=0}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \tfrac{16L\gamma^2}{\delta^2}(2 + \tfrac{9}{n}) \sum_{j=0}^{k} \left(\mathbb{E}[P(x^j) - P(x^*)] \right) + \tfrac{16\gamma^2(k+1)}{\delta} \left(\tfrac{3\sigma^2}{n} + \tfrac{(2+9/n)L}{\delta}(P(w^0) - P(x^*)) \right).$$

If $p > 0$, we have

$$\sum_{j=0}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \tfrac{32L\gamma^2(2+\frac{9}{n})}{\delta^2} \sum_{j=0}^{k} \left(\mathbb{E}[P(x^j) - P(x^*)] \right) + \tfrac{16L\gamma^2(2+\frac{9}{n})}{\delta^2 p} \left(P(w^0) - P(x^*)\right) + \tfrac{48}{n\delta}\sigma^2\gamma^2(k+1).$$

Appendix E. Proofs of Lemmas 5, 6, and 7

E.1. Proof of Lemma 5

Since $\tilde{x}^{k+1} = \tilde{x}^k - \gamma(\tilde{g}^k + \partial\psi(x^{k+1}))$, we have

$$\begin{aligned}
\langle \gamma\tilde{g}^k, x^* - x^{k+1} \rangle &= \langle \tilde{x}^k - \tilde{x}^{k+1} - \gamma\partial\psi(x^{k+1}), x^* - x^{k+1} \rangle \\
&= \langle \tilde{x}^k - x^{k+1}, x^* - x^{k+1} \rangle + \langle x^{k+1} - \tilde{x}^{k+1}, x^* - x^{k+1} \rangle - \gamma\langle \partial\psi(x^{k+1}), x^* - x^{k+1} \rangle \\
&\geq \frac{1}{2}\left(-\|\tilde{x}^k - x^*\|^2 + \|\tilde{x}^k - x^{k+1}\|^2 + \|x^{k+1} - x^*\|^2 \right) + \frac{1}{2}\left(\|\tilde{x}^{k+1} - x^*\|^2 \right. \\
&\quad \left. -\|x^{k+1} - \tilde{x}^{k+1}\|^2 - \|x^{k+1} - x^*\|^2 \right) + \gamma\left(\psi(x^{k+1}) - \psi(x^*) \right) \\
&= \frac{1}{2}\|\tilde{x}^{k+1} - x^*\|^2 - \frac{1}{2}\|\tilde{x}^k - x^*\|^2 + \frac{1}{2}\|\tilde{x}^k - x^{k+1}\|^2 - \frac{1}{2}\|\tilde{x}^{k+1} - x^{k+1}\|^2 \\
&\quad +\gamma\left(\psi(x^{k+1}) - \psi(x^*) \right).
\end{aligned}$$

From $\|\tilde{x}^k - x^{k+1}\|^2 \geq \frac{1}{2}\|x^{k+1} - x^k\|^2 - \|\tilde{x}^k - x^k\|^2$, and $\|x^{k+1} - x^*\|^2 \geq \frac{1}{2}\|\tilde{x}^{k+1} - x^*\|^2 - \|\tilde{x}^{k+1} - x^{k+1}\|^2$, we arrive at

$$\begin{aligned}
\langle \gamma\tilde{g}^k, x^* - x^{k+1} \rangle &\geq \frac{1}{2}\|\tilde{x}^{k+1} - x^*\|^2 - \frac{1}{2}\|\tilde{x}^k - x^*\|^2 + \frac{1}{4}\|x^{k+1} - x^k\|^2 - \frac{1}{2}\|\tilde{x}^k - x^k\|^2 \\
&\quad -\frac{1}{2}\|\tilde{x}^{k+1} - x^{k+1}\|^2 + \gamma(\psi(x^{k+1}) - \psi(x^*)).
\end{aligned} \tag{8}$$

Since f is convex and $\mathbb{E}_k[\tilde{g}^k] = \nabla f(x^k)$, we have

$$\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \\
&= f(x^k) + \mathbb{E}_k[\langle \tilde{g}^k, x^* - x^{k+1} + x^{k+1} - x^k \rangle] \\
&= f(x^k) + \mathbb{E}_k[\langle \tilde{g}^k, x^* - x^{k+1} \rangle] + \mathbb{E}_k[\langle \tilde{g}^k - \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\quad +\mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\geq \mathbb{E}_k[f(x^{k+1})] - \frac{L}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[\langle \tilde{g}^k, x^* - x^{k+1} \rangle] \\
&\quad +\mathbb{E}_k[\langle \tilde{g}^k - \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\geq \mathbb{E}_k[f(x^{k+1})] - \frac{L}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[\langle \tilde{g}^k, x^* - x^{k+1} \rangle] \\
&\quad -\frac{1}{2\beta}\mathbb{E}_k[\|\tilde{g}^k - \nabla f(x^k)\|^2] - \frac{\beta}{2}\mathbb{E}_k[x^{k+1} - x^k\|^2],
\end{aligned}$$

where the second inequality comes from that f is L-smooth and the last inequality comes from Young's inequality.

By choosing $\beta = \frac{1}{4\gamma}$, we can obtain

$$\begin{aligned} f(x^*) &\geq \mathbb{E}_k[f(x^{k+1})] - \left(\frac{L}{2} + \frac{1}{8\gamma}\right)\mathbb{E}_k\|x^{k+1} - x^k\|^2 + \mathbb{E}_k[\langle \tilde{g}^k, x^* - x^{k+1}\rangle] - 2\gamma\mathbb{E}_k\|\tilde{g}^k - \nabla f(x^k)\|^2 \\ &\overset{(8)}{\geq} \mathbb{E}_k[f(x^{k+1})] + \left(\frac{1}{4\gamma} - \frac{L}{2} - \frac{1}{8\gamma}\right)\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{1}{2\gamma}\mathbb{E}_k\|\tilde{x}^{k+1} - x^*\|^2 - \frac{1}{2\gamma}\|\tilde{x}^k - x^*\|^2 \\ &\quad -\frac{1}{2\gamma}\|\tilde{x}^k - x^k\|^2 - \frac{1}{2\gamma}\mathbb{E}_k\|\tilde{x}^{k+1} - x^{k+1}\|^2 + \mathbb{E}_k[\psi(x^{k+1})] - \psi(x^*) - 2\gamma\mathbb{E}_k\|\tilde{g}^k - \nabla f(x^k)\|^2. \end{aligned}$$

Noticing that $\frac{1}{4\gamma} - \frac{L}{2} - \frac{1}{8\gamma} \geq 0$ if $\gamma \leq \frac{1}{4L}$, we can get the result after rearrangement.

E.2. Proof of Lemma 6

Since f_i is L-smooth, we have

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L(f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle),$$

for any $x, y \in \mathbb{R}^d$. Moreover, $\mathbb{E}_k[\nabla f^{(\tau)}_{i_k^\tau}(x^k)] = \nabla f^\tau(x^k)$, and i_k^τ is sampled independently for $\tau = 1, ..., n$. Therefore,

$$\begin{aligned} \mathbb{E}\|\tilde{g}^k - \nabla f(x^k)\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^n \nabla f^{(\tau)}_{i_k^\tau}(x^k) - \frac{1}{n}\sum_{\tau=1}^n \nabla f^{(\tau)}(x^k)\right\|^2 \\ &= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}(x^k)\|^2. \end{aligned}$$

For $\mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}(x^k)\|^2$, we have

$$\begin{aligned} &\mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}(x^k)\|^2 \\ = \;& \mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}_{i_k^\tau}(x^*) + \nabla f^{(\tau)}_{i_k^\tau}(x^*) - \nabla f^{(\tau)}(x^*) + \nabla f^{(\tau)}(x^*) - \nabla f^{(\tau)}(x^k)\|^2 \\ \leq \;& 2\mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}_{i_k^\tau}(x^*) - (\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^*))\|^2 + 2\sum_{\tau=1}^n \mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^*) - \nabla f^{(\tau)}(x^*)\|^2 \\ \leq \;& 2\mathbb{E}\|\nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}_{i_k^\tau}(x^*)\|^2 + 2\sigma_\tau^2 \\ \leq \;& 4L\mathbb{E}[f^{(\tau)}(x^k) - f^{(\tau)}(x^*) - \langle \nabla f^{(\tau)}(x^*), x^k - x^*\rangle] + 2\sigma_\tau^2. \end{aligned} \tag{9}$$

Hence,

$$\begin{aligned} \mathbb{E}\|\tilde{g}^k - \nabla f(x^k)\|^2 &\leq \frac{4L}{n^2}\sum_{\tau=1}^n \mathbb{E}[f^{(\tau)}(x^k) - f^{(\tau)}(x^*) - \langle \nabla f^{(\tau)}(x^*), x^k - x^*\rangle] + \frac{2}{n^2}\sum_{\tau=1}^n \sigma_\tau^2 \\ &\leq \frac{4L}{n}\mathbb{E}[f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle] + \frac{2}{n}\sigma^2. \end{aligned}$$

Since x^* is an optimal solution, we have $-\nabla f(x^*) \in \partial\psi(x^*)$, which implies that

$$f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle \leq P(x^k) - P(x^*). \tag{10}$$

Thus,

$$\mathbb{E}\|\tilde{g}^k - \nabla f(x^k)\|^2 \leq \frac{4L}{n}\mathbb{E}[P(x^k) - P(x^*)] + \frac{2}{n}\sigma^2.$$

E.3. Proof of Lemma 7

(i) First, we have

$$\begin{aligned}
& \mathbb{E}[\|e^{k+1}\|^2] \\
= \;& \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^{n} e_\tau^{k+1}\right\|^2 \\
\leq \;& \frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq \;& \frac{1}{n}\sum_{\tau=1}^{n}(1-\delta)\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2 \\
= \;& \frac{1-\delta}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma(\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)) + \gamma(\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k))\|^2 \\
= \;& \frac{1-\delta}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma(\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k))\|^2 + \frac{1-\delta}{n}\sum_{\tau=1}^{n}\gamma^2\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k)\|^2 \\
\leq \;& \frac{1}{n}(1-\delta)(1+\beta)\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 + \frac{1}{n}(1-\delta)(1+\frac{1}{\beta})\gamma^2\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \\
& + \frac{1}{n}\gamma^2\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k)\|^2 \\
\leq \;& \frac{1}{n}(1-\frac{\delta}{2})\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 + \frac{2(1-\delta)}{n\delta}\gamma^2\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 + \frac{1}{n}\gamma^2\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k)\|^2,
\end{aligned}$$

where we choose $\beta = \frac{\delta}{2(1-\delta)}$.

For $\mathbb{E}\|\nabla f(x^k) - \nabla f(w^k)\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 &= \mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^*) + \nabla f^{(\tau)}(x^*) - \nabla f^{(\tau)}(w^k)\|^2 \\
&\leq 2\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^*)\|^2 + 2\mathbb{E}\|\nabla f^{(\tau)}(w^k) - \nabla f^{(\tau)}(x^*)\|^2 \\
&\leq 4L\mathbb{E}[f^{(\tau)}(x^k) - f^{(\tau)}(x^*) - \langle\nabla f^{(\tau)}(x^*), x^k - x^*\rangle] \\
&\quad + 4L\mathbb{E}[f^{(\tau)}(w^k) - f^{(\tau)}(x^*) - \langle\nabla f^{(\tau)}(x^*), w^k - x^*\rangle]. \qquad (11)
\end{aligned}$$

Thus, combining (9), we can obtain

$$\begin{aligned}
& \mathbb{E}[\|e^{k+1}\|^2] \leq \frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq \;& \left(\frac{2}{\delta} - 2\right)\gamma^2\left(4L\mathbb{E}[f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle] + 4L\mathbb{E}[f(w^k) - f(x^*) - \langle\nabla f(x^*), w^k - x^*\rangle]\right) \\
& + \frac{1}{n}(1-\frac{\delta}{2})\sum_{\tau=1}^{n}\mathbb{E}\|e^k\|^2 + 4L\gamma^2\mathbb{E}[f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle] + 2\gamma^2\sigma^2 \\
\overset{(10)}{\leq} \;& (1-\frac{\delta}{2})\cdot\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 + \frac{8L\gamma^2}{\delta}\left(\mathbb{E}[P(x^k) - P(x^*)] + \mathbb{E}[P(w^k) - P(x^*)]\right) + 2\gamma^2\sigma^2 \qquad (12) \\
\leq \;& \frac{8L}{\delta}\sum_{i=0}^{k}(1-\frac{\delta}{2})^{k-i}\gamma^2\left(\mathbb{E}[P(x^i) - P(x^*)] + \mathbb{E}[P(w^i) - P(x^*)]\right) + 2\sigma^2\sum_{i=0}^{k}(1-\frac{\delta}{2})^{k-i}\gamma^2 \\
\leq \;& \frac{8L\gamma^2}{\delta}\sum_{i=0}^{k}(1-\frac{\delta}{2})^{k-i}\left(\mathbb{E}[P(x^i) - P(x^*)] + \mathbb{E}[P(w^i) - P(x^*)]\right) + \frac{4}{\delta}\sigma^2\gamma^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}[\|e^{j+1}\|^2] \\
\leq\quad & \frac{8L\gamma^2}{\delta}\sum_{j=0}^{k}\sum_{i=0}^{j}(1-\frac{\delta}{2})^{j-i}\Big(\mathbb{E}[P(x^i)-P(x^*)]+\mathbb{E}[P(w^i)-P(x^*)]\Big)+\frac{4}{\delta}\sigma^2\gamma^2(k+1) \\
\leq\quad & \frac{8L\gamma^2}{\delta}\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]+\mathbb{E}[P(w^j)-P(x^*)]\Big)\sum_{i=0}^{+\infty}(1-\frac{\delta}{2})^{i}+\frac{4}{\delta}\sigma^2\gamma^2(k+1) \\
\leq\quad & \frac{16L\gamma^2}{\delta^2}\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]+\mathbb{E}[P(w^j)-P(x^*)]\Big)+\frac{4}{\delta}\sigma^2\gamma^2(k+1). \qquad (13)
\end{aligned}$$

If $p=0$, then $w^j=w^0$ for $j\geq 0$. Hence,

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}[\|e^{j+1}\|^2] \\
\leq\quad & \frac{16L\gamma^2}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)]+\frac{4\gamma^2(k+1)}{\delta}\left(\sigma^2+\frac{4L}{\delta}(P(w^0)-P(x^*))\right).
\end{aligned}$$

If $p>0$, then

$$\mathbb{E}[P(w^{k+1})-P(x^*)]=p\mathbb{E}[P(x^k)-P(x^*)]+(1-p)\mathbb{E}[P(w^k)-P(x^*)],$$

which indicates that

$$\sum_{j=0}^{k}\mathbb{E}[P(w^{j+1})-P(x^*)]=p\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)]+(1-p)\sum_{j=0}^{k}\mathbb{E}[P(w^j)-P(x^*)].$$

On the other hand,

$$\begin{aligned}
\sum_{j=0}^{k}\mathbb{E}[P(w^{j+1})-P(x^*)] \quad &= \quad \sum_{j=0}^{k}\mathbb{E}[P(w^j)-P(x^*)]+\mathbb{E}[P(w^{k+1})-P(x^*)]-\big(P(w^0)-P(x^*)\big) \\
&\geq \quad \sum_{j=0}^{k}\mathbb{E}[P(w^j)-P(x^*)]-\big(P(w^0)-P(x^*)\big).
\end{aligned}$$

Thus, we arrive at

$$\sum_{j=0}^{k}\mathbb{E}[P(w^j)-P(x^*)]\leq\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)]+\frac{1}{p}\big(P(w^0)-P(x^*)\big). \qquad (14)$$

Combining (13) and (14), we have

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}[\|e^{j+1}\|^2] \\
\leq\quad & \frac{32L\gamma^2}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)]+\frac{16L\gamma^2}{\delta^2 p}\big(P(w^0)-P(x^*)\big)+\frac{4}{\delta}\sigma^2\gamma^2(k+1).
\end{aligned}$$

(ii) Under Assumption 2.1, we have $\mathbb{E}[Q(x)] = \delta x$, and

$$\begin{aligned}
\mathbb{E}\|e^{k+1}\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^{n} e_\tau^{k+1}\right\|^2 \\
&= \frac{1}{n^2}\sum_{i,j}\mathbb{E}\langle e_i^{k+1}, e_j^{k+1}\rangle \\
&= \frac{1}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2 + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\langle e_i^{k+1}, e_j^{k+1}\rangle \\
&\leq \frac{1-\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2 + \frac{(1-\delta)^2}{n^2}\sum_{i\neq j}\mathbb{E}\langle e_i^k + \gamma g_i^k, e_j^k + g_j^k\rangle \\
&= \frac{(1-\delta)^2}{n^2}\mathbb{E}\left\|\sum_{\tau=1}^{n}(e_\tau^k + \gamma g_\tau^k)\right\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2 \\
&\leq (1-\delta)\mathbb{E}\|e^k + \gamma g^k\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2.
\end{aligned}$$

Under Assumption 2.2, we have

$$\begin{aligned}
\mathbb{E}\|e^{k+1}\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^{n} e_\tau^{k+1}\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^{n}\left(e_\tau^k + \gamma g_\tau^k - Q(\gamma g_\tau^k + e_\tau^k)\right)\right\|^2 \\
&\overset{Assumption 2.2}{\leq} (1-\delta')\mathbb{E}\|e^k + \gamma g^k\|^2 \\
&\leq (1-\delta)\mathbb{E}\|e^k + \gamma g^k\|^2.
\end{aligned}$$

Overall, under Assumption 2.1 or Assumption 2.2, we have

$$\mathbb{E}\|e^{k+1}\|^2 \leq (1-\delta)\mathbb{E}\|e^k + \gamma g^k\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2.$$

For $\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2$, we have

$$\begin{aligned}
&\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma g_\tau^k\|^2 \\
= &\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma(\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)) + \gamma(\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k))\|^2 \\
= &\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k + \gamma(\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k))\|^2 + \sum_{\tau=1}^{n}\gamma^2\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k)\|^2 \\
\leq &(1+\delta)\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 + (1+\frac{1}{\delta})\gamma^2\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 + \sum_{\tau=1}^{n}\gamma^2\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(x^k)\|^2 \\
\leq &(1+\delta)\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 + 4nL(1+\frac{1}{\delta})\gamma^2\left(\mathbb{E}[P(x^k) - P(x^*)] + \mathbb{E}[P(w^k) - P(x^*)]\right) \\
&+4nL\gamma^2\mathbb{E}[P(x^k) - P(x^*)] + 2n\gamma^2\sigma^2,
\end{aligned}$$

where the last inequality comes from (9), (10), and (11).

For $(1-\delta)\mathbb{E}\|e^k+\gamma g^k\|^2$, we have

$$
\begin{aligned}
& (1-\delta)\mathbb{E}\|e^k+\gamma g^k\|^2 \\
= \quad & (1-\delta)\mathbb{E}\|e^k+\gamma\tilde{g}^k-\gamma\nabla f(w^k)\|^2 \\
= \quad & (1-\delta)\mathbb{E}\|e^k+\gamma\nabla f(x^k)-\gamma\nabla f(w^k)+\gamma\tilde{g}^k-\gamma\nabla f(x^k)\|^2 \\
= \quad & (1-\delta)\mathbb{E}\|e^k+\gamma\nabla f(x^k)-\gamma\nabla f(w^k)\|^2+\gamma^2\mathbb{E}\|\tilde{g}^k-\nabla f(x^k)\|^2 \\
\leq \quad & (1-\delta)(1+\frac{\delta}{2(1-\delta)})\mathbb{E}\|e^k\|^2+(1-\delta)(1+\frac{2(1-\delta)}{\delta})\gamma^2\mathbb{E}\|\nabla f(x^k)-\nabla f(w^k)\|^2+\gamma^2\mathbb{E}\|\tilde{g}^k-\nabla f(x^k)\|^2 \\
\overset{(7)}{\leq} \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+(\frac{2}{\delta}-2)\gamma^2\mathbb{E}\|\nabla f(x^k)-\nabla f(w^k)\|^2+\frac{4L}{n}\gamma^2\mathbb{E}[P(x^k)-P(x^*)]+\frac{2}{n}\gamma^2\sigma^2 \\
\overset{(10)}{\leq} \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+(\frac{2}{\delta}-2)4\gamma^2L\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right) \\
& +\frac{4L}{n}\gamma^2\mathbb{E}[P(x^k)-P(x^*)]+\frac{2}{n}\gamma^2\sigma^2 \\
\leq \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+\frac{8L\gamma^2}{\delta}\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right)+\frac{2}{n}\gamma^2\sigma^2.
\end{aligned}
$$

Thus, we arrive at

$$
\begin{aligned}
\mathbb{E}\|e^{k+1}\|^2 \quad \leq \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+\frac{8L\gamma^2}{\delta}\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right)+\frac{2}{n}\gamma^2\sigma^2 \\
& +\frac{\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{4L}{n}(1-\delta)^2\gamma^2\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right) \\
& +\frac{4L}{n}(1-\delta)\delta\gamma^2\mathbb{E}[P(x^k)-P(x^*)]+\frac{2(1-\delta)\delta}{n}\gamma^2\sigma^2 \\
\leq \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+\frac{\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{4}{n}\gamma^2\sigma^2 \\
& +\left(\frac{8L\gamma^2}{\delta}+\frac{4L\gamma^2}{n}\right)\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right).
\end{aligned}
$$

From (12), we have

$$
\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2\leq(1-\frac{\delta}{2})\cdot\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{8L\gamma^2}{\delta}\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right)+2\gamma^2\sigma^2.
$$

Thereofore, we have

$$
\begin{aligned}
& \mathbb{E}\|e^{k+1}\|^2+\frac{4}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq \quad & (1-\frac{\delta}{2})\mathbb{E}\|e^k\|^2+\left(\frac{\delta}{n^2}+(1-\frac{\delta}{2})\frac{4}{n^2}\right)\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{12}{n}\gamma^2\sigma^2 \\
& +\frac{L\gamma^2}{\delta}(8+\frac{36}{n})\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right) \\
\leq \quad & (1-\frac{\delta}{4})\left(\mathbb{E}\|e^k\|^2+\frac{4}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2\right)+\frac{L\gamma^2}{\delta}(8+\frac{36}{n})\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right)+\frac{12}{n}\gamma^2\sigma^2 \\
\leq \quad & \frac{L\gamma^2}{\delta}(8+\frac{36}{n})\sum_{i=0}^{k}(1-\frac{\delta}{4})^{k-i}\left(\mathbb{E}[P(x^i)-P(x^*)]+\mathbb{E}[P(w^i)-P(x^*)]\right)+\frac{12\sigma^2}{n}\sum_{i=0}^{k}(1-\frac{\delta}{4})^{k-i}\gamma^2 \\
\leq \quad & \frac{L\gamma^2}{\delta}(8+\frac{36}{n})\sum_{i=0}^{k}(1-\frac{\delta}{4})^{k-i}\left(\mathbb{E}[P(x^i)-P(x^*)]+\mathbb{E}[P(w^i)-P(x^*)]\right)+\frac{48}{n\delta}\sigma^2\gamma^2,
\end{aligned}
$$

which implies that

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}\|e^{j+1}\|^2\\
\leq\ & \frac{L\gamma^2}{\delta}(8+\frac{36}{n})\sum_{j=0}^{k}\sum_{i=0}^{j}(1-\frac{\delta}{4})^{j-i}\Big(\mathbb{E}[P(x^i)-P(x^*)]+\mathbb{E}[P(w^i)-P(x^*)]\Big)+\frac{48}{n\delta}\sigma^2\gamma^2(k+1)\\
\leq\ & \frac{L\gamma^2}{\delta}(8+\frac{36}{n})\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]+\mathbb{E}[P(w^j)-P(x^*)]\Big)\sum_{i=0}^{+\infty}(1-\frac{\delta}{4})^{i}+\frac{48}{n\delta}\sigma^2\gamma^2(k+1)\\
\leq\ & \frac{4L\gamma^2}{\delta^2}(8+\frac{36}{n})\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]+\mathbb{E}[P(w^j)-P(x^*)]\Big)+\frac{48}{n\delta}\sigma^2\gamma^2(k+1).
\end{aligned}$$

If $p=0$, then $w^j=w^0$ for $j\geq 0$, which implies that

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}[\|e^{j+1}\|^2]\\
\leq\ & \frac{16L\gamma^2}{\delta^2}(2+\frac{9}{n})\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]\Big)+\frac{16\gamma^2(k+1)}{\delta}\left(\frac{3\sigma^2}{n}+(2+\frac{9}{n})\frac{L}{\delta}(P(w^0)-P(x^*))\right)
\end{aligned}$$

If $p>0$, then combining (14), we can obtain

$$\begin{aligned}
&\sum_{j=0}^{k}\mathbb{E}[\|e^{j+1}\|^2]\\
\leq\ & \frac{32L\gamma^2}{\delta^2}(2+\frac{9}{n})\sum_{j=0}^{k}\Big(\mathbb{E}[P(x^j)-P(x^*)]\Big)+\frac{16L\gamma^2}{\delta^2 p}(2+\frac{9}{n})\Big(P(w^0)-P(x^*)\Big)+\frac{48}{n\delta}\sigma^2\gamma^2(k+1).
\end{aligned}$$

Appendix F. Proof of Theorem 1

From the convexity of P, we have

$$\mathbb{E}[P(\bar{x}^k)-P(x^*)]\leq\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)],$$

for $\bar{x}^k=\frac{1}{k}\sum_{j=1}^{k}x^j$. Hence, we only need to estimate $\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)]$.

(i) From Lemma 5, we have

$$\begin{aligned}
2\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] \quad\leq\quad & \sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j-x^*\|^2}{\gamma}-\frac{\mathbb{E}\|\tilde{x}^{j+1}-x^*\|^2}{\gamma}\right)+2\sum_{j=0}^{k}\frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma}\\
&+4\gamma\sum_{j=0}^{k}\mathbb{E}\|\nabla f_{i_j}(x^j)-\nabla f(x^j)\|^2\\
\overset{(7)}{\leq}\quad & \sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j-x^*\|^2}{\gamma}-\frac{\mathbb{E}\|\tilde{x}^{j+1}-x^*\|^2}{\gamma}\right)+2\sum_{j=0}^{k}\frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma}\\
&+\frac{16L\gamma}{n}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)]+\frac{8\sigma^2}{n}\gamma(k+1).
\end{aligned}\tag{15}$$

If $p = 0$, then from Lemma 7, we can obtain

$$
\begin{aligned}
& 2\sum_{j=0}^{k} \frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma} + \frac{16L\gamma}{n}\sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{8\sigma^2}{n}\gamma(k+1) \\
\leq\ & (\frac{32}{\delta^2} + \frac{16}{n})L\gamma \sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{8\gamma(k+1)}{\delta}\left(\sigma^2 + \frac{\sigma^2\delta}{n} + \frac{4L}{\delta}(P(w^0) - P(x^*))\right) \\
\leq\ & \frac{48L\gamma}{\delta^2}\sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{16\gamma(k+1)}{\delta}\left(\sigma^2 + \frac{2L}{\delta}(P(w^0) - P(x^*))\right) \\
\leq\ & \frac{48L\gamma}{\delta^2}\sum_{j=0}^{k} \mathbb{E}[P(x^{j+1}) - P(x^*)] + \frac{16\gamma(k+1)}{\delta}\left(\sigma^2 + \frac{(2 + 3/(k+1))L}{\delta}(P(w^0) - P(x^*))\right) \\
\leq\ & \frac{48L\gamma}{\delta^2}\sum_{j=0}^{k} \mathbb{E}[P(x^{j+1}) - P(x^*)] + \frac{16\gamma(k+1)}{\delta}\left(\sigma^2 + \frac{5L}{\delta}(P(w^0) - P(x^*))\right).
\end{aligned}
$$

Thus, from (15), if $\gamma \leq \frac{\delta^2}{48L}$, we have

$$
\begin{aligned}
& \frac{1}{k+1}\sum_{j=0}^{k} \mathbb{E}[P(x^{j+1}) - P(x^*)] \\
\leq\ & \frac{1}{k+1}\sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j - x^*\|^2}{\gamma} - \frac{\mathbb{E}\|\tilde{x}^{j+1} - x^*\|^2}{\gamma} + \frac{16\gamma}{\delta}\left(\sigma^2 + \frac{5L}{\delta}(P(w^0) - P(x^*))\right)\right)
\end{aligned}
$$

Therefore, from Lemma 13 in [22], there exists a constant stepsize $\gamma \leq \frac{\delta^2}{48L}$ such that

$$
\frac{1}{k+1}\sum_{j=1}^{k+1} \mathbb{E}[P(x^j) - P(x^*)] \leq \frac{48L\|x^0 - x^*\|^2}{\delta^2(k+1)} + \frac{2\|x^0 - x^*\|\sqrt{16\sigma^2/\delta + 80L(P(w^0) - P(x^*))/\delta^2}}{\sqrt{k+1}}.
$$

If $p > 0$, then from Lemma 7, we have

$$
\begin{aligned}
& 2\sum_{j=0}^{k} \frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma} + \frac{16L\gamma}{n}\sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{8\sigma^2}{n}\gamma(k+1) \\
\leq\ & (\frac{64}{\delta^2} + \frac{16}{n})L\gamma \sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{32L\gamma}{\delta^2 p}(P(w^0) - P(x^*)) + (\frac{1}{n} + \frac{1}{\delta})8\sigma^2\gamma(k+1) \\
\leq\ & \frac{80L\gamma}{\delta^2}\sum_{j=0}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{32L\gamma}{\delta^2 p}(P(w^0) - P(x^*)) + \frac{16}{\delta}\sigma^2\gamma(k+1) \\
\leq\ & \frac{80L\gamma}{\delta^2}\sum_{j=0}^{k} \mathbb{E}[P(x^{j+1}) - P(x^*)] + \frac{112L\gamma}{\delta^2 p}(P(w^0) - P(x^*)) + \frac{16}{\delta}\sigma^2\gamma(k+1).
\end{aligned}
$$

Hence, from (15), if $\gamma \leq \frac{\delta^2}{80L}$, we have

$$
\begin{aligned}
& \frac{1}{k+1}\sum_{j=0}^{k} \mathbb{E}[P(x^{j+1}) - P(x^*)] \\
\leq\ & \frac{1}{k+1}\sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j - x^*\|^2}{\gamma} - \frac{\mathbb{E}\|\tilde{x}^{j+1} - x^*\|^2}{\gamma} + \frac{16\sigma^2\gamma}{\delta}\right) + \frac{2(P(w^0) - P(x^*))/p}{k+1}.
\end{aligned}
$$

Therefore, from Lemma 13 in [22], there exists a constant stepsize $\gamma \leq \frac{\delta^2}{80L}$ such that

$$
\frac{1}{k+1}\sum_{j=1}^{k+1} \mathbb{E}[P(x^j) - P(x^*)] \leq \frac{1}{k+1}\left(\frac{80L\|x^0 - x^*\|^2}{\delta^2} + \frac{2(P(w^0) - P(x^*))}{p}\right) + \frac{8\sigma\|x^0 - x^*\|}{\sqrt{\delta(k+1)}}.
$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$, then from Lemma 7 (ii), we can obtain

$$\begin{aligned}
& 2\sum_{j=0}^{k}\frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma} + \frac{16L\gamma}{n}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{8\sigma^2}{n}\gamma(k+1) \\
\leq \; & (\frac{64}{\delta^2}+\frac{288}{\delta^2 n}+\frac{16}{n})L\gamma\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{8\gamma(k+1)}{\delta}\left(\frac{12\sigma^2}{n}+\frac{\sigma^2\delta}{n}+\frac{L}{\delta}(8+\frac{36}{n})(P(w^0)-P(x^*))\right) \\
\leq \; & \frac{(64+304/n)L\gamma}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{8\gamma(k+1)}{\delta}\left(\frac{13\sigma^2}{n}+\frac{L}{\delta}(8+\frac{36}{n})(P(w^0)-P(x^*))\right) \\
\leq \; & \frac{(64+304/n)L\gamma}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] + \frac{8\gamma(k+1)}{\delta}\left(\frac{13\sigma^2}{n}+\frac{L}{\delta}(8+\frac{36}{n}+\frac{8+38/n}{k+1})(P(w^0)-P(x^*))\right) \\
\leq \; & \frac{(64+304/n)L\gamma}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] + \frac{8\gamma(k+1)}{\delta}\left(\frac{13\sigma^2}{n}+\frac{4L}{\delta}(4+\frac{19}{n})(P(w^0)-P(x^*))\right).
\end{aligned}$$

Thus, from (15), if $\gamma \leq \frac{\delta^2}{(64+304/n)L}$, we have

$$\begin{aligned}
& \frac{1}{k+1}\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] \\
\leq \; & \frac{1}{k+1}\sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j-x^*\|^2}{\gamma}-\frac{\mathbb{E}\|\tilde{x}^{j+1}-x^*\|^2}{\gamma}+\frac{8\gamma}{\delta}\left(\frac{13\sigma^2}{n}+\frac{4L}{\delta}(4+\frac{19}{n})(P(w^0)-P(x^*))\right)\right)
\end{aligned}$$

Therefore, from Lemma 13 in [22], there exists a constant stepsize $\gamma \leq \frac{\delta^2}{(64+304/n)L}$ such that

$$\frac{1}{k+1}\sum_{j=1}^{k+1}\mathbb{E}[P(x^j)-P(x^*)] \leq O\left(\frac{L\|x^0-x^*\|^2}{\delta^2(k+1)}+\frac{\|x^0-x^*\|\sqrt{\sigma^2/(n\delta)+L(P(w^0)-P(x^*))/\delta^2}}{\sqrt{k+1}}\right).$$

If $p > 0$, then from Lemma 7 (ii), we have

$$\begin{aligned}
& 2\sum_{j=0}^{k}\frac{\mathbb{E}\|e^{j+1}\|^2}{\gamma} + \frac{16L\gamma}{n}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{8\sigma^2}{n}\gamma(k+1) \\
\leq \; & (\frac{128}{\delta^2}+\frac{576}{\delta^2 n}+\frac{16}{n})L\gamma\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{L\gamma}{\delta^2 p}(64+\frac{288}{n})(P(w^0)-P(x^*)) + (\frac{1}{n}+\frac{12}{n\delta})8\sigma^2\gamma(k+1) \\
\leq \; & \frac{(128+592/n)L\gamma}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^j)-P(x^*)] + \frac{L\gamma}{\delta^2 p}(64+\frac{288}{n})(P(w^0)-P(x^*)) + \frac{13}{n\delta}\sigma^2\gamma(k+1) \\
\leq \; & \frac{(128+592/n)L\gamma}{\delta^2}\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] + \frac{L\gamma}{\delta^2 p}(192+\frac{880}{n})(P(w^0)-P(x^*)) + \frac{104}{n\delta}\sigma^2\gamma(k+1).
\end{aligned}$$

Hence, from (15), if $\gamma \leq \frac{\delta^2}{(128+592/n)L}$, we have

$$\begin{aligned}
& \frac{1}{k+1}\sum_{j=0}^{k}\mathbb{E}[P(x^{j+1})-P(x^*)] \\
\leq \; & \frac{1}{k+1}\sum_{j=0}^{k}\left(\frac{\mathbb{E}\|\tilde{x}^j-x^*\|^2}{\gamma}-\frac{\mathbb{E}\|\tilde{x}^{j+1}-x^*\|^2}{\gamma}+\frac{104\sigma^2\gamma}{n\delta}\right)+\frac{2(P(w^0)-P(x^*))/p}{k+1}.
\end{aligned}$$

Therefore, from Lemma 13 in [22], there exists a constant stepsize $\gamma \leq \frac{\delta^2}{(128+592/n)L}$ such that

$$\frac{1}{k+1}\sum_{j=1}^{k+1}\mathbb{E}[P(x^j)-P(x^*)] \leq O\left(\frac{1}{k+1}\left(\frac{L\|x^0-x^*\|^2}{\delta^2}+\frac{(P(w^0)-P(x^*))}{p}\right)+\frac{\sigma\|x^0-x^*\|}{\sqrt{n\delta(k+1)}}\right).$$

Appendix G. Lemmas 8, 9, and 10

Noticing that the iteration in the error compensated RDA starts from $k = 1$, let $\beta_0 = \beta_1$ and $x^0 = x^1$. We also define two conjugate-type functions as in [31] for $k \geq 0$:

$$U_k(s) = \max_{x \in \mathcal{F}_D} \{\langle s, x - x^0 \rangle - k\psi(x)\},$$

$$V_k(s) = \max_{x} \{\langle s, x - x^0 \rangle - k\psi(x) - \beta_k h(x)\},$$

where $\mathcal{F}_D = \{x \mid h(x) \leq D^2\}$.

Lemma 8 *For each $k \geq 1$, we have*

$$V_k(-\tilde{s}^k) + \psi(x^{k+1}) \leq V_{k-1}(-\tilde{s}^k) + \langle \bar{g}^k + \frac{\beta_k}{k}\partial h(x^{k+1}), \tilde{x}^{k+1} - x^{k+1} \rangle. \tag{16}$$

Lemma 9 *For $j \geq 1$ and any $\alpha > 0$, we have*

$$\begin{aligned}
& \langle \tilde{g}^j, x^j - x^0 \rangle + \psi(x^{j+1}) \\
\leq \quad & V_{j-1}(-\tilde{s}^{j-1}) - V_j(-\tilde{s}^j) + \frac{\alpha \|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{1}{2\alpha\beta_{j-1}}\|e^j\|^2 \\
& + \frac{\alpha \|\tilde{s}^j / j\|^2}{2\beta_j} + \frac{1}{2\alpha\beta_j}\|e^{j+1}\|^2 + \frac{\|e^{j+1}\|^2}{j\beta_j} + \frac{1}{2j}\|\partial h(x^{j+1})\|^2 + \frac{1}{2j}\|e^{j+1}\|^2. \qquad (17)
\end{aligned}$$

Lemma 10 *(i)For $k \geq 1$, we have*

$$\mathbb{E}\|e^k\|^2 \leq \frac{16}{\delta^2} G^2. \tag{18}$$

(ii) If $p = 0$, we have

$$\sum_{j=1}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \frac{16L}{\delta^2} \sum_{j=1}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{4k}{\delta}\left(\sigma^2 + \frac{4L}{\delta}(P(w^1) - P(x^*))\right).$$

If $p > 0$, we have

$$\sum_{j=1}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \frac{32L}{\delta^2} \sum_{j=1}^{k} \mathbb{E}[P(x^j) - P(x^*)] + \frac{16L}{\delta^2 p}\left(P(w^1) - P(x^*)\right) + \frac{4}{\delta}\sigma^2\gamma^2 k.$$

(iii) Under Assumption 2.1 or Assumption 2.2. If $p = 0$, we have

$$\sum_{j=1}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \tfrac{16L}{\delta^2}(2 + \tfrac{9}{n}) \sum_{j=1}^{k} \Big(\mathbb{E}[P(x^j) - P(x^*)]\Big) + \tfrac{16k}{\delta}\left(\tfrac{3\sigma^2}{n} + \tfrac{(2+9/n)L}{\delta}(P(w^1) - P(x^*))\right).$$

If $p > 0$, we have

$$\sum_{j=1}^{k} \mathbb{E}[\|e^{j+1}\|^2] \leq \tfrac{32L(2+\frac{9}{n})}{\delta^2} \sum_{j=1}^{k} \Big(\mathbb{E}[P(x^j) - P(x^*)]\Big) + \tfrac{16L(2+\frac{9}{n})}{\delta^2 p}\left(P(w^1) - P(x^*)\right) + \tfrac{48}{n\delta}\sigma^2 k.$$

Appendix H. Proofs of Lemmas 8, 9, and 10

H.1. Proof of Lemma 8

From Lemma 11 in [31], we have

$$V_k(-\tilde{s}^k) + \psi(\tilde{x}^{k+1}) \leq V_{k-1}(-\tilde{s}^k) + (\beta_{k-1} - \beta_k)h(\tilde{x}^{k+1}).$$

Since $\beta_0 = \beta_1$, we know $\{\beta_k\}_{k \geq 0}$ is a nondecreasing sequence. Hence, by assumption $h(x) \geq 0$, we have

$$V_k(-\tilde{s}^k) + \psi(\tilde{x}^{k+1}) \leq V_{k-1}(-\tilde{s}^k). \tag{19}$$

From the convexity of ψ, we have

$$\psi(\tilde{x}^{k+1}) \geq \psi(x^{k+1}) + \langle \partial\psi(x^{k+1}), \tilde{x}^{k+1} - x^{k+1} \rangle. \tag{20}$$

From the definition of x^{k+1}, we have

$$\partial\psi(x^{k+1}) = -\bar{g}^k - \frac{\beta_k}{k}\partial h(x^{k+1}). \tag{21}$$

Combining (20) and (21), we can obtain

$$\psi(x^{k+1}) \leq \psi(\tilde{x}^{k+1}) + \langle \bar{g}^k + \frac{\beta_k}{k}\partial h(x^{k+1}), \tilde{x}^{k+1} - x^{k+1} \rangle. \tag{22}$$

From (19) and (22), we can get the result.

H.2. Proof of Lemma 9

From (16) in Lemma 8, for any $j \geq 1$, we can obtain

$$\begin{aligned}
& V_j(-\tilde{s}^j) + \psi(x^{j+1}) \\
\leq \quad & V_{j-1}(-\tilde{s}^j) + \langle \bar{g}^j + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle \\
\leq \quad & V_{j-1}(-\tilde{s}^{j-1}) + \langle -\tilde{g}^j, \nabla V_{j-1}(-\tilde{s}^{j-1}) + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} \\
& + \langle \bar{g}^j + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle \\
= \quad & V_{j-1}(-\tilde{s}^{j-1}) + \langle -\tilde{g}^j, x^j - x^0 \rangle - \langle \tilde{g}^j, \nabla V_{j-1}(-\tilde{s}^{j-1}) - \nabla V_{j-1}(-s^{j-1}) \rangle + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} \\
& + \langle \bar{g}^j + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle \\
= \quad & V_{j-1}(-\tilde{s}^{j-1}) + \langle -\tilde{g}^j, x^j - x^0 \rangle - \langle \tilde{g}^j, \tilde{x}^j - x^j \rangle + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} \\
& + \langle \bar{g}^j + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle,
\end{aligned} \tag{23}$$

where the second inequality comes from (48) in [31] and (4), the last equality comes from (47) in [31]. From Lemma 10 in [31], we have

$$\|\tilde{x}^k - x^k\|^2 = \|\nabla V_{k-1}(-\tilde{s}^{k-1}) - \nabla V_{k-1}(-s^{k-1})\|^2 \leq \frac{1}{\beta_{k-1}^2}\|e^k\|^2,$$

for $k \geq 2$. Moreover, $\tilde{x}^1 = x^1$ and $e^1 = 0$. Hence

$$\|\tilde{x}^k - x^k\|^2 \leq \frac{1}{\beta_{k-1}^2}\|e^k\|^2, \tag{24}$$

for $k \geq 1$. Then we can get

$$\begin{aligned}
& \langle -\tilde{g}^j, \tilde{x}^j - x^j \rangle + \langle \bar{g}^j + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle \\
\leq \quad & \frac{\alpha\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\beta_{j-1}}{2\alpha}\|\tilde{x}^j - x^j\|^2 + \langle \frac{\tilde{s}^j - e^{j+1}}{j} + \frac{\beta_j}{j}\partial h(x^{j+1}), \tilde{x}^{j+1} - x^{j+1} \rangle \\
\leq \quad & \frac{\alpha\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\beta_{j-1}}{2\alpha}\|\tilde{x}^j - x^j\|^2 + \frac{\alpha\|\tilde{s}^j/j\|^2}{2\beta_j} + \frac{\beta_j}{2\alpha}\|\tilde{x}^{j+1} - x^{j+1}\|^2 \\
& + \frac{\|e^{j+1}\|^2}{2j\beta_j} + \frac{\beta_j}{2j}\|\tilde{x}^{j+1} - x^{j+1}\|^2 + \frac{1}{2j}\|\partial h(x^{j+1})\|^2 + \frac{\beta_j^2}{2j}\|\tilde{x}^{j+1} - x^{j+1}\|^2 \\
\overset{(24)}{\leq} \quad & \frac{\alpha\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{1}{2\alpha\beta_{j-1}}\|e^j\|^2 + \frac{\alpha\|\tilde{s}^j/j\|^2}{2\beta_j} + \frac{1}{2\alpha\beta_j}\|e^{j+1}\|^2 \\
& + \frac{\|e^{j+1}\|^2}{j\beta_j} + \frac{1}{2j}\|\partial h(x^{j+1})\|^2 + \frac{1}{2j}\|e^{j+1}\|^2.
\end{aligned}$$

Combining the above inequality and (23), we can obtain

$$\begin{aligned} & \langle \tilde{g}^j, x^j - x^0 \rangle + \psi(x^{j+1}) \\ \leq \quad & V_{j-1}(-\tilde{s}^{j-1}) - V_j(-\tilde{s}^j) + \frac{\alpha\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{1}{2\alpha\beta_{j-1}}\|e^j\|^2 \\ & + \frac{\alpha\|\tilde{s}^j/j\|^2}{2\beta_j} + \frac{1}{2\alpha\beta_j}\|e^{j+1}\|^2 + \frac{\|e^{j+1}\|^2}{j\beta_j} + \frac{1}{2j}\|\partial h(x^{j+1})\|^2 + \frac{1}{2j}\|e^{j+1}\|^2. \end{aligned}$$

H.3. Proof of Lemma 10

Since $e^1 = 0$, (18) holds for $k = 1$. For $k \geq 2$, we have

$$\begin{aligned} \mathbb{E}\|e^k\|^2 \quad & \leq \quad \frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 \\ & = \quad \frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k-1} + g_\tau^{k-1} - y_\tau^{k-1}\|^2 \\ & \leq \quad \frac{1-\delta}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k-1} + g_\tau^{k-1}\|^2 \\ & \leq \quad \frac{(1-\delta)}{n}\sum_{\tau=1}^{n}(1+\beta)\mathbb{E}\|e_\tau^{k-1}\|^2 + \frac{(1-\delta)}{n}(1+\frac{1}{\beta})\sum_{\tau=1}^{n}\mathbb{E}\|g_\tau^{k-1}\|^2 \\ & \leq \quad (1-\frac{\delta}{2})\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k-1}\|^2 + \frac{2}{\delta}\cdot\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|g_\tau^{k-1}\|^2, \end{aligned}$$

where we choose $\beta = \frac{\delta}{2(1-\delta)}$. For $\mathbb{E}\|g_\tau^{k-1}\|^2$, we have

$$\mathbb{E}\|g_\tau^{k-1}\|^2 = \mathbb{E}\|\nabla f_{i_{k-1}^\tau}^{(\tau)}(x^{k-1}) - \nabla f^{(\tau)}(w^{k-1})\|^2 \leq 4G^2.$$

Therefore,

$$\begin{aligned} \mathbb{E}\|e^k\|^2 \quad & \leq \quad \frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 \\ & \leq \quad (1-\frac{\delta}{2})\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k-1}\|^2 + \frac{8}{\delta}G^2 \\ & \leq \quad \frac{8}{\delta}\sum_{j=0}^{k-1}(1-\frac{\delta}{2})^{k-1-j}G^2 \\ & \leq \quad \frac{16}{\delta^2}G^2. \end{aligned}$$

For the rest results, the proof is the same as that of Lemma 7.

Appendix I. Proof of Theorem 3

From the convexity of P, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j) - P(x^*)],$$

for $\bar{x}^k = \frac{1}{k}\sum_{j=1}^{k}x^j$. Hence, we only need to estimate $\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j) - P(x^*)]$.

First, we define the regret $R_k(x)$ similar to that in [31] :

$$R_k(x) = \sum_{j=1}^{k}\sum_{\tau=1}^{n}\frac{1}{n}\left(f_{i_j^\tau}^{(\tau)}(x^j) + \psi(x^j)\right) - \sum_{j=1}^{k}\sum_{\tau=1}^{n}\frac{1}{n}\left(f_{i_j^\tau}^{(\tau)}(x) + \psi(x)\right).$$

From (4) and $\tilde{s}^0 = 0$, we have $\tilde{s}^k = \sum_{j=1}^k \tilde{g}^j = \sum_{j=1}^k \sum_{\tau=1}^n \frac{1}{n} \nabla f_{i_j^\tau}^{(\tau)}(x^j)$. Then similar to (53), (54) and (55) in [31], we have

$$R_k(x) \leq \sum_{j=1}^k \Big(\langle \tilde{g}^j, x^j - x^0 \rangle + \psi(x^j) \Big) + V_k(-\tilde{s}^k) + \beta_k D^2. \tag{25}$$

Summing the inequality (17) for $j = 1, ..., k$, and noting $V_0(-\tilde{s}^0) = V_0(0) = 0$, we arrive at

$$\begin{aligned}
& \sum_{j=1}^k \Big(\langle \tilde{g}^j, x^j - x^0 \rangle + \psi(x^{j+1}) \Big) + V_k(-\tilde{s}^k) \\
\leq \quad & \sum_{j=1}^k \left(\frac{\alpha \|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\|\tilde{g}^j\|^2}{2\beta_{j-1}} + \frac{\alpha \|\tilde{s}^j / j\|^2}{2\beta_j} + \frac{1}{2\alpha\beta_{j-1}} \|e^j\|^2 + \frac{1}{2\alpha\beta_j} \|e^{j+1}\|^2 \right) \\
& + \sum_{j=1}^k \left(\frac{\|e^{j+1}\|^2}{j\beta_j} + \frac{1}{2j} \|\partial h(x^{j+1})\|^2 + \frac{1}{2j} \|e^{j+1}\|^2 \right) \\
\overset{(18)}{\leq} \quad & \sum_{j=1}^k \left(\frac{(\alpha+1)G^2}{2\beta_{j-1}} + \frac{\alpha G^2}{2\beta_j} + \frac{1}{2\alpha\beta_{j-1}} \|e^j\|^2 + \frac{1}{2\alpha\beta_j} \|e^{j+1}\|^2 \right) \\
& + \sum_{j=1}^k \left(\frac{16G^2}{\delta^2 j \beta_j} + \frac{1}{2j} H^2 + \frac{8}{\delta^2 j} G^2 \right).
\end{aligned}$$

By adding the nonpositive quantity $\psi(x^1) - \psi(x^{k+1})$ to the left-hand side of the above inequality and combining (25), we can obtain

$$\begin{aligned}
R_k(x) \quad \leq \quad & \beta_k D^2 + \sum_{j=1}^k \left(\frac{\alpha G^2}{2\beta_{j-1}} + \frac{G^2}{2\beta_{j-1}} + \frac{\alpha G^2}{2\beta_j} + \frac{1}{2\alpha\beta_{j-1}} \|e^j\|^2 + \frac{1}{2\alpha\beta_j} \|e^{j+1}\|^2 \right) \\
& + \sum_{j=1}^k \left(\frac{16G^2}{\delta^2 j \beta_j} + \frac{1}{2j} H^2 + \frac{8}{\delta^2 j} G^2 \right).
\end{aligned}$$

By choosing $x = x^*$, $\beta_j = \beta$, and taking expectation in the above inequality, we can get

$$\begin{aligned}
& \sum_{j=1}^k \mathbb{E}[P(x^j) - P(x^*)] \\
\leq \quad & \beta D^2 + \sum_{j=1}^k \left(\frac{\alpha G^2}{2\beta} + \frac{G^2}{2\beta} + \frac{\alpha G^2}{2\beta} + \frac{1}{2\alpha\beta} \mathbb{E}\|e^j\|^2 + \frac{1}{2\alpha\beta} \mathbb{E}\|e^{j+1}\|^2 \right) + \sum_{j=1}^k \left(\frac{16G^2}{\delta^2 j \beta} + \frac{1}{2j} H^2 + \frac{8}{\delta^2 j} G^2 \right) \\
\leq \quad & \beta D^2 + \frac{(2\alpha+1)k}{2\beta} G^2 + \frac{1}{\alpha\beta} \sum_{j=1}^k \mathbb{E}\|e^{j+1}\|^2 + \left(\frac{16G^2}{\beta\delta^2} + \frac{H^2}{2} + \frac{8G^2}{\delta^2} \right) \ln k.
\end{aligned} \tag{26}$$

(i) If $p = 0$, from Lemma 10 and (26), and letting $\alpha = \frac{4}{\delta}$, we have

$$\begin{aligned}
& \left(1 - \frac{4L}{\beta\delta}\right) \sum_{j=1}^k \mathbb{E}[P(x^j) - P(x^*)] \\
\leq \quad & \beta D^2 + \frac{(2\alpha+1)k}{2\beta} G^2 + \frac{4k}{\alpha\beta\delta} \left(\sigma^2 + \frac{4L}{\delta} (P(w^1) - P(x^*)) \right) + \left(\frac{16G^2}{\beta\delta^2} + \frac{H^2}{2} + \frac{8G^2}{\delta^2} \right) \ln k \\
\leq \quad & \beta D^2 + \frac{5k}{\beta\delta} \left(G^2 + L(P(w^1) - P(x^*)) + \delta\sigma^2/4 \right) + \left(\frac{16G^2}{\beta\delta^2} + \frac{H^2}{2} + \frac{8G^2}{\delta^2} \right) \ln k.
\end{aligned}$$

Let $\beta = 4\sqrt{\frac{k}{\delta}} \frac{\sqrt{G^2 + L(P(w^1) - P(x^*)) + \delta\sigma^2/4}}{D}$. Then we have $\frac{4L}{\beta\delta} \leq \frac{1}{2}$, if

$$k \geq \frac{4L^2 D^2}{\delta(G^2 + L(P(w^1) - P(x^*)) + \delta\sigma^2/4)}.$$

Thus

$$\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \leq \frac{12D}{\sqrt{\delta k}}\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2/4}+\left(\frac{8DG}{\delta\sqrt{\delta k}}+H^2+\frac{16G^2}{\delta^2}\right)\frac{\ln k}{k}.$$

If $p>0$, from Lemma 10 and (26), and letting $\alpha=\frac{8}{\sqrt{\delta}}$, we have

$$\begin{aligned}
&\left(1-\frac{4L}{\beta\delta^{\frac{3}{2}}}\right)\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \\
\leq\quad & \beta D^2+\frac{(2\alpha+1)k}{2\beta}G^2+\frac{16L}{\alpha\beta\delta^2 p}\left(P(w^1)-P(x^*)\right)+\frac{4\sigma^2 k}{\alpha\beta\delta}+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k \\
\leq\quad & \beta D^2+\frac{k}{2\beta\sqrt{\delta}}\left(\sigma^2+24G^2\right)+\frac{2L}{\beta\delta^{\frac{3}{2}}p}\left(P(w^1)-P(x^*)\right)+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k.
\end{aligned}$$

Let $\beta=\frac{4\sqrt{k}}{\delta^{\frac{1}{4}}}\frac{\sqrt{\sigma^2+24G^2}}{D}$. Then we have $\frac{4L}{\beta\delta^{\frac{3}{2}}}\leq\frac{1}{2}$, if

$$k\geq\frac{4L^2D^2}{\delta^{\frac{5}{2}}(\sigma^2+24G^2)}.$$

Thus

$$\begin{aligned}
\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \quad\leq\quad & \frac{10D}{\delta^{\frac{1}{4}}\sqrt{k}}\sqrt{\sigma^2+24G^2}+\frac{LD}{k\sqrt{k}\delta^{\frac{5}{4}}p\sqrt{\sigma^2+16G^2}}\left(P(w^1)-P(x^*)\right) \\
&+\left(\frac{2DG}{\sqrt{k}\delta^{\frac{7}{4}}}+H^2+\frac{16G^2}{\delta^2}\right)\frac{\ln k}{k}.
\end{aligned}$$

(ii) Under Assumption 2.1 or Assumption 2.2. If $p=0$, from Lemma 10 and (26), and letting $\alpha=\frac{4}{\delta}$, we have

$$\begin{aligned}
&\left(1-\frac{4L(2+9/n)}{\beta\delta}\right)\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \\
\leq\quad & \beta D^2+\frac{(2\alpha+1)k}{2\beta}G^2+\frac{16k}{\alpha\beta\delta}\left(\frac{3\sigma^2}{n}+\frac{(2+9/n)L}{\delta}(P(w^1)-P(x^*))\right)+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k \\
\leq\quad & \beta D^2+\frac{5k}{\beta\delta}\left(G^2+(2+\frac{9}{n})L(P(w^1)-P(x^*))+3\delta\sigma^2/n\right)+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k.
\end{aligned}$$

Let $\beta=4\sqrt{\frac{k}{\delta}}\frac{\sqrt{G^2+(2+9/n)L(P(w^1)-P(x^*))+3\delta\sigma^2/n}}{D}$. Then we have $\frac{4L(2+9/n)}{\beta\delta}\leq\frac{1}{2}$, if

$$k\geq\frac{4(2+9/n)^2L^2D^2}{\delta(G^2+(2+9/n)L(P(w^1)-P(x^*))+3\delta\sigma^2/n)}.$$

Thus

$$\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)]\leq O\left(\frac{D}{\sqrt{\delta k}}\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2/n}+\left(\frac{DG}{\delta\sqrt{\delta k}}+H^2+\frac{G^2}{\delta^2}\right)\frac{\ln k}{k}\right).$$

If $p>0$, from Lemma 10 and (26), and letting $\alpha=\frac{8}{\sqrt{n\delta}}$, we have

$$\begin{aligned}
&\left(1-\frac{4\sqrt{n}L(2+9/n)}{\beta\delta^{\frac{3}{2}}}\right)\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \\
\leq\quad & \beta D^2+\frac{(2\alpha+1)k}{2\beta}G^2+\frac{16L(2+9/n)}{\alpha\beta\delta^2 p}\left(P(w^1)-P(x^*)\right)+\frac{48\sigma^2 k}{n\alpha\beta\delta}+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k \\
\leq\quad & \beta D^2+\frac{k}{\beta\sqrt{n\delta}}\left(6\sigma^2+12G^2\right)+\frac{2\sqrt{n}L(2+9/n)}{\beta\delta^{\frac{3}{2}}p}\left(P(w^1)-P(x^*)\right)+\left(\frac{16G^2}{\beta\delta^2}+\frac{H^2}{2}+\frac{8G^2}{\delta^2}\right)\ln k,
\end{aligned}$$

where in the last inequality, we use $n\delta \leq 1$. Let $\beta = \frac{4\sqrt{k}}{(n\delta)^{\frac{1}{4}}}\frac{\sqrt{6\sigma^2+12G^2}}{D}$. Then we have $\frac{4\sqrt{n}L(2+9/n)}{\beta\delta^{\frac{3}{2}}} \leq \frac{1}{2}$, if

$$k \geq \frac{4L^2D^2n^{\frac{3}{2}}(2+9/n)^2}{\delta^{\frac{5}{2}}(6\sigma^2+12G^2)}.$$

Thus

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[P(x^j)-P(x^*)] \quad &\leq \quad O\left(\frac{D}{(n\delta)^{\frac{1}{4}}\sqrt{k}}\sqrt{\sigma^2+G^2}+\frac{n^{\frac{3}{4}}LD}{k\sqrt{k}\delta^{\frac{5}{4}}p\sqrt{\sigma^2+G^2}}\left(P(w^1)-P(x^*)\right)\right.\\ &\qquad \left.+\left(\frac{n^{\frac{1}{4}}DG}{\sqrt{k}\delta^{\frac{7}{4}}}+H^2+\frac{16G^2}{\delta^2}\right)\frac{\ln k}{k}\right)\end{aligned}$$

Appendix J. Other Datasets in Appendix

In the appendix, we add some experiments to supplement the effectiveness of our algorithm. Due to the consideration of computing power and dataset characteristics, we will add additional datasets here and have attached information about several datasets that will be used in the appendix. It should be noted that syn-128 and syn-256 are linear regression problems we obtain from the normal distribution sampling, so each dimension is equivalent.

Table 3: Dataset statistics

DATASET	SPARSE FORMAT	TASK	d	n	DENSITY
A5A	✓	LOGISTIC REGRESSION	122	6,414	13%
MUSHROOMS		LOGISTIC REGRESSION	119	8,124	19%
SYN-k		LASSO REGRESSION	k	10,000	100%

Appendix K. General Convex Case Experiment (L_1 regularization)

Due to the need to compare with DIANA in the main part, we adopted the penalty of L_1-L_2. In fact, our algorithm supports all convex problems, so we did a logistic regression experiment with L_1 penalty here. The performance of the EC-Prox algorithm is not much different from the L_1-L_2 penalty.

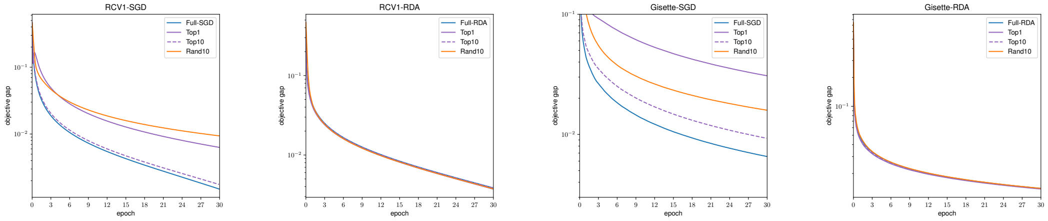


Figure 5: Error Compesated Proximal and full gradient SGD/RDA (General Convex)

Appendix L. Dense Dataset

We also verified the performance of EC-Prox with full dense dataset (linear regression, L_1 regularizer). As shown in Figure 6, TopK is still much more efficient than RandK. Moreover, even under such datasets, our algorithm can achieve a compression ratio of 20-200 times. Such potential is difficult to reach by quantization-based algorithms.

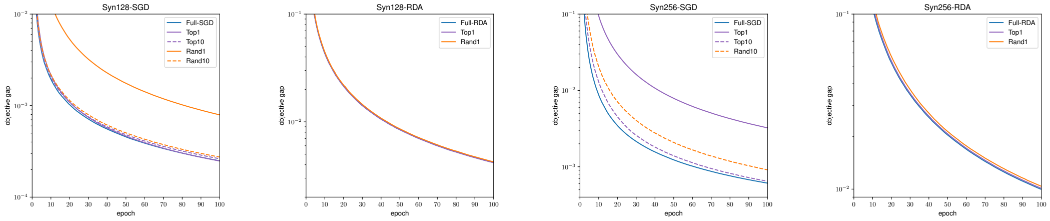
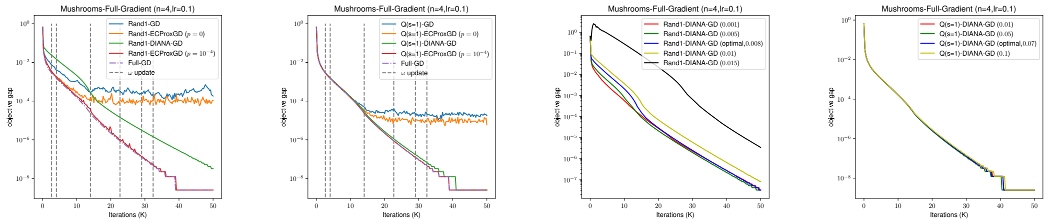


Figure 6: Error Compesated Proximal and full gradient SGD/RDA (Dense data, $\lambda_1 = 0.1$)

Appendix M. Further Comparison with DIANA

In the body part, we emphasize that EC-Prox shows a large advantage over quantization-based/RandK DIANA. On the other hand, the gap between DIANA and EC-Prox is slight in theory, so we hope to explore this difference further regardless of compressors. To make the variance of the gradient smaller, we use GD instead of SGD here for more stable results. In particular, for each node, we compute the gradient of coresponding objective rather than stochastic gradient.

We perform our algorithm on mushrooms dataset ($\lambda_1 = \lambda_2 = 10^{-3}$). Figure 7 shows that the EC-Prox algorithm with $p > \delta^2$ is slightly better than DIANA and even very close to non-compressed GD. Note that we only update w^k 6 times, which means the communication cost changes very little. In this setting, the EC-Prox with $p = 0$ and compressed-GD would not converge linearly, but EC-Prox is still better. From (c) and (d), unless too large, we notice that the α of DIANA is not sensitive, so we choose the theoretical optimal in other experiments.



(a) (b) (c) (d)

Figure 7: Error Compesated and DIANA Gradient Descent on mushrooms dataset ($\lambda_1 = \lambda_2 = 10^{-3}$)

In this problem, the optimal learning rate of GD – $1/L$ is about 0.4. In Figure 8, we compare DIANA and EC-Prox with large learning rate. As the learning rate increases, we can find that EC-Prox goes down harder with significant noise. In such a circumstance, we have to increase p to make the convergence of EC-Prox faster. In contrast, DIANA shows its strengths that makes the loss function converge smoothly with large learning rate. In fact, since δ is very small, theoretical learning rate of EC-Prox in our analysis is much smaller than $1/L$, this is also consistent with the instability of EC-Prox under large lr from the side. However, practically, we may not find the best learning rate of DIANA efficiently and larger p can easily improve the performance of EC-Prox. As a result, Figure 7 might be more common in real world.

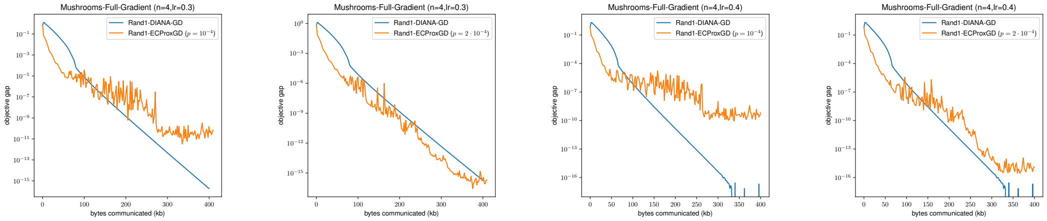


Figure 8: Error Compesated and DIANA Gradient Descent on mushrooms dataset (large lr)

Appendix N. Comparison of EC-Prox/DIANA for different regularizer weights

In this part, we compare our algorithm and DIANA for different regularizer weights. Specifically, we implement $L_1 - L_2$ regularized logistic regression on a5a and mushrooms datasets. From Figure 9, we can find that our algorithm outperforms DIANA significantly in most cases. In particular, when the regularizer weight is relative small, the gap is wider. In addition, when the regularizer is too large, the convergence of RDA will be very fast. This was also observed in our previous gistette experiment.

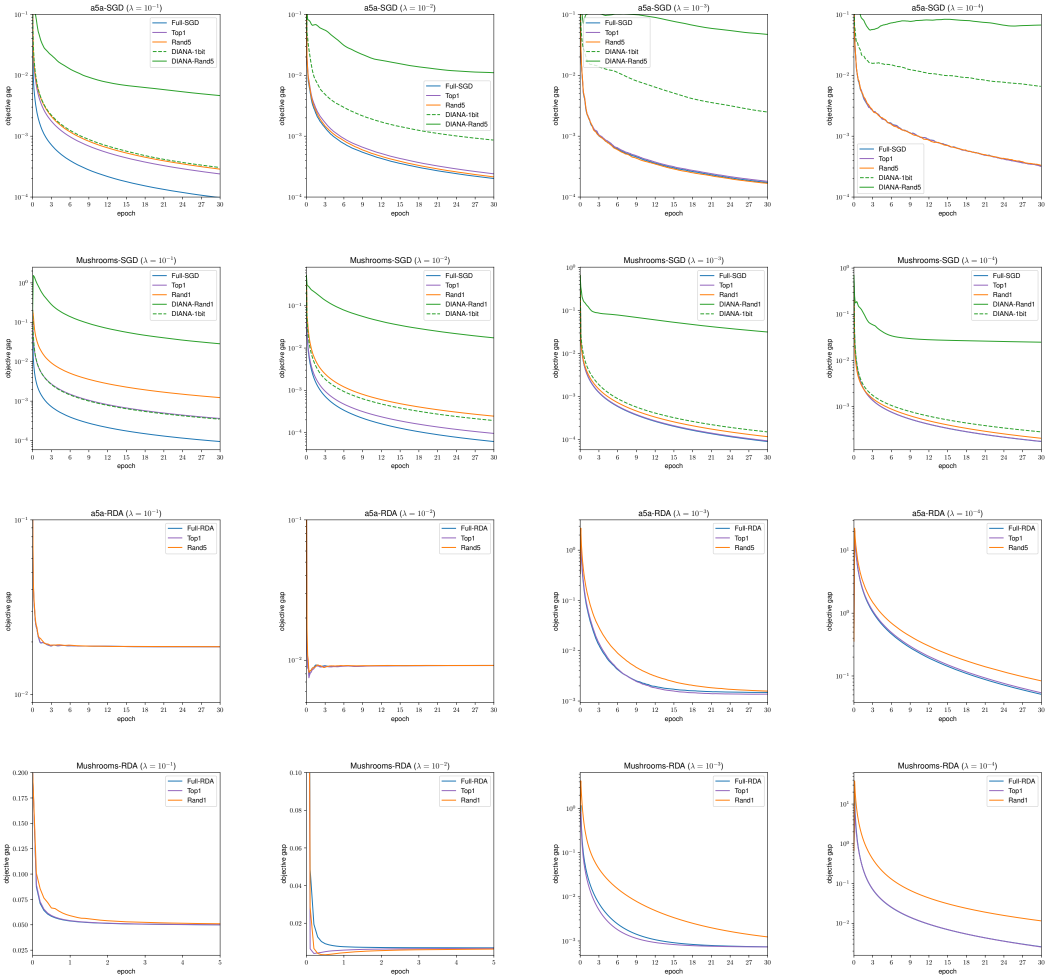


Figure 9: Error Compesated Proximal SGD/RDA and DIANA with different regularizers

Appendix O. Additional Distributed Experiments

Due to space limitations in the main part, we will show more results of distributed experiments here. As mentioned earlier, the RDA algorithm's initial convergence speed is too fast under large regularities, which is not convenient for us to observe the acceleration ratio. Thus, we choose $\lambda_1 = \lambda_2 = 10^{-3}$ here. The result is similar to EC-Prox SGD, which performs well in the multi-core scenario.

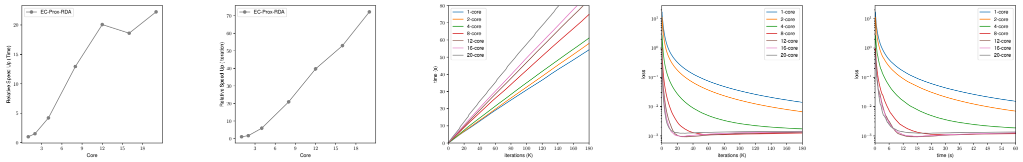


Figure 10: Distributed Error Compesated Proximal RDA (Top10) on Gisette ($\lambda_1 = \lambda_2 = 10^{-3}$)