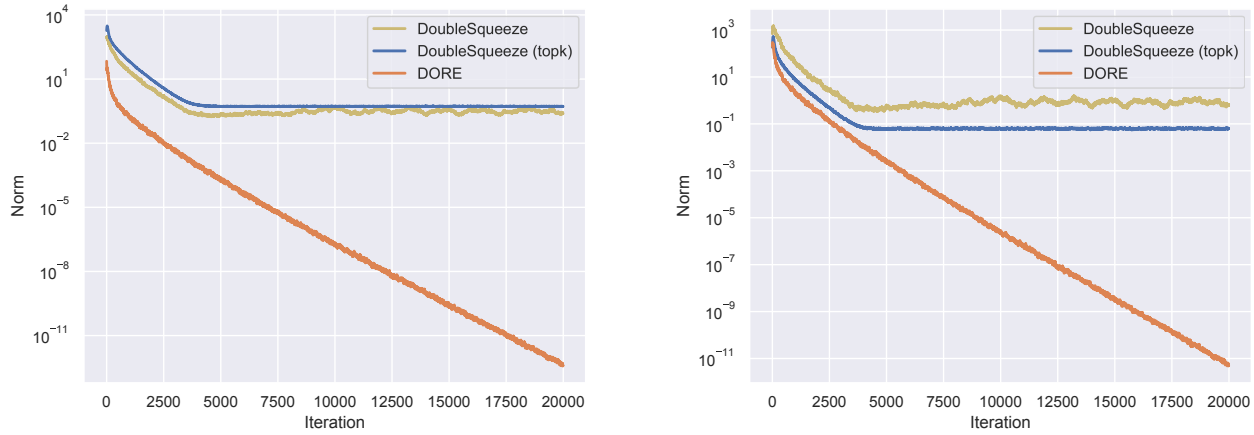
A Supplementary materials

A.1 Compression error

The property of the compression operator indicates that the compression error is linearly proportional to the norm of the variable being compressed:

$$\mathbb{E}\|Q(\mathbf{x}) - \mathbf{x}\|^2 \leq C\|\mathbf{x}\|^2.$$

We visualize the norm of the variables being compressed, i.e., the gradient residual (the worker side) and model residual (the master side) for DORE as well as error compensated gradient (the worker side) and averaged gradient (the master side) for DoubleSqueeze. As showed in Figure 6, the gradient and model residual of DORE decrease exponentially and the compression errors vanish. However, for DoubleSqueeze, their norms only decrease to some certain value and the compression error doesn't vanish. It explains why algorithms without residual compression cannot converge linearly to the $\mathcal{O}(\sigma)$ neighborhood of the optimal solution in the strongly convex case.



(a) Worker side

(b) Master side

Figure 6: The norm of variable being compressed in the linear regression experiment.

A.2 Communication Efficiency

To make an explicit comparison of communication efficiency, we report the training loss convergence with respect to communication bits in Figure 7, 8 and 9 for the experiments on synthetic data, MNIST and CIFAR10 dataset respectively. These results are independent of the system architectures and network bandwidth. It suggests that the proposed DORE reduce the communication cost significantly while maintaining good convergence speed.

Furthermore, we also test the running time of ResNet18 trained on CIFAR10 dataset under two different network bandwidth configurations, i.e. 1Gbps and 200Mbps, as showed in Figure 10 and 11. Due to its superior communication efficiency, the proposed DORE runs faster in both configurations. Moreover, when the network bandwidth reduces from 1Gbps to 200Mbps, the running time of DORE only increases slightly, which indicates that DORE is more robust to network bandwidth change and can work more efficiently under limited bandwidth. These results clearly suggest the advantages of the proposed algorithm.

All the experiments in this section are under the exactly same setting as described in Section 5. The running time is tested in a High Performance Computing Cluster with NVIDIA Tesla K80 GPUs and the computing nodes are connected by Gigabit Ethernet interfaces and we use mpi4py as the communication backend. All algorithms in this paper are implemented with PyTorch.

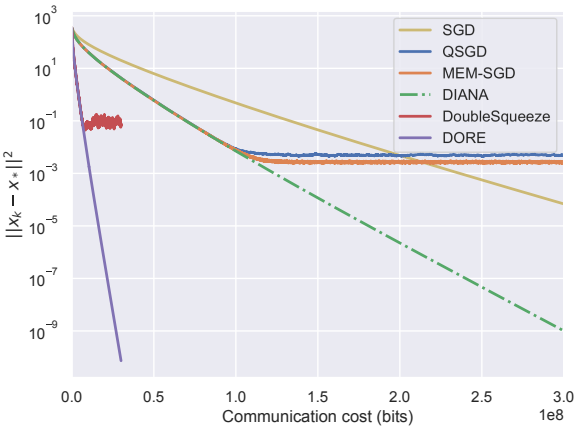


Figure 7: Linear regression on synthetic data.

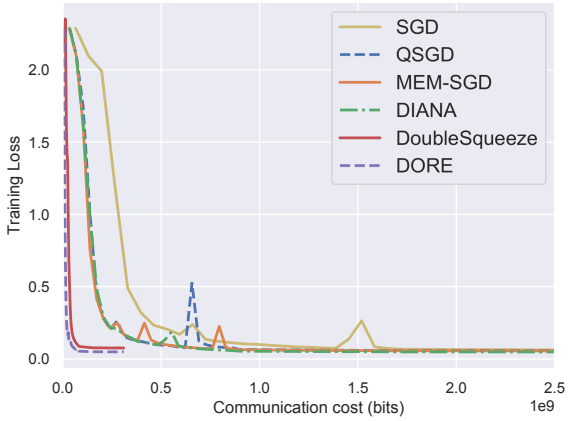


Figure 8: LeNet trained on MNIST dataset.

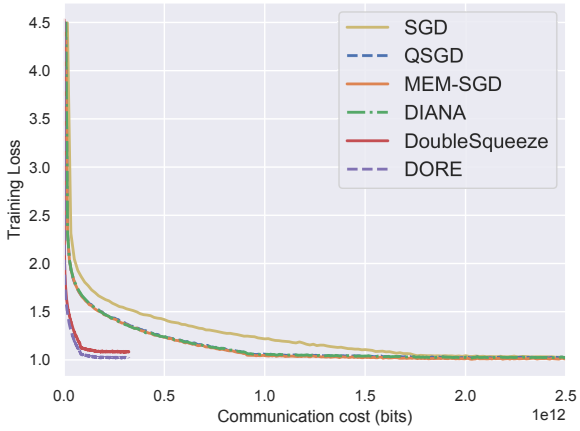


Figure 9: Resnet18 trained on CIFAR10 dataset.

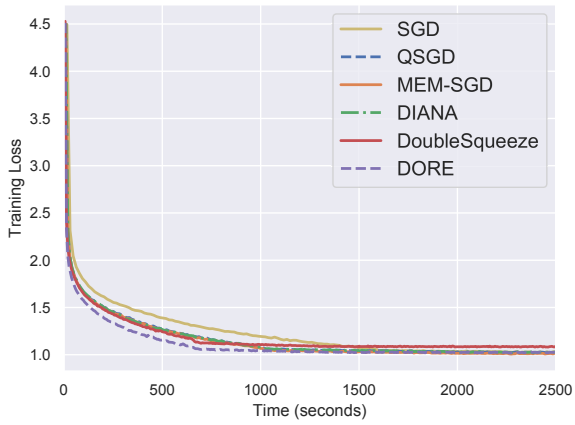


Figure 10: Resnet18 trained on CIFAR10 dataset with 1Gbps network bandwidth.

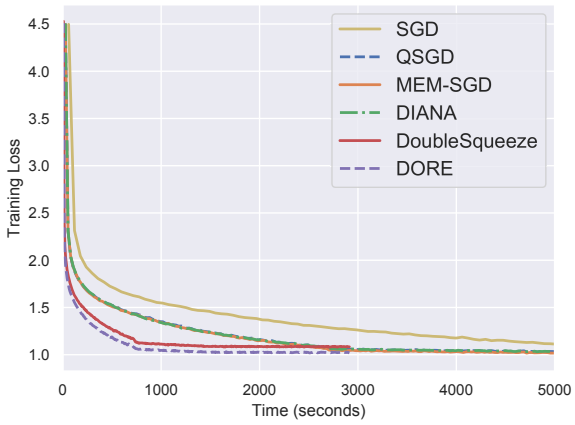
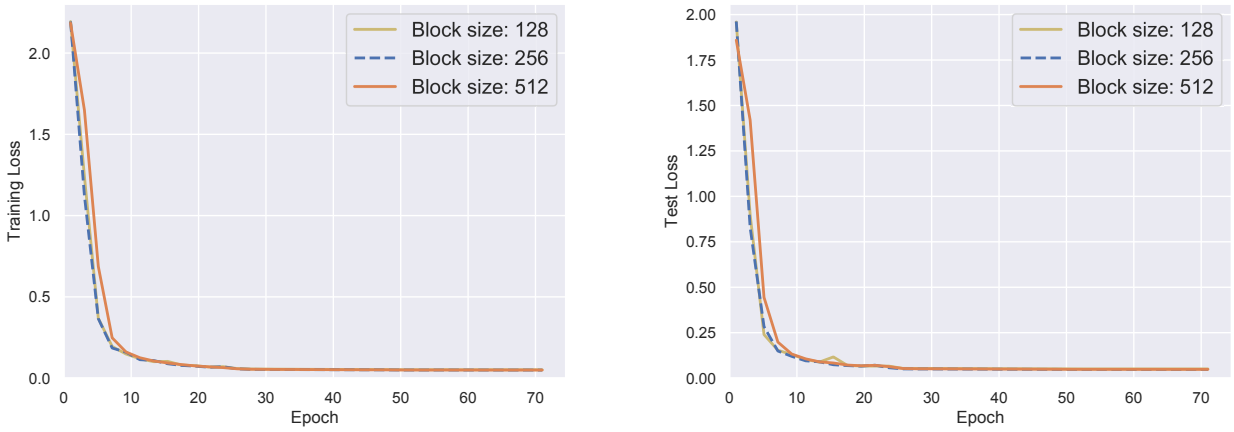


Figure 11: Resnet18 trained on CIFAR10 dataset with 200Mbps network bandwidth.

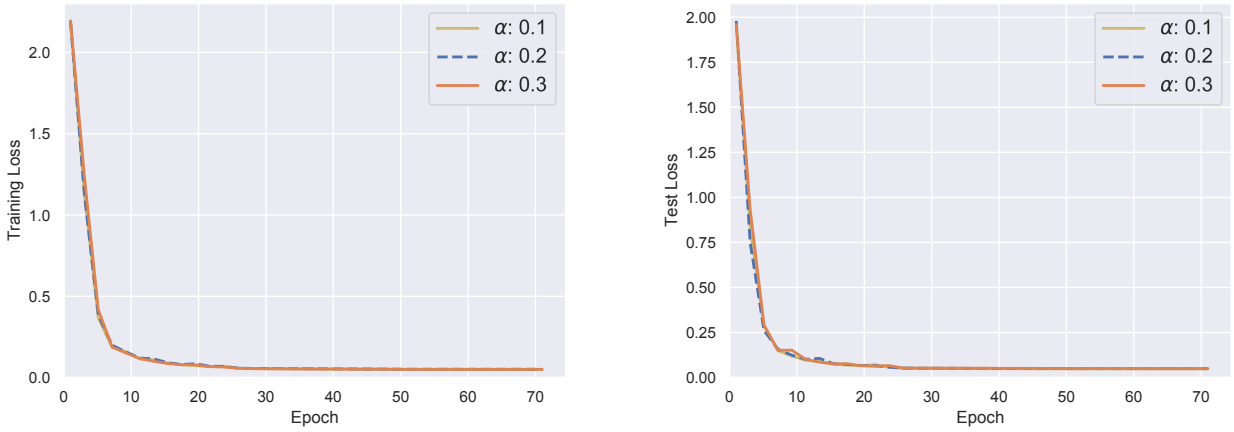
A.3 Parameter sensitivity

Continuing the MNIST experiment in Section 5, we further conduct parameter analysis on DORE. The basic setting for block size, learning rate, α, β and η are 256, 0.1, 0.1, 1, 1, respectively. We change each parameter individually. Figures 12, 13, 14, and 15 demonstrate that DORE performs consistently well under different parameter settings.



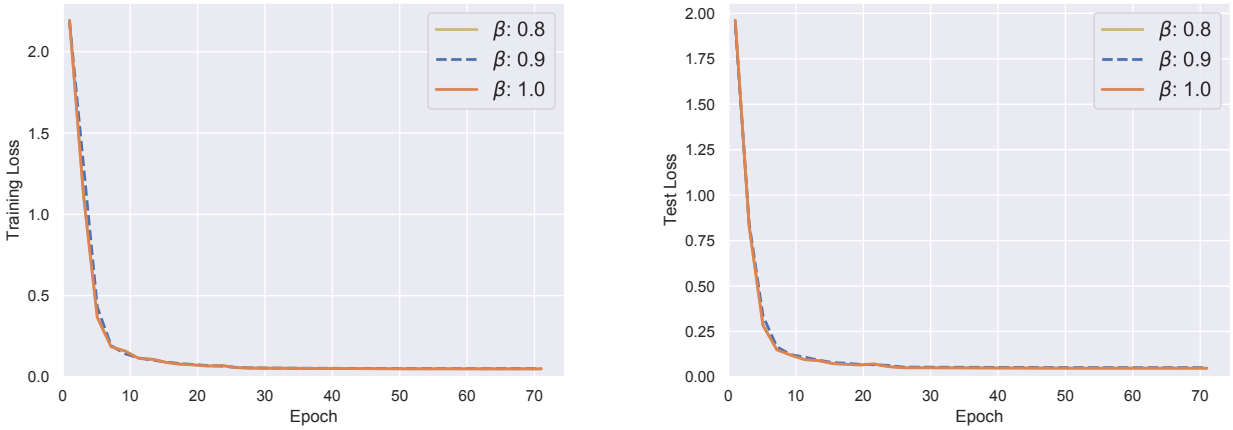
(a) Training loss (b) Test loss

Figure 12: Training under different compression block sizes.



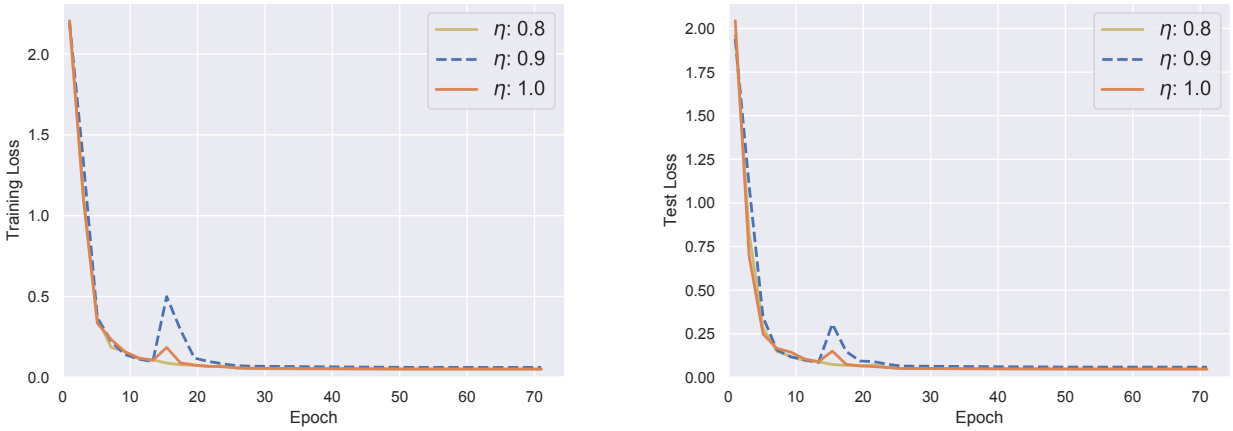
(a) Training loss (b) Test loss

Figure 13: Training under different α



(a) Training loss (b) Test loss

Figure 14: Training under different β



(a) Training loss (b) Test loss

Figure 15: Training under different η

A.4 DORE in the smooth case

Algorithm 2 DORE with $R(\mathbf{x}) = 0$

Input: Stepsize $\alpha, \beta, \gamma, \eta$, initialize $\mathbf{h}^0 = \mathbf{h}_i^0 = \mathbf{0}^d$, $\hat{\mathbf{x}}_i^0 = \hat{\mathbf{x}}^0$, $\forall i \in \{1, \dots, n\}$.
for $k = 1, 2, \cdots, K-1$ **do**

For each worker $\{i = 1, 2, \cdots, n\}$:
- Sample $\mathbf{g}_i^k$ such that $\mathbb{E}[\mathbf{g}_i^k | \hat{\mathbf{x}}_i^k] = \nabla f_i(\hat{\mathbf{x}}_i^k)$
- Gradient residual: $\Delta_i^k = \mathbf{g}_i^k - \mathbf{h}_i^k$
- Compression: $\hat{\Delta}_i^k = Q(\Delta_i^k)$
- $\mathbf{h}_i^{k+1} = \mathbf{h}_i^k + \alpha \hat{\Delta}_i^k$
- $\{\ \hat{\mathbf{g}}_i^k = \mathbf{h}_i^k + \hat{\Delta}_i^k\ \}$
- Sent $\hat{\Delta}_i^k$ to the master
- Receive $\hat{\mathbf{q}}^k$ from the master
- $\hat{\mathbf{x}}_i^{k+1} = \hat{\mathbf{x}}_i^k + \beta \hat{\mathbf{q}}^k$

For the master:
- Receive $\hat{\Delta}_i^k$s from workers
- $\hat{\Delta}^k = 1/n \sum_i^n \hat{\Delta}_i^k$
- $\hat{\mathbf{g}}^k = \mathbf{h}^k + \hat{\Delta}^k\ \ \{= 1/n \sum_i^n \hat{\mathbf{g}}_i^k\ \}$
- $\mathbf{h}^{k+1} = \mathbf{h}^k + \alpha \hat{\Delta}^k$
- $\mathbf{q}^k = -\gamma \hat{\mathbf{g}}^k + \eta \mathbf{e}^k$
- Compression: $\hat{\mathbf{q}}^k = Q(\mathbf{q}^k)$
- $\mathbf{e}^{k+1} = \mathbf{q}^k - \hat{\mathbf{q}}^k$
- Broadcast $\hat{\mathbf{q}}^k$ to workers

end for
Output: any $\hat{\mathbf{x}}_i^K$

A.5 Proof of Theorem 1

We first provide two lemmas. We define $\mathbb{E}_Q$, $\mathbb{E}_k$, and $\mathbb{E}$ be the expectation taken over the quantization, the kth iteration based on $\hat{\mathbf{x}}^k$, and the overall expectation, respectively.

Lemma 1. *For every i, we can estimate the first two moments of $\mathbf{h}_i^{k+1}$ as*

$$\mathbb{E}_Q \mathbf{h}_i^{k+1} = (1-\alpha)\mathbf{h}_i^k + \alpha \mathbf{g}_i^k, \tag{15}$$

$$\mathbb{E}_Q \|\mathbf{h}_i^{k+1} - \mathbf{s}_i\|^2 \le (1-\alpha)\|\mathbf{h}_i^k - \mathbf{s}_i\|^2 + \alpha \|\mathbf{g}_i^k - \mathbf{s}_i\|^2 + \alpha[(C_q+1)\alpha - 1]\|\Delta_i^k\|^2. \tag{16}$$

Proof. The first equality follows from lines 5-7 of Algorithm 1 and Assumption 1. For the second equation, we have the following variance decomposition

$$\mathbb{E}\|X\|^2 = \|\mathbb{E}X\|^2 + \mathbb{E}\|X - \mathbb{E}X\|^2 \tag{17}$$

for any random vector X. By taking $X = \mathbf{h}_i^{k+1} - \mathbf{s}_i$, we get

$$\mathbb{E}_Q \|\mathbf{h}_i^{k+1} - \mathbf{s}_i\|^2 = \|(1-\alpha)(\mathbf{h}_i^k - \mathbf{s}_i) + \alpha(\mathbf{g}_i^k - \mathbf{s}_i)\|^2 + \alpha^2 \mathbb{E}_Q \|\hat{\Delta}_i^k - \Delta_i^k\|^2. \tag{18}$$

Using the basic equality

$$\|\lambda \mathbf{a} + (1-\lambda)\mathbf{b}\|^2 + \lambda(1-\lambda)\|\mathbf{a} - \mathbf{b}\|^2 = \lambda\|\mathbf{a}\|^2 + (1-\lambda)\|\mathbf{b}\|^2 \tag{19}$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\lambda \in [0, 1]$, as well as Assumption 1, we have

$$\mathbb{E}_Q \|\mathbf{h}_i^{k+1} - \mathbf{s}_i\|^2 \le (1-\alpha)\|\mathbf{h}_i^k - \mathbf{s}_i\|^2 + \alpha \|\mathbf{g}_i^k - \mathbf{s}_i\|^2 - \alpha(1-\alpha)\|\Delta_i^k\|^2 + \alpha^2 C_q \|\Delta_i^k\|^2, \tag{20}$$

which is the inequality (16). □

Next, from the variance decomposition (17), we also derive Lemma 2.

Lemma 2. *The following inequality holds*

$$\mathbb{E}[\|\hat{\mathbf{g}}^k - \mathbf{h}^*\|^2] \le \mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k) - \mathbf{h}^*\|^2 + \frac{C_q}{n^2} \sum_{i=1}^n \mathbb{E}\|\Delta_i^k\|^2 + \frac{\sigma^2}{n}, \tag{21}$$

where $\mathbf{h}^* = \nabla f(\mathbf{x}^*) = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i^*$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

Proof. By taking the expectation over the quantization of $\mathbf{g}$, we have

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{g}}^k-\mathbf{h}^*\|^2&=\mathbb{E}\|\mathbf{g}^k-\mathbf{h}^*\|^2+\mathbb{E}\|\hat{\mathbf{g}}^k-\mathbf{g}^k\|^2\\&\leq\mathbb{E}\|\mathbf{g}^k-\mathbf{h}^*\|^2+\frac{C_q}{n^2}\sum_{i=1}^n\mathbb{E}\|\Delta_i^k\|^2,\end{aligned}\tag{22}$$

where the inequality is from Assumption 1.

For $\|\mathbf{g}^k-\mathbf{h}^*\|$, we take the expectation over the sampling of gradients and derive

$$\begin{aligned}\mathbb{E}\|\mathbf{g}^k-\mathbf{h}^*\|^2&=\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\|^2+\mathbb{E}\|\mathbf{g}^k-\nabla f(\hat{\mathbf{x}}^k)\|^2\\&\leq\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\|^2+\frac{\sigma^2}{n}\end{aligned}\tag{23}$$

by Assumption 2.

Combining (22) with (23) gives (21). □

Proof of Theorem 1. We consider $\mathbf{x}^{k+1}-\mathbf{x}^*$ first. Since $\mathbf{x}^*$ is the solution of (1), it satisfies

$$\mathbf{x}^*=\mathbf{prox}_{\gamma R}(\mathbf{x}^*-\gamma\mathbf{h}^*).\tag{24}$$

Hence

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2=&\mathbb{E}\|\mathbf{prox}_{\gamma R}(\hat{\mathbf{x}}^k-\gamma\hat{\mathbf{g}}^k)-\mathbf{prox}_{\gamma R}(\mathbf{x}^*-\gamma\mathbf{h}^*)\|^2\\\leq&\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*-\gamma(\hat{\mathbf{g}}^k-\mathbf{h}^*)\|^2\\=&\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\gamma\mathbb{E}\langle\hat{\mathbf{x}}^k-\mathbf{x}^*,\hat{\mathbf{g}}^k-\mathbf{h}^*\rangle+\gamma^2\mathbb{E}\|\hat{\mathbf{g}}^k-\mathbf{h}^*\|^2\\=&\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\gamma\mathbb{E}\langle\hat{\mathbf{x}}^k-\mathbf{x}^*,\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\rangle+\gamma^2\mathbb{E}\|\hat{\mathbf{g}}^k-\mathbf{h}^*\|^2,\end{aligned}\tag{25}$$

where the inequality comes from the non-expansiveness of the proximal operator and the last equality is derived by taking the expectation of the stochastic gradient $\hat{\mathbf{g}}^k$. Combining (21) and (25), we have

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2\leq&\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\gamma\mathbb{E}\langle\hat{\mathbf{x}}^k-\mathbf{x}^*,\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\rangle\\&+\frac{\gamma^2}{n}\sum_{i=1}^n\mathbb{E}\|\nabla f_i(\hat{\mathbf{x}}^k)-\mathbf{h}_i^*\|^2+\frac{C_q\gamma^2}{n^2}\sum_{i=1}^n\mathbb{E}\|\Delta_i^k\|^2+\frac{\gamma^2}{n}\sigma^2.\end{aligned}\tag{26}$$

Then we consider $\mathbb{E}\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2$. According to Algorithm 1, we have:

$$\begin{aligned}\mathbb{E}_Q[\hat{\mathbf{x}}^{k+1}-\mathbf{x}^*]&=\hat{\mathbf{x}}^k+\beta\mathbf{q}^k-\mathbf{x}^*\\&=(1-\beta)(\hat{\mathbf{x}}^k-\mathbf{x}^*)+\beta(\mathbf{x}^{k+1}-\mathbf{x}^*+\eta\mathbf{e}^k)\end{aligned}\tag{27}$$

where the expectation is taken on the quantization of $\mathbf{q}^k$.

By variance decomposition (17) and the basic equality (19),

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2\\\leq&(1-\beta)\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+\beta\mathbb{E}\|\mathbf{x}^{k+1}+\eta\mathbf{e}^k-\mathbf{x}^*\|^2-\beta(1-\beta)\mathbb{E}\|\mathbf{q}^k\|^2+\beta^2C_q^m\mathbb{E}\|\mathbf{q}^k\|^2\\\leq&(1-\beta)\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+(1+\eta^2\epsilon)\beta\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2-\beta(1-(C_q^m+1)\beta)\mathbb{E}\|\mathbf{q}^k\|^2+(\eta^2+\frac{1}{\epsilon})\beta C_q^m\mathbb{E}\|\mathbf{q}^{k-1}\|^2,\end{aligned}\tag{28}$$

where ϵ is generated from Cauchy inequality of inner product. For convenience, we let $\epsilon=\frac{1}{\eta}$.

Choose a β such that $0<\beta\leq\frac{1}{1+C_q^m}$. Then we have

$$\begin{aligned}&\beta(1-(C_q^m+1)\beta)\mathbb{E}\|\mathbf{q}^k\|^2+\mathbb{E}\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2\\\leq&(1-\beta)\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+(1+\eta)\beta\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2+(\eta^2+\eta)\beta C_q^m\mathbb{E}\|\mathbf{q}^{k-1}\|^2.\end{aligned}\tag{29}$$

Letting $\mathbf{s}_i = \mathbf{h}_i^*$ in (16), we have

$$\begin{aligned}
&\frac{(1+\eta)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^{k+1}-\mathbf{h}_i^*\|^2\\
\leq&\frac{(1+\eta)(1-\alpha)c\beta\gamma^2}{n}\sum_{i=1}^n \|\mathbf{h}_i^{k}-\mathbf{h}_i^*\|^2+\frac{(1+\eta)\alpha c\beta\gamma^2}{n}\sum_{i=1}^n \|\mathbf{g}_i^{k}-\mathbf{h}_i^*\|^2\\
&+\frac{(1+\eta)\alpha[(C_q+1)\alpha-1]c\beta\gamma^2}{n}\sum_{i=1}^n \|\Delta_i^k\|^2.
\end{aligned} \tag{30}$$

Then we let $\mathbf{R}^k = \beta(1-(C_q^m+1)\beta)\mathbb{E}\|\mathbf{q}^k\|^2$ and define $\mathbf{V}^k = \mathbf{R}^{k-1}+\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+\frac{(1+\eta)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^k-\mathbf{h}_i^*\|^2$. Thus, we obtain

$$\begin{aligned}
\mathbf{V}^{k+1}\leq&(\eta^2+\eta)\beta C_q^m\mathbb{E}\|\mathbf{q}^{k-1}\|^2+(1+\eta\beta)\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2-2(1+\eta)\beta\gamma\mathbb{E}\langle\hat{\mathbf{x}}^k-\mathbf{x}^*,\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\rangle\\
&+\frac{(1+\eta)(1-\alpha)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^{k}-\mathbf{h}_i^*\|^2+\frac{(1+\eta)\beta\gamma^2}{n^2}\Big[nc(C_q+1)\alpha^2-nc\alpha+C_q\Big]\sum_{i=1}^n \mathbb{E}\|\Delta_i^k\|^2\\
&+\frac{(1+\eta)(1+c\alpha)}{n}\beta\gamma^2\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\hat{\mathbf{x}}^k)-\mathbf{h}_i^*\|^2+\frac{(1+\eta)(1+nc\alpha)}{n}\beta\gamma^2\sigma^2.
\end{aligned} \tag{31}$$

The $\mathbb{E}\|\Delta_i^k\|^2$-term can be ignored if $nc(C_q+1)\alpha^2-nc\alpha+C_q\leq 0$, which can be guaranteed by $c\geq\frac{4C_q(C_q+1)}{n}$ and

$$\alpha\in\left(\frac{1-\sqrt{1-\frac{4C_q(C_q+1)}{nc}}}{2(C_q+1)},\frac{1+\sqrt{1-\frac{4C_q(C_q+1)}{nc}}}{2(C_q+1)}\right).$$

Given that each f_i is L-Lipschitz differentiable and μ-strongly convex, we have

$$\mathbb{E}\langle\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*,\hat{\mathbf{x}}^k-\mathbf{x}^*\rangle\geq\frac{\mu L}{\mu+L}\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+\frac{1}{\mu+L}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\|\nabla f_i(\hat{\mathbf{x}}^k)-\mathbf{h}_i^*\|^2. \tag{32}$$

Hence

$$\begin{aligned}
\mathbf{V}^{k+1}\leq&\rho_1\mathbf{R}^{k-1}+(1+\eta\beta)\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2-2(1+\eta)\beta\gamma\mathbb{E}\langle\hat{\mathbf{x}}^k-\mathbf{x}^*,\nabla f(\hat{\mathbf{x}}^k)-\mathbf{h}^*\rangle\\
&+\frac{(1+\eta)(1-\alpha)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^{k}-\mathbf{h}_i^*\|^2+\frac{(1+\eta)(1+c\alpha)}{n}\beta\gamma^2\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\hat{\mathbf{x}}^k)-\mathbf{h}_i^*\|^2+\frac{(1+\eta)(1+nc\alpha)}{n}\beta\gamma^2\sigma^2\\
\leq&\rho_1\mathbf{R}^{k-1}+\Big[1+\eta\beta-\frac{2(1+\eta)\beta\gamma\mu L}{\mu+L}\Big]\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+\frac{(1+\eta)(1-\alpha)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^{k}-\mathbf{h}_i^*\|^2\\
&+\Big[(1+\eta)(1+c\alpha)\beta\gamma^2-\frac{2(1+\eta)\beta\gamma}{\mu+L}\Big]\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\hat{\mathbf{x}}^k)-\mathbf{h}_i^*\|^2+\frac{(1+\eta)(1+nc\alpha)}{n}\beta\gamma^2\sigma^2\\
\leq&\rho_1\mathbf{R}^{k-1}+\rho_2\mathbb{E}\|\hat{\mathbf{x}}^k-\mathbf{x}^*\|^2+\frac{(1+\eta)(1-\alpha)c\beta\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^{k}-\mathbf{h}_i^*\|^2+\frac{(1+\eta)(1+nc\alpha)}{n}\beta\gamma^2\sigma^2
\end{aligned} \tag{33}$$

where

$$\begin{aligned}
\rho_1=&\frac{(\eta^2+\eta)C_q^m}{1-(C_q^m+1)\beta},\\
\rho_2=&1+\eta\beta-\frac{2(1+\eta)\beta\gamma\mu L}{\mu+L}.
\end{aligned}$$

Here we let $\gamma \leq \frac{2}{(1+c\alpha)(\mu+L)}$ such that $(1+\eta)(1+c\alpha)\beta\gamma^2 - \frac{2(1+\eta)\beta\gamma}{\mu+L} \leq 0$ and the last inequality holds. In order to get $\max(\rho_1, \rho_2, 1-\alpha) < 1$, we have the following conditions

$$\begin{aligned} 0 \leq (\eta^2+\eta)C_q^m \leq & 1-(C_q^m+1)\beta, \\ \eta < & \frac{2(1+\eta)\gamma\mu L}{\mu+L}. \end{aligned}$$

Therefore, the condition for γ is

$$\frac{\eta(\mu+L)}{2(1+\eta)\mu L} \leq \gamma \leq \frac{2}{(1+c\alpha)(\mu+L)},$$

which implies an additional condition for η. Therefore, the condition for η is

$$\eta \in \left[0, \min\left(\frac{-C_q^m + \sqrt{(C_q^m)^2 + 4(1-(C_q^m+1)\beta)}}{2C_q^m}, \frac{4\mu L}{(\mu+L)^2(1+c\alpha) - 4\mu L}\right)\right).$$

where $\eta \leq \frac{4\mu L}{(\mu+L)^2(1+c\alpha)-4\mu L}$ is to ensure $\frac{\eta(\mu+L)}{2(1+\eta)\mu L} \leq \frac{2}{(1+c\alpha)(\mu+L)}$ such that we don't get an empty set for γ.

If we define $\rho = \max\{\rho_1, \rho_2, 1-\alpha\}$, we obtain

$$\mathbf{V}^{k+1} \leq \rho \mathbf{V}^k + \frac{(1+\eta)(1+nc\alpha)}{n}\beta\gamma^2\sigma^2 \tag{34}$$

and the proof is completed by applying (34) recurrently. □

A.6 Proof of Theorem 2

Proof. In Algorithm 2, we can show

$$\begin{aligned} \mathbb{E}\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\|^2 &= \beta^2\mathbb{E}\|\hat{\mathbf{q}}^k\|^2 = \beta^2\mathbb{E}\|\mathbb{E}\hat{\mathbf{q}}^k\|^2 + \beta^2\mathbb{E}\|\hat{\mathbf{q}}^k - \mathbb{E}\hat{\mathbf{q}}^k\|^2 \\ &= \beta^2\mathbb{E}\|\mathbf{q}^k\|^2 + \beta^2\mathbb{E}\|\hat{\mathbf{q}}^k - \mathbf{q}^k\|^2 \\ &\leq (1+C_q^m)\beta^2\mathbb{E}\|\mathbf{q}^k\|^2. \end{aligned} \tag{35}$$

and

$$\mathbb{E}\|\mathbf{q}^k\|^2 = \mathbb{E}\| - \gamma\hat{\mathbf{g}}^k + \eta\mathbf{e}^k\|^2 \leq 2\gamma^2\mathbb{E}\|\hat{\mathbf{g}}^k\|^2 + 2\eta^2\mathbb{E}\|\mathbf{e}^k\|^2 \leq 2\gamma^2\mathbb{E}\|\hat{\mathbf{g}}^k\|^2 + 2C_q^m\eta^2\mathbb{E}\|\mathbf{q}^{k-1}\|^2. \tag{36}$$

Using (35)(36) and the Lipschitz continuity of $\nabla f(\mathbf{x})$, we have

$$\begin{aligned} &\mathbb{E}f(\hat{\mathbf{x}}^{k+1}) + (C_q^m+1)L\beta^2\mathbb{E}\|\mathbf{q}^k\|^2 \\ \leq &\mathbb{E}f(\hat{\mathbf{x}}^k) + \mathbb{E}\langle\nabla f(\hat{\mathbf{x}}^k), \hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\rangle + \frac{L}{2}\mathbb{E}\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\|^2 + (C_q^m+1)L\beta^2\mathbb{E}\|\mathbf{q}^k\|^2 \\ = &\mathbb{E}f(\hat{\mathbf{x}}^k) + \beta\mathbb{E}\langle\nabla f(\hat{\mathbf{x}}^k), -\gamma\hat{\mathbf{g}}^k + \eta\mathbf{e}^k\rangle + \frac{(1+C_q^m)L\beta^2}{2}\mathbb{E}\|\mathbf{q}^k\|^2 + (C_q^m+1)L\beta^2\mathbb{E}\|\mathbf{q}^k\|^2 \\ = &\mathbb{E}f(\hat{\mathbf{x}}^k) + \beta\mathbb{E}\langle\nabla f(\hat{\mathbf{x}}^k), -\gamma\nabla f(\hat{\mathbf{x}}^k) + \eta\mathbf{e}^k\rangle + \frac{3(C_q^m+1)L\beta^2}{2}\mathbb{E}\|\mathbf{q}^k\|^2 \\ \leq &\mathbb{E}f(\hat{\mathbf{x}}^k) - \beta\gamma\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 + \frac{\beta\eta}{2}\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 + \frac{\beta\eta}{2}\mathbb{E}\|\mathbf{e}^k\|^2 \\ &\quad + 3(C_q^m+1)L\beta^2\left[\gamma^2\mathbb{E}\|\hat{\mathbf{g}}^k\|^2 + C_q^m\eta^2\mathbb{E}\|\mathbf{q}^{k-1}\|^2\right] \\ \leq &\mathbb{E}f(\hat{\mathbf{x}}^k) - \left[\beta\gamma - \frac{\beta\eta}{2} - 3(C_q^m+1)L\beta^2\gamma^2\right]\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 \\ &+ \frac{3C_q(C_q^m+1)L\beta^2\gamma^2}{n^2}\sum_{i=1}^n\mathbb{E}\|\Delta_i^k\|^2 + \frac{3(C_q^m+1)L\beta^2\gamma^2}{n}\sigma^2 \\ &+ \left[\frac{\beta\eta C_q^m}{2} + (3C_q^m+1)C_q^mL\beta^2\eta^2\right]\mathbb{E}\|\mathbf{q}^{k-1}\|^2, \end{aligned} \tag{37}$$

where the last inequality is from (21) with $\mathbf{h}^* = \mathbf{0}$.

Letting $\mathbf{s}_i = \mathbf{0}$ in (16), we have

$$\mathbb{E}_Q\|\mathbf{h}_i^{k+1}\|^2 \leq (1-\alpha)\|\mathbf{h}_i^k\|^2 + \alpha\|\mathbf{g}_i^k\|^2 + \alpha[(C_q+1)\alpha - 1]\|\Delta_i^k\|^2. \tag{38}$$

Due to the assumption that each worker samples the gradient from the full dataset, we have

$$\mathbb{E}\mathbf{g}_i^k = \mathbb{E}\nabla f(\hat{\mathbf{x}}^k), \quad \mathbb{E}\|\mathbf{g}_i^k\|^2 \leq \mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 + \sigma_i^2. \tag{39}$$

Define $\Lambda^k = (C_q^m+1)L\beta^2\|\mathbf{q}^{k-1}\|^2 + f(\hat{\mathbf{x}}^k) - f^* + 3c(C_q^m+1)L\beta^2\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^k\|^2$, and from (37), (38), and (39), we have

$$\begin{aligned}
\mathbb{E}\Lambda^{k+1} \leq & \mathbb{E}f(\hat{\mathbf{x}}^k) - f^* + 3(1-\alpha)c(C_q^m+1)L\beta^2\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^k\|^2 \\
& - \Big[\beta\gamma - \frac{\beta\eta}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma^2\Big]\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 \\
& + \frac{(C_q^m+1)L\beta^2\gamma^2}{n^2}\Big[3nc(C_q+1)\alpha^2 - 3nc\alpha + 3C_q\Big]\sum_{i=1}^n \mathbb{E}\|\Delta_i^k\|^2 \\
& + 3(1+nc\alpha)\frac{(C_q^m+1)L\beta^2\gamma^2\sigma^2}{n} \\
& + \Big[\frac{\beta\eta C_q^m}{2} + 3(C_q^m+1)C_q^m L\beta^2\eta^2\Big]\mathbb{E}\|\mathbf{q}^{k-1}\|^2.
\end{aligned} \tag{40}$$

If we let $c = \frac{4C_q(C_q+1)}{n}$, then the condition of α in (5) gives $3nc(C_q+1)\alpha^2 - 3nc\alpha + 3C_q \leq 0$ and

$$\begin{aligned}
\mathbb{E}\Lambda^{k+1} \leq & \mathbb{E}f(\hat{\mathbf{x}}^k) - f^* + 3(1-\alpha)c(C_q^m+1)L\beta^2\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\mathbf{h}_i^k\|^2 \\
& - \Big[\beta\gamma - \frac{\beta\eta}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma^2\Big]\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 \\
& + 3(1+nc\alpha)\frac{(C_q^m+1)L\beta^2\gamma^2\sigma^2}{n} \\
& + [\frac{\beta\eta C_q^m}{2} + 3(C_q^m+1)C_q^m L\beta^2\eta^2]\mathbb{E}\|\mathbf{q}^{k-1}\|^2.
\end{aligned} \tag{41}$$

Let $\eta = \gamma$ and $\beta\gamma \leq \frac{1}{6(1+c\alpha)(C_q^m+1)L}$, we have

$$\beta\gamma - \frac{\beta\eta}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma^2 = \frac{\beta\gamma}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma^2 \geq 0.$$

Take $\gamma \leq \min\Big\{\frac{-1+\sqrt{1+\frac{48L^2\beta^2(C_q^m+1)^2}{C_q^m}}}{12L\beta(C_q^m+1)}, \frac{1}{6L\beta(1+c\alpha)(C_q^m+1)}\Big\}$ will guarantee

$$\Big[\frac{\beta\eta C_q^m}{2} + 3(C_q^m+1)C_q^m L\beta^2\eta^2\Big] \leq (C_q^m+1)L\beta^2.$$

Hence we obtain

$$\mathbb{E}\Lambda^{k+1} \leq \mathbb{E}\Lambda^k - \Big[\frac{\beta\gamma}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma^2\Big]\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 + 3(1+nc\alpha)\frac{(C_q^m+1)L\beta^2\gamma^2\sigma^2}{n}. \tag{42}$$

Taking the telescoping sum and plugging the initial conditions, we derive (12). □

A.7 Proof of Corollary 2

Proof. With $\alpha = \frac{1}{2(C_q+1)}$ and $c = \frac{4C_q(C_q+1)}{n}$, $1 + nc\alpha = 1 + 2C_q$ is a constant. We set $\beta = \frac{1}{C_q^m+1}$ and $\gamma = \min\left\{\frac{-1+\sqrt{1+\frac{48L^2}{C_q^m}}}{12L}, \frac{1}{12L(1+c\alpha)(1+\sqrt{K/n})}\right\}$. In general, C_q^m is bounded which makes the first bound negligible, i.e., $\gamma = \frac{1}{12L(1+c\alpha)(1+\sqrt{K/n})}$ when K is large enough. Therefore, we have

$$\frac{\beta}{2} - 3(1+c\alpha)(C_q^m+1)L\beta^2\gamma = \frac{1-6(1+c\alpha)L\gamma}{2(C_q^m+1)} \leq \frac{1}{4(C_q^m+1)}. \tag{43}$$

From Theorem 2, we derive

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^k)\|^2 \\
\leq&\frac{4(C_q^m+1)(\mathbb{E}\Lambda^1 - \mathbb{E}\Lambda^{K+1})}{\gamma K} + \frac{12(1+nc\alpha)L\sigma^2\gamma}{n} \\
\leq&48L(C_q^m+1)(1+c\alpha)(\mathbb{E}\Lambda^1 - \mathbb{E}\Lambda^{K+1})(\frac{1}{K} + \frac{1}{\sqrt{nK}}) + \frac{(1+nc\alpha)\sigma^2}{(1+c\alpha)}\frac{1}{\sqrt{nK}},
\end{aligned} \tag{44}$$

which completes the proof.

□