Supplementary Material - SADDLe: Sharpness-Aware Decentralized Deep Learning with Heterogeneous Data

1. Theoretical Analysis

The update rule for Q-SADDLe with SAM-based gradient $\widetilde{\mathbf{G}}$ is as follows:

$$\begin{aligned}\mathbf{X}^{(t+1)} &= \mathbf{W}\left(\mathbf{X}^{(t)} - \eta\left(\beta\mathbf{M}^{(t)} + \widetilde{\mathbf{G}}^{(t)}\right)\right)\\ \mathbf{M}^{(t+1)} &= \mu\mathbf{M}^{(t)} + (1-\mu)\frac{\mathbf{X}^{(t)} - \mathbf{X}^{(t+1)}}{\eta}\\ &= (\mu + (1-\mu)\beta\mathbf{W})\,\mathbf{M}^{(t)} + (1-\mu)\mathbf{W}\widetilde{\mathbf{G}}^{(t)}\\ &\quad + \frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)},\end{aligned} \tag{1}$$

For a doubly stochastic mixing matrix $\mathbf{W}$, we can simplify the updates as follows:

$$\begin{aligned}\bar{\mathbf{x}}^{(t+1)} &= \bar{\mathbf{x}}^{(t)} - \eta\left(\beta\bar{\mathbf{m}}^{(t)} + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i^t\right),\\ \bar{\mathbf{m}}^{(t+1)} &= \mu\bar{\mathbf{m}}^{(t)} + (1-\mu)\frac{\bar{\mathbf{x}}^{(t)} - \bar{\mathbf{x}}^{(t+1)}}{\eta}\\ &= (1-(1-\mu)(1-\beta))\bar{\mathbf{m}}^{(t)} + (1-\mu)\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i^t.\end{aligned} \tag{2}$$

Here, $\widetilde{\mathbf{g}}_i^t$ is the SAM-based gradient update, which we reiterate for ease of understanding :

$$\widetilde{\mathbf{g}}_i^t = \nabla F_i(\mathbf{x}_i^t + \xi(\mathbf{x}_i^t); d_i^t), \quad \text{where} \quad \xi(\mathbf{x}_i^t) = \rho\frac{\mathbf{g}_i^t}{\|\mathbf{g}_i^t\|} \tag{3}$$

For the rest of the analysis, we use $\xi(\mathbf{x}_i^t) = \xi_i^t$ for simplicity of notation. We introduce the following lemma to define an upper bound on the stochastic variance of SAM-based updates.

Lemma 1 *Given assumptions 1-3, the stochastic variance of local gradients with perturbation can be bounded as*

$$\|\nabla F_i(\mathbf{x}_i + \xi_i) - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2 \leq 3\sigma^2 + 6L^2\rho^2 \tag{4}$$

Proof:

$$\begin{aligned}&\|\nabla F_i(\mathbf{x}_i + \xi_i) - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2 =\\ &\|\nabla F_i(\mathbf{x}_i + \xi_i) - \nabla F_i(\mathbf{x}_i) + \nabla F_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i) + \nabla f_i(\mathbf{x}_i)\\ &- \nabla f_i(\mathbf{x}_i + \xi_i)\|^2 \overset{a}{\leq} 3\|\nabla F_i(\mathbf{x}_i + \xi_i) - \nabla F_i(\mathbf{x}_i)\|^2\\ &+ 3\|\nabla F_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i)\|^2 + 3\|\nabla f_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2\\ &\overset{b}{\leq} 3\|\nabla F_i(\mathbf{x}_i + \xi_i) - \nabla F_i(\mathbf{x}_i)\|^2 + 3\sigma^2\\ &+ 3\|\nabla f_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2 \overset{c}{\leq} 3\sigma^2 + 6L^2\rho^2\end{aligned} \tag{5}$$

(a) follows from the property $\|x_1 + x_2 + ...x_n\|^2 \leq n[\|x_1\|^2 + \|x_2\|^2...\|x_n\|^2]$ for random variables $x_1, x_2, ...x_n$. (b) follows from Assumption 2 in the main paper. (c) follows from Assumption 1 and the perturbation ξ_i being bounded by the perturbation radius ρ.

Lemma 2 *Given assumptions 1-3 and $\widetilde{\mathbf{g}}_i = \nabla F_i(\mathbf{x}_i + \xi_i)$, the following relationship holds*

$$\mathbb{E}\|[\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i\|^2] \leq \frac{3\sigma^2}{n} + \frac{6L^2\rho^2}{n} + \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i + \xi_i)\|^2] \tag{6}$$

Proof:

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i\|^2] = \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2]\\ &+ \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i + \xi_i)\|^2]\\ &= \frac{1}{n^2}\|\sum_{i=1}^{n}\mathbb{E}[\|\widetilde{\mathbf{g}}_i - \nabla f_i(\mathbf{x}_i + \xi_i)\|^2]\\ &+ \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i + \xi_i)\|^2] \leq \frac{3\sigma^2}{n} + \frac{6L^2\rho^2}{n}\\ &+ \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i + \xi_i)\|^2]\end{aligned} \tag{7}$$

As a first step, we simplify our convergence analysis by

defining another sequence of parameters $\mathbf{z}^{(t)}$ with the following update rule:

$$\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \left(\frac{\eta}{1-\beta}\right)\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i^t \tag{8}$$

Inspired by QGM [8], this sequence has a simpler SAM update rule, while our parameters $\bar{\mathbf{x}}^{(t)}$ follow SAM-based gradient updates along with a momentum buffer $\mathbf{m}_i^t$. We use $\bar{\mathbf{g}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{g}}_i^t$ and $\tilde{\eta} = \frac{\eta}{1-\beta}$ for rest of the analysis. We begin by proving that the error $\mathbf{e}^{(t)} = \mathbf{z}^{(t)} - \bar{\mathbf{x}}^{(t)}$ remains bounded.

Lemma 3 *Given Assumptions 1-3, the sequence of iterates generated by Q-SADDLe satisfy*

$$\mathbb{E}\left\|\mathbf{e}^{(t+1)}\right\|^2 \leq (1-(1-\mu)(1-\beta))\mathbb{E}\left\|\mathbf{e}^{(t)}\right\|^2 + \frac{2\tilde{\eta}^2\beta^2}{(1-\beta)(1-\mu)}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i+\xi_i)\right\|^2 + 3\tilde{\eta}^2\beta^2\sigma^2 + 6\tilde{\eta}^2\beta^2L^2\rho^2\,.$$

Proof: For $\mathbf{e}^{(0)} = 0$, specifying $\mathbf{e}^{(t+1)}$ in terms of update sequences $\mathbf{z}^{(t+1)}$ and $\bar{\mathbf{x}}^{(t+1)}$:

$$\mathbf{e}^{(t+1)} = \mathbf{z}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} = \left(\mathbf{z}^{(t)} - \frac{\eta}{1-\beta}\bar{\mathbf{g}}^{(t)}\right) - (\bar{\mathbf{x}}^{(t)} - \eta(\beta\bar{\mathbf{m}}^{(t)} + \bar{\mathbf{g}}^{(t)})) = \mathbf{e}^{(t)} - \eta\beta(\frac{1}{1-\beta}\bar{\mathbf{g}}^{(t)} - \bar{\mathbf{m}}^{(t)}) = \sum_{k=0}^{t} -\eta\beta(\frac{1}{1-\beta}\bar{\mathbf{g}}^{(k)} - \bar{\mathbf{m}}^{(k)})\,. \tag{9}$$

Using equation (2), we have [8]:

$$\mathbf{e}^{(t+1)} = \sum_{k=0}^{(t)} -\eta\beta(\frac{1}{1-\beta}\bar{\mathbf{g}}^{(k)} - ((1-(1-\mu)(1-\beta))\bar{\mathbf{m}}^{(k-1)} + (1-\mu)\bar{\mathbf{g}}^{(k-1)})) = (1-(1-\mu)(1-\beta))\sum_{k=0}^{t} -\eta\beta\left(\frac{1}{1-\beta}\bar{\mathbf{g}}^{(k-1)} - \bar{\mathbf{m}}^{(k-1)}\right) + \sum_{k=0}^{(t)} -\frac{\eta\beta}{1-\beta}(\bar{\mathbf{g}}^{(k)} - \bar{\mathbf{g}}^{(k-1)}) = (1-(1-\mu)(1-\beta))\mathbf{e}^{(t)} - \frac{\eta\beta}{1-\beta}\bar{\mathbf{g}}^{(t)}\,. \tag{10}$$

Taking expectation of $\|\mathbf{e}^{(t+1)}\|^2$:

$$\mathbb{E}\left\|\mathbf{e}^{(t+1)}\right\|^2 = \mathbb{E}\left\|(1-(1-\mu)(1-\beta))\mathbf{e}^{(t)} - \frac{\eta\beta}{1-\beta}\bar{\mathbf{g}}^{(t)}\right\|^2 \overset{a}{\leq} \mathbb{E}\left\|(1-(1-\mu)(1-\beta))\mathbf{e}^{(t)} - \frac{\eta\beta}{1-\beta}\mathbb{E}_t[\bar{\mathbf{g}}^{(t)}]\right\|^2 + \left(\frac{\eta^2\beta^2}{(1-\beta)^2}\right)(3\sigma^2 + 6L^2\rho^2) \leq (1-(1-\mu)(1-\beta))\mathbb{E}\left\|\mathbf{e}^{(t)}\right\|^2 + \frac{2\tilde{\eta}^2\beta^2}{(1-\beta)(1-\mu)}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i+\xi_i)\right\|^2 + 3\tilde{\eta}^2\beta^2\sigma^2 + 6\tilde{\eta}^2\beta^2L^2\rho^2\,. \tag{11}$$

(a) is the result of Lemma 1.

We now proceed to bound the consensus error.

Lemma 4 *Given Assumptions 1-3, the sequence of iterates generated by Q-SADDLe satisfy,*

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\right\|^2 \leq \frac{(1-\lambda/4)}{n}\mathbb{E}\left\|\mathbf{X}^{t} - \bar{\mathbf{X}}^{t}\right\|^2 + \frac{24\eta^2L^2\rho^2}{\lambda} + \frac{12\eta^2\delta^2}{\lambda} + 12\eta^2(1-\lambda)(\sigma^2+2L^2\rho^2) + \frac{6\eta^2\beta^2}{\lambda n}\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2\,. \tag{12}$$

Proof: We start by describing $\mathbf{X}^{t+1}$ and $\bar{\mathbf{X}}^{t+1}$ in terms of the update rule in equation 1:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\right\|^2 = \frac{1}{n}\mathbb{E}\|\mathbf{W}(\mathbf{X}^{(t)} - \eta(\beta\mathbf{M}^{(t)} + \widetilde{\mathbf{G}}^{(t)})) - (\bar{\mathbf{X}}^{(t)} - \eta(\beta\bar{\mathbf{M}}^{(t)} + \bar{\mathbf{G}}^{(t)}))\|^2 \overset{a}{\leq} \frac{1-\lambda}{n}\mathbb{E}\|(\mathbf{X}^{(t)} - \eta(\beta\mathbf{M}^{(t)} + \widetilde{\mathbf{G}}^{(t)})) - (\bar{\mathbf{X}}^{(t)} - \eta(\beta\bar{\mathbf{M}}^{(t)} + \bar{\mathbf{G}}^{(t)}))\|^2 \overset{b}{\leq} \frac{1-\lambda}{n}\mathbb{E}\|(\mathbf{X}^{(t)} - \eta(\beta\mathbf{M}^{(t)} + \mathbb{E}[\widetilde{\mathbf{G}}^{(t)}])) - (\bar{\mathbf{X}}^{(t)} - \eta(\beta\bar{\mathbf{M}}^{(t)} + \mathbb{E}[\bar{\mathbf{G}}^{(t)}]))\|^2 + 12\eta^2(1-\lambda)(\sigma^2+2L^2\rho^2) \leq \frac{(1-\lambda)(1+\lambda/2)}{n}\mathbb{E}\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\|^2 + \frac{6\eta^2\beta^2}{\lambda n}\mathbb{E}\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\|^2 + \frac{6\eta^2}{\lambda n}\underbrace{\mathbb{E}\|\mathbb{E}_t[\widetilde{\mathbf{G}}^{(t)}] - \mathbb{E}_t[\bar{\mathbf{G}}^{(t)}]\|^2}_{\star} + 12\eta^2(1-\lambda)(\sigma^2+2L^2\rho^2)\,. \tag{13}$$

(a) comes from Assumption 3 on the Mixing matrix. (b) results from $\widetilde{\mathbf{G}}^{(t)} = \mathbb{E}_t[\widetilde{\mathbf{G}}^{(t)}] + \widetilde{\mathbf{G}}^{(t)} - \mathbb{E}_t[\widetilde{\mathbf{G}}^{(t)}]$ and Lemma 1.

We first analyze $\star$:

$$\mathbb{E}\|\mathbb{E}_t[\widetilde{\mathbf{G}}^{(t)}] - \mathbb{E}_t[\bar{\mathbf{G}}^{(t)}]\|^2 = \sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(\mathbf{x}_i^{(t)} + \xi_i^{(t)}) \pm \nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f(\bar{\mathbf{x}}^{(t)})\|^2 \leq 2\sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(\mathbf{x}_i^{(t)} + \xi_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\|^2 + 2\sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2$$
$$\overset{a}{\leq} 2L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_i^{(t)} + \xi_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + 2n\delta^2$$
$$\overset{b}{\leq} 4L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + 4nL^2\rho^2 + 2n\delta^2 \tag{14}$$

(a) follows from Assumption 1, and (b) is the result of perturbation being bounded by the perturbation radius ρ.

Substituting the result of equation 14 in 13:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\right\|^2 \leq \frac{(1-\lambda/2)}{n}\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{6\eta^2\beta^2}{\lambda n}\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2 + \frac{24\eta^2L^2}{\lambda n}\left(\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2\right) + \frac{24\eta^2L^2\rho^2}{\lambda} + \frac{12\eta^2\delta^2}{\lambda} + 12\eta^2(1-\lambda)(\sigma^2 + 2L^2\rho^2) \tag{15}$$

The assumption that learning rate $\eta \leq \frac{\lambda}{10L}$ ensures that $24\eta^2L^2 \leq \lambda^2/4$. Modifying the above equation through this and rearranging the terms we have:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\right\|^2 \leq \frac{(1-\lambda/4)}{n}\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{6\eta^2\beta^2}{\lambda n}\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2 + \frac{24\eta^2L^2\rho^2}{\lambda} + \frac{12\eta^2\delta^2}{\lambda} + 12\eta^2(1-\lambda)(\sigma^2 + 2L^2\rho^2) \tag{16}$$

In the above bound on the consensus error, we have a momentum error term $\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2$. We present the following lemma to provide an upper bound on this error:

Lemma 5 *Given Assumptions 1-3, the sequence of iterates generated by Q-SADDLe for* $\frac{\beta}{1-\beta} \leq \frac{\lambda}{21}$,

$$\frac{6\eta^2\beta^2}{n\lambda(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{M}^{t+1} - \bar{\mathbf{M}}^{t+1}\right\|^2 \leq \left(\frac{6\eta^2\beta^2}{n\lambda(1-\mu)(1-\beta)} - \frac{6\eta^2\beta^2}{n\lambda}\right)\mathbb{E}\left\|(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)})\right\|^2 + \frac{\lambda}{8n}\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{\lambda\eta^2\delta^2}{8} + \left(\frac{3(1-\beta)}{(1-\mu)} + \frac{1}{2}\right)\frac{\lambda\eta^2L^2\rho^2}{4} + \frac{\lambda\eta^2\sigma^2(1-\beta)}{8(1-\mu)}.$$

Proof: Starting from the update (1), we have:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{M}^{(t+1)} - \bar{\mathbf{M}}^{(t+1)}\right\|^2 = \frac{1}{n}\mathbb{E}\|\left(\mu\mathbf{I} + (1-\mu)\beta\mathbf{W}\right)(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}) + (1-\mu)\mathbf{W}(\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}) + \frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)}\|^2$$
$$= \frac{1}{n}\mathbb{E}\|(\mu\mathbf{I} + (1-\mu)\beta\mathbf{W})(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}) + \frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)} + (1-\mu)\mathbf{W}(\mathbb{E}[\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}])\|^2 + \frac{1}{n}\mathbb{E}\|(1-\mu)\mathbf{W}\left(\widetilde{\mathbf{G}}^{(t)} - \mathbb{E}[\widetilde{\mathbf{G}}^{(t)}] - (\bar{\mathbf{G}}^{(t)} - \mathbb{E}[\bar{\mathbf{G}}^{(t)}])\right)\|^2$$
$$\overset{a}{\leq} \frac{1}{n}\mathbb{E}\|(\mu\mathbf{I} + (1-\mu)\beta\mathbf{W})(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}) + \frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)} + (1-\mu)\mathbf{W}(\mathbb{E}[\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}])\|^2 + 4(3\sigma^2 + 6L^2\rho^2)$$
$$\overset{b}{\leq} \frac{1}{n}\left(1 + \frac{(1-\mu)(1-\beta)}{1-(1-\mu)(1-\beta)}\right)\mathbb{E}\|(\mu\mathbf{I} + (1-\mu)\beta\mathbf{W})(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)})\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{1}{n}\left(1 + \frac{1-(1-\mu)(1-\beta)}{(1-\mu)(1-\beta)}\right)\mathbb{E}\|\frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)} + (1-\mu)\mathbf{W}(\mathbb{E}[\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}])\|^2. \tag{17}$$

(a) follows from Lemma 1, and (b) follows from the inequality $\|x_i + x_j\|^2 \leq (1+a)\|x_i\|^2 + (1+\frac{1}{a})\|x_j\|^2$ for any $a > 0$. Since $\mathbf{W} < \mathbf{I}$, we have $(\mu\mathbf{I} + (1-\mu)\beta\mathbf{W}) < (\mu + (1-\mu)\beta)\mathbf{I} = (1-(1-\beta)(1-\mu))\mathbf{I}$. Further, we have $\mathbf{I} - \mathbf{W} < 2\mathbf{I}$ [8]. With these observations:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{M}^{(t+1)} - \bar{\mathbf{M}}^{(t+1)}\right\|^2 \leq \frac{1}{n}\left(1 + \frac{(1-\mu)(1-\beta)}{1-(1-\mu)(1-\beta)}\right)\mathbb{E}\left\|(1-(1-\mu)(1-\beta))(\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)})\right\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{1}{n}\left(1 + \frac{1-(1-\mu)(1-\beta)}{(1-\mu)(1-\beta)}\right)\mathbb{E}\|\frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)} + (1-\mu)\mathbf{W}\left(\mathbb{E}\left[\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}\right]\right)\|^2$$
$$\leq \frac{1}{n}(1-(1-\mu)(1-\beta))\,\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{1}{(1-\mu)(1-\beta)n}\mathbb{E}\|\frac{1-\mu}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{(t)} + (1-\mu)\mathbf{W}(\mathbb{E}[\widetilde{\mathbf{G}}^{(t)} - \bar{\mathbf{G}}^{(t)}])\|^2 \leq \frac{1}{n}(1-(1-\mu)(1-\beta))\,\mathbb{E}\left\|\mathbf{M}^{(t)} - \bar{\mathbf{M}}^{(t)}\right\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{4(1-\mu)}{(1-\beta)n\eta^2}\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{2(1-\mu)}{(1-\beta)n}\mathbb{E}\left\|\mathbb{E}\left[\widetilde{\mathbf{G}}^{(t)}\right] - \mathbb{E}\left[\bar{\mathbf{G}}^{(t)}\right]\right\|^2. \tag{18}$$

Substituting equation 14 in the above equation:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{M}^{(t+1)}-\bar{\mathbf{M}}^{(t+1)}\right\|^2 \leq \frac{1}{n}\left(1-(1-\mu)(1-\beta)\right)$$
$$\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{4(1-\mu)}{(1-\beta)n\eta^2}$$
$$\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{8(1-\mu)L^2}{(1-\beta)n}\left(\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2\right)$$
$$+\frac{8(1-\mu)L^2\rho^2}{(1-\beta)} + \frac{4\delta^2(1-\mu)}{(1-\beta)} \leq \frac{1}{n}\left(1-(1-\mu)(1-\beta)\right)$$
$$\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|^2 + 12\sigma^2 + 24L^2\rho^2 + \frac{4(1-\mu)(1+2\eta^2L^2)}{(1-\beta)n\eta^2}$$
$$\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{8(1-\mu)L^2\rho^2}{(1-\beta)} + \frac{4\delta^2(1-\mu)}{(1-\beta)} \tag{19}$$

Multiplying both sides by $\frac{6\eta^2\beta^2}{\lambda(1-\mu)(1-\beta)}$ yields

$$\frac{6\eta^2\beta^2}{\lambda n(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{M}^{(t+1)}-\bar{\mathbf{M}}^{(t+1)}\right\|^2 \leq \frac{6\eta^2\beta^2}{\lambda n}$$
$$\left(\frac{1}{(1-\mu)(1-\beta)}-1\right)\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|^2$$
$$+\left(\frac{24\beta^2(1+2\eta^2L^2)}{n\lambda(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{72\eta^2\beta^2\sigma^2}{\lambda(1-\mu)(1-\beta)}$$
$$+\frac{144\eta^2\beta^2L^2\rho^2}{\lambda(1-\mu)(1-\beta)} + \frac{48L^2\rho^2\eta^2\beta^2}{\lambda(1-\beta)^2} + \frac{24\eta^2\beta^2\delta^2}{\lambda(1-\beta)^2}$$
$$\overset{a}{\leq} \frac{6\eta^2\beta^2}{\lambda n}\left(\frac{1}{(1-\mu)(1-\beta)}-1\right)\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|^2 + \frac{\lambda}{8n}$$
$$\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|^2 + \frac{\lambda\eta^2\sigma^2(1-\beta)}{6(1-\mu)} + \left(\frac{(1-\beta)}{3(1-\mu)}+\frac{1}{9}\right)$$
$$\lambda\eta^2L^2\rho^2 + \frac{\lambda\eta^2\delta^2}{18} \tag{20}$$

(a) follows from our assumption that the momentum parameter satisfies $\frac{\beta}{1-\beta} \leq \frac{\lambda}{21}$ and $\eta \leq \frac{1}{7L}$.

Adding the results of Lemma 4 and 5 and simplifying the coefficients, we describe the progress made in each gossip averaging consensus round:

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{t+1}-\bar{\mathbf{X}}^{t+1}\right\|^2 + \frac{6\eta^2\beta^2}{n\lambda(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{M}^{t+1}-\bar{\mathbf{M}}^{t+1}\right\|^2$$
$$\leq \frac{1-\lambda/8}{n}\mathbb{E}\left\|\mathbf{X}^{t}-\bar{\mathbf{X}}^{t}\right\|^2 + \frac{6\eta^2\beta^2}{n\lambda(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{M}^{t}-\bar{\mathbf{M}}^{t}\right\|^2$$
$$+\frac{13\eta^2\delta^2}{\lambda} + \frac{12\eta^2\sigma^2(2-\beta-\mu)}{(1-\mu)\lambda} + \frac{49\eta^2L^2\rho^2(2-\beta-\mu)}{(1-\mu)\lambda} \tag{21}$$

1.1. Proof for Theorem 1

We start with the following property for a L-smooth function $f(\mathbf{x})$:

$$\mathbb{E}f(\mathbf{z}^{(t+1)}) \leq \mathbb{E}f(\mathbf{z}^{(t)}) + \mathbb{E}\left\langle\nabla f(\mathbf{z}^{(t)}), \mathbf{z}^{(t+1)}-\mathbf{z}^{(t)}\right\rangle +$$
$$\frac{L}{2}\mathbb{E}\left\|\mathbf{z}^{(t+1)}-\mathbf{z}^{(t)}\right\|^2 = \mathbb{E}f(\mathbf{z}^{(t)}) -$$
$$\underbrace{\tilde{\eta}\,\mathbb{E}\left\langle\nabla f(\mathbf{z}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\rangle}_{I} + \frac{L}{2}\underbrace{\mathbb{E}\left\|\mathbf{z}^{(t+1)}-\mathbf{z}^{(t)}\right\|^2}_{II} \tag{22}$$

We first focus on finding an upper bound for I:

$$I: \frac{1}{2}\left(\mathbb{E}\left\|\nabla f(\mathbf{z}^t)\right\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2\right)$$
$$-\frac{1}{2}\underbrace{\left(\mathbb{E}\left\|\nabla f(\mathbf{z}^t)) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2\right)}_{\star} \tag{23}$$

To bound $\star$:

$$\star: \left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}^t) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2\right)$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2 \tag{24}$$

Substituting equation 24 in 23:

$$I \geq \frac{1}{2}\left(\mathbb{E}\left\|\nabla f(\mathbf{z}^t)\right\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2\right) -$$
$$\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2 \tag{25}$$

Now, we find an upper bound for II:

$$\mathbb{E}\left\|\mathbf{z}^{(t+1)}-\mathbf{z}^{(t)}\right\|^2 = \tilde{\eta}^2\mathbb{E}\left\|\bar{\mathbf{g}}\right\|^2$$
$$\overset{a}{\leq} \tilde{\eta}^2\left(\frac{3\sigma^2}{n} + \frac{6L^2\rho^2}{n} + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2\right) \tag{26}$$

Here, (a) is the result of Lemma 2. Putting equation 25

and 26 in 22:

$$\mathbb{E}f(\mathbf{z}^{(t+1)}) \leq \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\mathbf{z}^t)\right\|^2 - \frac{\tilde{\eta}}{2}\mathbb{E}\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\|^2 + \frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\|^2 + \frac{\tilde{\eta}^2 L}{2}(\frac{3\sigma^2}{n} + \frac{6L^2\rho^2}{n} + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2) \tag{27}$$

Rearranging the above terms we get:

$$\mathbb{E}f(\mathbf{z}^{(t+1)}) \leq \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\mathbf{z}^t)\right\|^2 + \left(\frac{\tilde{\eta}^2 L}{2} - \frac{\tilde{\eta}}{2}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\|^2 + \frac{3\tilde{\eta}^2 L^3\rho^2}{n} + \frac{3\tilde{\eta}^2 L\sigma^2}{2n} \leq \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 \underbrace{+\frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\mathbf{z}^t) - \nabla f(\bar{\mathbf{x}}^t)\right\|^2}_{\star} \underbrace{+\frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2}_{\star} + \left(\frac{\tilde{\eta}^2 L}{2} - \frac{\tilde{\eta}}{2}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{3\tilde{\eta}^2 L^3\rho^2}{n} + \frac{3\tilde{\eta}^2 L\sigma^2}{2n} \tag{28}$$

Now we simplify $\star$:

$$\star : \frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\|^2 \leq \frac{\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\tilde{\eta}}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\tilde{\eta}}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\|^2 = \frac{3\tilde{\eta}}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\tilde{\eta}}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 \tag{29}$$

Putting this back into equation 28:

$$\mathbb{E}f(\mathbf{z}^{(t+1)}) \leq \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 + \left(\frac{\tilde{\eta}^2 L}{2} - \frac{\tilde{\eta}}{2}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{3\tilde{\eta}}{2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}^t) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\tilde{\eta}}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{3\tilde{\eta}^2 L^3\rho^2}{n} + \frac{3\tilde{\eta}^2 L\sigma^2}{2n} \tag{30}$$

Using our assumption $\tilde{\eta} \leq \frac{1}{4L}$ and Assumption 1, we have:

$$\mathbb{E}f(\mathbf{z}^{(t+1)}) \leq \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{3\tilde{\eta}L^2}{2}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}^t - \bar{\mathbf{x}}^t\right\|^2 + \frac{\tilde{\eta}L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t - \xi_i^t\right\|^2 + \frac{3\tilde{\eta}^2 L^3\rho^2}{n} + \frac{3\tilde{\eta}^2 L\sigma^2}{2n} \overset{a}{\leq} \mathbb{E}f(\mathbf{z}^{(t)}) - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 - \frac{\tilde{\eta}}{4}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t + \xi_i^t)\right\|^2 + \frac{3\tilde{\eta}L^2}{2}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}^t - \bar{\mathbf{x}}^t\right\|^2 + \frac{2\tilde{\eta}L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 + 2\tilde{\eta}L^2\rho^2 + \frac{3\tilde{\eta}^2 L^3\rho^2}{n} + \frac{3\tilde{\eta}^2 L\sigma^2}{2n} \tag{31}$$

(a) follows from the perturbation ξ_i being bounded by the perturbation radius ρ. Now we see that the terms $\|\mathbf{z}^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2$ and $\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2$, which we bound in Lemma 3 and 4 respectively, appear in the above equation. We start by Lemma 3, and scale both sides by $\frac{3L^2\tilde{\eta}}{2(1-\mu)(1-\beta)}$:

$$\frac{3L^2\tilde{\eta}}{2(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{e}^{(t+1)}\right\|^2 \leq \left(\frac{3L^2\tilde{\eta}}{2(1-\mu)(1-\beta)} - \frac{3L^2\tilde{\eta}}{2}\right)\mathbb{E}\left\|\mathbf{e}^{(t)}\right\|^2 + \frac{3L^2\tilde{\eta}^3\beta^2}{(1-\beta)^2(1-\mu)^2}\mathbb{E}\left\|\mathbb{E}_t[\bar{\mathbf{g}}^{(t)}]\right\|^2 + \frac{9L^2\tilde{\eta}^3\beta^2\sigma^2}{2(1-\mu)(1-\beta)} + \frac{9L^4\tilde{\eta}^3\beta^2\rho^2}{(1-\mu)(1-\beta)} \tag{32}$$

Next, we take the total consensus change from equation

21, and scale it with $\frac{16L^2\tilde{\eta}}{\lambda}$:

$$\frac{16L^2\tilde{\eta}}{\lambda n}\mathbb{E}\left\|\mathbf{X}^{t+1}-\bar{\mathbf{X}}^{t+1}\right\|^2+\frac{96\tilde{\eta}^3\beta^2(1-\beta)}{n\lambda^2(1-\mu)}\mathbb{E}\|\mathbf{M}^{t+1}-\bar{\mathbf{M}}^{t+1}\|^2\leq\frac{16L^2\tilde{\eta}}{\lambda n}\mathbb{E}\left\|\mathbf{X}^{t}-\bar{\mathbf{X}}^{t}\right\|^2+\frac{96\tilde{\eta}^3\beta^2(1-\beta)}{n\lambda^2(1-\mu)}\mathbb{E}\left\|\mathbf{M}^{t}-\bar{\mathbf{M}}^{t}\right\|^2-\frac{2L^2\tilde{\eta}}{n}\mathbb{E}\left\|\mathbf{X}^{t}-\bar{\mathbf{X}}^{t}\right\|^2+\frac{208L^2\tilde{\eta}^3(1-\beta)^2\delta^2}{\lambda^2}+\frac{192L^2\tilde{\eta}^3(2-\beta-\mu)(1-\beta)^2\sigma^2}{(1-\mu)\lambda^2}+\frac{784L^4\tilde{\eta}^3(1-\beta)^2(2-\beta-\mu)\rho^2}{(1-\mu)\lambda^2} \tag{33}$$

Through equation 32 and 33, we define another sequence $\phi^t\geq 0$ such that $\phi^0=\mathbb{E}[f(\bar{\mathbf{x}}^0)-f^*]$:

$$\phi^t:\frac{16L^2\tilde{\eta}}{\lambda n}\mathbb{E}\left\|\mathbf{X}^{t}-\bar{\mathbf{X}}^{t}\right\|^2+\frac{96L^2\tilde{\eta}^3\beta^2(1-\beta)}{n\lambda^2(1-\mu)}\mathbb{E}\left\|\mathbf{M}^{t}-\bar{\mathbf{M}}^{t}\right\|^2+\frac{3L^2\tilde{\eta}}{2(1-\mu)(1-\beta)}\mathbb{E}\left\|\mathbf{e}^{(t)}\right\|^2+\mathbb{E}[f(\bar{\mathbf{x}}^t)-f^\star]$$

Adding the right hand sides of equation 31, 32 and 33, and bounding ϕ^{t+1} in terms of ϕ^t:

$$\phi^{t+1}\leq\phi^t-\frac{\tilde{\eta}}{4}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2+\left(\frac{3L^2\tilde{\eta}^3\beta^2}{(1-\beta)^2(1-\mu)^2}-\frac{\tilde{\eta}}{4}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2+2\tilde{\eta}L^2\rho^2+\frac{3\tilde{\eta}^2L^3\rho^2}{n}+\frac{3\tilde{\eta}^2L\sigma^2}{2n}+\frac{9L^2\tilde{\eta}^3\beta^2\sigma^2}{2(1-\mu)(1-\beta)}+\frac{9L^4\tilde{\eta}^3\beta^2\rho^2}{(1-\mu)(1-\beta)}+\frac{208L^2\tilde{\eta}^3(1-\beta)^2\delta^2}{\lambda^2}+\frac{192L^2\tilde{\eta}^3(2-\beta-\mu)(1-\beta)^2\sigma^2}{(1-\mu)\lambda^2}+\frac{784L^4\tilde{\eta}^3(1-\beta)^2(2-\beta-\mu)\rho^2}{(1-\mu)\lambda^2} \tag{34}$$

Simplifying the above equation by rearranging terms and approximating some coefficients:

$$\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\leq\frac{4}{\tilde{\eta}}(\phi^t-\phi^{t+1})+\left(\frac{12L^2\tilde{\eta}^2\beta^2}{(1-\beta)^2(1-\mu)^2}-1\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{x}_i^t+\xi_i^t)\right\|^2+\underbrace{\frac{6\tilde{\eta}L}{n}+18\tilde{\eta}^2L^2C_2+\frac{768\tilde{\eta}^2L^2C_1}{\lambda^2}}_{C_\sigma}\sigma^2+\underbrace{\frac{832L^2\tilde{\eta}^2(1-\beta)^2}{\lambda^2}}_{C_\delta}\delta^2+\underbrace{8L^2+\frac{12\tilde{\eta}L^3}{n}+36\tilde{\eta}^2L^4C_2+\frac{3136L^4\tilde{\eta}^2C_1}{\lambda^2}}_{C_\rho}\rho^2 \tag{35}$$

Here, $C_1=\frac{(2-\beta-\mu)(1-\beta)^2}{(1-\mu)}$ and $C_2=\frac{\beta^2}{(1-\mu)(1-\beta)}$.

For $\frac{12L^2\tilde{\eta}^2\beta^2}{(1-\beta)^2(1-\mu)^2}-1\leq 0$:

$$\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\leq\frac{4}{\tilde{\eta}}(\phi^t-\phi^{t+1})+C_\sigma\sigma^2+C_\delta\delta^2+C_\rho\rho^2 \tag{36}$$

Averaging over T, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\leq\frac{4}{\tilde{\eta}T}(f(\bar{\mathbf{x}}^0)-f^*)+C_\sigma\sigma^2+C_\delta\delta^2+C_\rho\rho^2 \tag{37}$$

1.2. Proof for Corollary 2

To find the convergence rate with a learning rate $\eta=\mathcal{O}\left(\sqrt{\frac{n}{T}}\right)$ and perturbation radius $\rho=\mathcal{O}\left(\sqrt{\frac{1}{T}}\right)$, we find the order of all the terms in equation 37:

- $\frac{4}{\tilde{\eta}T}(f(\bar{\mathbf{x}}^0)-f^*)=\mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$
- $C_\sigma\sigma^2=\mathcal{O}\left(\frac{\eta}{n}+\eta^2\right)=\mathcal{O}\left(\frac{1}{\sqrt{nT}}+\frac{n}{T}\right)$
- $C_\delta\delta^2=\mathcal{O}\left(\eta^2\right)=\mathcal{O}\left(\frac{n}{T}\right)$
- $C_\rho\rho^2=\mathcal{O}\left(\rho^2+\frac{\eta\rho^2}{n}+\eta^2\rho^2\right)=\mathcal{O}\left(\frac{1}{T}+\frac{1}{n^{1/2}T^{3/2}}+\frac{n}{T^2}\right)$

Adding all the terms and ignoring n in higher order terms:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\leq\mathcal{O}\left(\frac{1}{\sqrt{nT}}+\frac{1}{T}+\frac{1}{T^{3/2}}+\frac{1}{T^2}\right) \tag{38}$$

This implies that when T is sufficiently large, Q-SADDLe converges at the rate of $\mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$.

1.3. Condition on Learning Rate η and Momentum Coefficient β

In Lemma 4, we assume $\eta\leq\frac{\lambda}{10L}$ and in Lemma 5, we assume $\eta\leq\frac{1}{7L}$. Combining both bounds results in $\eta\leq\min(\frac{1}{7L},\frac{\lambda}{10L})\leq\min(\frac{1}{7L},\frac{\lambda}{7L})\leq\frac{\lambda}{7L}$.In Theorem 1 proof, we assume $\eta\leq\frac{(1-\beta)}{4L}$ to simplify equation 30. Further to

simplify equation 35, we have the following upper bound on η:

$$\begin{aligned} &\frac{12L^2\tilde{\eta}^2\beta^2}{(1-\beta)^2(1-\mu)^2} - 1 \leq 0 \\ &12L^2\tilde{\eta}^2\beta^2 - (1-\beta)^2(1-\mu)^2 \leq 0 \\ &\eta \leq \frac{(1-\beta)^2(1-\mu)}{\sqrt{12}L\beta} \end{aligned} \tag{39}$$

Combining all the above mentioned bounds, we can describe $\eta \leq \min\left(\frac{\lambda}{7L}, \frac{1-\beta}{4L}, \frac{(1-\beta)^2(1-\mu)}{\sqrt{12}L\beta}\right)$.

Similarly, for momentum coefficient β, we assume $\frac{\beta}{1-\beta} \leq \frac{\lambda}{21}$ in Lemma 5. Note that we don't abide by these constraints and still achieve competitive performance for our results in Section 6 (main paper) and Section 3 (Supplementary).

2. Algorithmic Details

2.1. Background

To highlight that SADDLe can improve the generalization and communication efficiency of existing decentralized algorithms, we choose two state-of-the-art techniques for our evaluation: Quasi-Global Momentum (QGM) [8] and Neighborhood Gradient Mean (NGM) [1]. QGM improves the performance of D-PSGD [7] without introducing any extra communication. However, as shown in our results in Section 6, it performs poorly with extreme data heterogeneity. To achieve competitive performance with higher degrees of non-IIDness, NGM proposes to boost the performance through cross-gradients, which require 2x communication (i.e., an extra round of communication) as compared to D-PSGD [7].

Quasi-Global Momentum (QGM): The authors in QGM [8] show that local momentum acceleration is hindered by data heterogeneity. Inspired by this, QGM updates the momentum buffer by computing the difference between two consecutive models $\mathbf{x}_i^{t+1}$ and $\mathbf{x}_i^t$ to approximate the global optimization direction locally. The following equation illustrates the update rule for QGM:

$$\begin{aligned} &\text{QGM: } \mathbf{x}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} w_{ij}[\mathbf{x}_j^t - \eta(\mathbf{g}_j^t + \beta\mathbf{m}_j^{t-1})] \\ &\text{where, } \mathbf{m}_i^t = \mu\mathbf{m}_i^{t-1} + (1-\mu)\frac{\mathbf{x}_i^t - \mathbf{x}_i^{(t+1)}}{\eta}. \end{aligned} \tag{40}$$

Neighborhood Gradient Mean (NGM): NGM [1] modifies the local gradient update with the aid of self and cross-gradients. The self-gradients are computed at each agent through its model parameters and the local dataset. The data variant cross-gradients are derivatives of the local model with respect to the dataset of neighbors. These gradients are obtained through an additional round of communication. The update rule for NGM is shown in equation 41, where each gradient update $\mathbf{g}_j^t$ is a weighted average of the self and received cross-gradients.

$$\begin{aligned} &\text{NGM: } \mathbf{x}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} w_{ij}\mathbf{x}_j^t - \eta\mathbf{g}_j^t; \\ &\mathbf{g}_j^t = \sum_{j\in\mathcal{N}(i)} w_{ij}\nabla F_j(\mathbf{x}_i^t; d_j^t). \end{aligned} \tag{41}$$

Algorithm 1 NGM vs N-SADDLe

Input: Each agent $i \in [1, n]$ initializes model weights $\mathbf{x}_i$, step size η, momentum coefficient β, averaging rate γ, mixing matrix $\mathbf{W} = [w_{ij}]_{i,j\in[1,n]}$, and I_{ij} are elements of $n \times n$ identity matrix, $\mathcal{N}(i)$ represents neighbors of i including itself.

procedure TRAIN( ) $\forall i$
1. **for** t = $1, 2, \ldots, T$ **do**
2. $d_i^t \sim D^i$
3. $\mathbf{g}_{ii}^t = \nabla F_i(\mathbf{x}_i^t; d_i^t)$
4. $\tilde{\mathbf{g}}_{ii}^t = \nabla F_i(\mathbf{x}_i^t + \xi(\mathbf{x}_i^t); d_i^t)$, where $\xi(\mathbf{x}_i^t) = \rho\frac{\mathbf{g}_{ii}^t}{\|\mathbf{g}_{ii}^t\|}$
5. SENDRECEIVE($\mathbf{x}_i^t$)
6. **for** each neighbor $j \in \{\mathcal{N}(i) - i\}$ **do**
7. $\mathbf{g}_{ji}^t = \nabla F_i(\mathbf{x}_j^t; d_i^t)$
8. $\tilde{\mathbf{g}}_{ji}^t = \nabla F_i(\mathbf{x}_j^t + \xi(\mathbf{x}_j^t); d_i^t)$, where $\xi(\mathbf{x}_j^t) = \rho\frac{\mathbf{g}_{ji}^t}{\|\mathbf{g}_{ji}^t\|}$
9. SENDRECEIVE $(\mathbf{g}_{ji}^t)$ $(\tilde{\mathbf{g}}_{ji}^t)$
10. **end**
11. $\mathbf{g}_i^t = \sum_{j\in\mathcal{N}(i)} w_{ij}\mathbf{g}_{ij}^t$
12. $\mathbf{m}_i^t = \beta\mathbf{m}_i^{(t-1)} + \mathbf{g}_i^t$
13. $\tilde{\mathbf{g}}_i^t = \sum_{j\in\mathcal{N}(i)} w_{ij}\tilde{\mathbf{g}}_{ij}^t$
14. $\mathbf{m}_i^t = \beta\mathbf{m}_i^{(t-1)} + \tilde{\mathbf{g}}_i^t$
15. $\mathbf{x}_i^{(t+1/2)} = \mathbf{x}_i^t - \eta\mathbf{m}_i^t$
16. $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+1/2)} + \gamma\sum_{j\in\mathcal{N}(i)}(w_{ij} - I_{ij})\mathbf{x}_j^t$
17. **end**

return $\mathbf{x}_i^T$

2.2. N-SADDLe and Comp N-SADDLe

Algorithm 1 highlights the difference between NGM and N-SADDLe. Specifically, N-SADDLe computes SAM-based gradient updates for self and cross gradients (lines 4 and 8). Similarly, please refer to Algorithm 2 to understand the difference between the compressed versions of NGM and N-SADDLe (i.e., Comp NGM and Comp N-SADDLe). The error between the original gradients and their compressed version is added as feedback to the gradients before

Algorithm 2 Comp NGM vs Comp N-SADDLe

Input: Each agent i initializes model weights $\mathbf{x}_i$, step size η, averaging rate γ, mixing matrix $\mathbf{W} = [w_{ij}]_{i,j\in[1,n]}$, $Q(.)$ is the compression operator, $\mathcal{N}(i)$ represents neighbors of i.

procedure TRAIN( ) $\forall i$

1. **for** t=$1, 2, \ldots, T$ **do**
2. $d_i^t \sim D_i$
3. $\mathbf{g}_{ii}^t = \nabla F_i(\mathbf{x}_i^t; d_i^t)$
4. $\widetilde{\mathbf{g}}_{ii}^t = \nabla F_i(\mathbf{x}_i^t + \xi(\mathbf{x}_i^t); d_i^t)$, where $\xi(\mathbf{x}_i^t) = \rho \frac{\mathbf{g}_{ii}^t}{\|\mathbf{g}_{ii}^t\|}$
5. $\mathbf{p}_{ii}^t = \mathbf{g}_{ii}^t + \mathbf{e}_{ii}^t$
6. $\mathbf{p}_{ii}^t = \widetilde{\mathbf{g}}_{ii}^t + \mathbf{e}_{ii}^t$
7. $\delta_{ii}^t = Q(\mathbf{p}_{ii}^t)$
8. $\mathbf{e}_{ii}^{t+1} = \mathbf{p}_{ii}^t - \delta_{ii}^t$
9. SENDRECEIVE($\mathbf{x}_i^t$)
10. **for** each neighbor $j \in \{N(i) - i\}$ **do**
11. $\mathbf{g}_{ji}^t = \nabla F_i(\mathbf{x}_j^t; d_i^t)$
12. $\widetilde{\mathbf{g}}_{ji}^t = \nabla F_i(\mathbf{x}_j^t + \xi(\mathbf{x}_j^t); d_i^t)$, where $\xi(\mathbf{x}_j^t) = \rho \frac{\mathbf{g}_{ji}^t}{\|\mathbf{g}_{ji}^t\|}$
13. $\mathbf{p}_{ji}^t = \mathbf{g}_{ji}^t + \mathbf{e}_{ji}^t$
14. $\mathbf{p}_{ji}^t = \widetilde{\mathbf{g}}_{ji}^t + \mathbf{e}_{ji}^t$
15. $\delta_{ji}^t = Q(\mathbf{p}_{ji}^t)$
16. $\mathbf{e}_{ji}^{t+1} = \mathbf{p}_{ji}^t - \delta_{ji}^t$
17. SENDRECEIVE(δ_{ji}^t)
18. **end**
19. **end**
20. $\mathbf{g}_i^t = \sum_{j\in\mathcal{N}(i)} w_{ij}\delta_{ij}^t$
21. $\mathbf{m}_i^t = \beta \mathbf{m}_i^{(t-1)} + \mathbf{g}_i^t$
22. $\mathbf{x}_i^{(t+1/2)} = \mathbf{x}_i^t - \eta \mathbf{m}_i^t$
23. $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+1/2)} + \gamma \sum_{j\in\mathcal{N}(i)} (w_{ij} - I_{ij}) \mathbf{x}_j^t$
24. **end**

return $\mathbf{x}_i^T$

compressing them in the next iteration (lines 5, 6, 13, and 14 in Algorithm 2).

3. Additional Results

3.1. SADDLe with DPSGD

A natural question that arises is, does SADDLe improve the performance of DPSGD [7] in the presence of data heterogeneity? Note that DPSGD assumes the data distribution to be IID and has been shown to incur significant performance drop with non-IID data [8]. Algorithm 3 shows the difference between DPSGD and D-SADDLe, a version incorporating SAM-based updates within DPSGD. D-SADDLe leads to an average improvement of 10% and 5.4% over DPSGD for CIFAR-10 and CIFAR-100 datasets, respectively, as shown in Table 1.

Algorithm 3 DPSGD vs D-SADDLe

Input: Each agent $i \in [1, n]$ initializes model weights $\mathbf{x}_i^{(0)}$, learning rate η, perturbation radius ρ, and mixing matrix $\mathbf{W} = [w_{ij}]_{i,j\in[1,n]}$, $\mathcal{N}(i)$ represents neighbors of i.

procedure TRAIN( ) $\forall i$

1. **for** t=$0, 1, \ldots, T-1$ **do**
2. $d_i^k \sim D_i$
3. $\mathbf{g}_i^t = \nabla F_i(d_i^t; \mathbf{x}_i^t)$
4. $\widetilde{\mathbf{g}}_i^t = \nabla F_i(\mathbf{x}_i^t + \xi(\mathbf{x}_i^t); d_i^t)$, where $\xi(\mathbf{x}_i^t) = \rho \frac{\mathbf{g}_i^t}{\|\mathbf{g}_i^t\|}$
5. $\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^t - \eta \mathbf{g}_i^t$
6. $\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^t - \eta \widetilde{\mathbf{g}}_i^t$
7. SENDRECEIVE($\mathbf{x}_i^{(t+\frac{1}{2})}$)
8. $\mathbf{x}_i^{(t+1)} = \sum_{j\in\mathcal{N}_i} w_{ij} \mathbf{x}_j^{t+\frac{1}{2}}$

return

3.2. Results with Top-k Sparsification

We present results for QGM and Q-SADDLe with Top-30% Sparsification-based compressor in Table 2. Note that Top-30% implies that only the top 30% of model updates for each layer are communicated to the peers. As shown in Table 2, QGM performs poorly in the presence of compression, with a significant drop of $\sim 5 - 57\%$, and the training even diverges for some cases. In contrast, Q-SADDLe is much more stable, with an accuracy drop of $\sim 0.6 - 18.5\%$ with compression.

3.3. Compression Error and Gradient Norms for N-SADDLe

Recall that the expectation of compression error for a compression operator $Q(.)$ has the following upper bound:

$$\mathbb{E}_Q \|Q(\theta) - \theta\|^2 \leq (1 - \zeta)\|\theta\|^2, \text{ where } \zeta > 0 \qquad (42)$$

For NGM and N-SADDLe, θ corresponds to the gradients $\mathbf{g}_i$ and $\widetilde{\mathbf{g}}_i$ respectively. In Figure 1, we compare the compression error ($\|Q(\theta) - \theta\|$) and gradient norms for NGM and N-SADDLe with a 1-bit Sign SGD-based compression scheme. Clearly, N-SADDLe leads to a lower compression error, as well as lower gradient norms throughout the training. Here, we plot the sum of layer-wise compression errors and the sum of gradient norms for each layer in the ResNet-20 model. Like Q-SADDLe, the bound in Equation 42 is tighter for N-SADDLe than NGM.

Table 1. Test accuracy of DPSGD and D-SADDLe evaluated on CIFAR-10 and CIFAR-100 over ResNet-20, distributed with different degrees of heterogeneity over ring topologies.

Agents	Method	CIFAR-10		CIFAR-100	
		$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.001$
5	DPSGD (IID)	91.05 ± 0.06		64.47 ± 0.48	
	DPSGD	82.15 ± 3.25	80.54 ± 4.36	47.30 ± 4.92	45.54 ± 0.71
	D-SADDLe (ours)	$\mathbf{85.38 \pm 0.84}$	$\mathbf{84.94 \pm 0.31}$	$\mathbf{54.35 \pm 0.48}$	$\mathbf{54.30 \pm 0.50}$
10	DPSGD (IID)	90.46 ± 0.33		62.73 ± 1.03	
	DPSGD	49.17 ± 17.38	40.74 ± 2.62	31.66 ± 0.84	29.79 ± 1.30
	D-SADDLe (ours)	$\mathbf{64.18 \pm 5.63}$	$\mathbf{61.30 \pm 0.79}$	$\mathbf{37.49 \pm 0.59}$	$\mathbf{35.31 \pm 0.77}$
20	DPSGD (IID)	89.46 ± 0.02		59.61 ± 1.15	
	DPSGD	40.49 ± 3.06	36.13 ± 5.67	24.45 ± 0.51	21.58 ± 1.00
	D-SADDLe (ours)	$\mathbf{52.14 \pm 2.02}$	$\mathbf{47.06 \pm 2.35}$	$\mathbf{26.39 \pm 0.17}$	$\mathbf{24.92 \pm 0.62}$

Table 2. Test accuracy (Acc) and accuracy drop (Drop) of QGM and Q-SADDLe with Sparsification (top-30%) based compression evaluated on CIFAR-10 distributed over ring topologies. $\star$ indicates 1 out of 3 runs diverged.

Agents	Comp	Method	CIFAR-10			
			$\alpha = 0.01$		$\alpha = 0.001$	
			Acc (%)	Drop(%)	Acc (%)	Drop(%)
5	✓	QGM	83.58 ± 2.96	4.86	67.04 ± 9.76	21.68
	✓	*Q-SADDLe (ours)*	$\mathbf{90.01 \pm 0.38}$	0.65	$\mathbf{89.49 \pm 0.38}$	1.18
10	✓	QGM	$52.23 \star$	25.18	23.00 ± 1.96	56.48
	✓	*Q-SADDLe (ours)*	$\mathbf{80.34 \pm 5.56}$	7.38	$\mathbf{71.01 \pm 3.75}$	15.32
20	✓	QGM	62.90 ± 5.89	9.3	32.92 ± 9.25	29.56
	✓	*Q-SADDLe (ours)*	$\mathbf{71.96 \pm 2.51}$	6.45	$\mathbf{64.31 \pm 2.14}$	18.50

Table 3. Test accuracy (Acc) and accuracy drop (Drop) of NGM and N-SADDLe with 2-bit quantization compression scheme [2] evaluated on CIFAR-10, with $\alpha = 0.01, 0.001$.

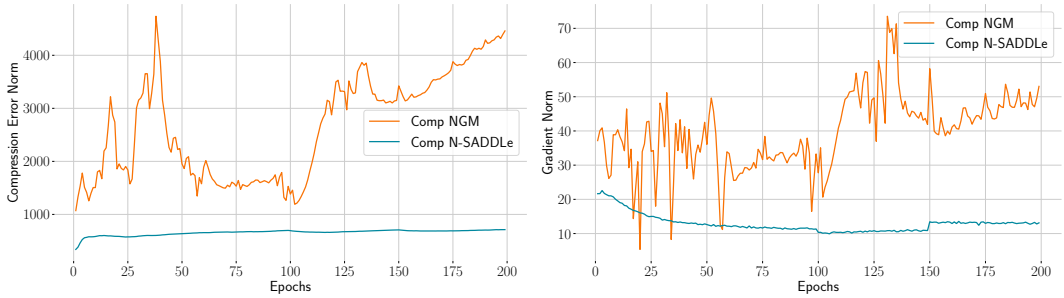
Agents	Comp	Method	CIFAR-10 (ResNet-20)			
			$\alpha = 0.01$		$\alpha = 0.001$	
			Acc (%)	Drop(%)	Acc (%)	Drop(%)
5	✓	NGM	87.38 ± 2.01	3.49	87.27 ± 0.56	3.46
	✓	*N-SADDLe (ours)*	$\mathbf{91.35 \pm 0.17}$	0.61	$\mathbf{91.18 \pm 0.25}$	0.51
10	✓	NGM	79.89 ± 8.74	5.19	79.20 ± 3.05	4.23
	✓	*N-SADDLe (ours)*	$\mathbf{87.25 \pm 1.65}$	1.18	$\mathbf{85.70 \pm 1.15}$	1.59
20	✓	NGM	81.87 ± 1.17	2.97	76.68 ± 0.95	6.90
	✓	*N-SADDLe (ours)*	$\mathbf{84.25 \pm 0.17}$	2.01	$\mathbf{85.09 \pm 0.31}$	1.52

3.4. Stochastic Quantization for NGM and N-SADDLe

The main paper uses Sign SGD [4] compression scheme with NGM and N-SADDLe since it has been shown to perform better than stochastic quantization for extreme compression [4, 5]. However, to demonstrate the generalizability of our approach, we present results on 2-bit stochastic quantization in Table 3. NGM incurs an average drop of 4.4%, while N-SADDLe incurs only a 1.2% average accuracy drop in the presence of this compression scheme.

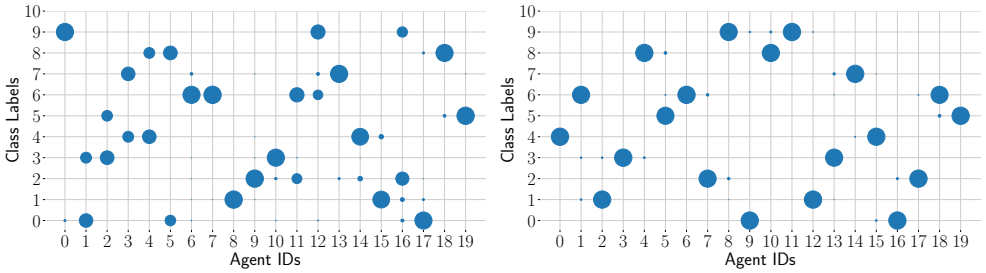
3.5. Loss Landscape Visualization

To visualize the loss landscape, we randomly sample two directions through orthogonal Gaussian perturbations [6] and plot the loss for ResNet-20 trained with CIFAR-10 distributed across 10 nodes with $\alpha = 0.001$. As shown in Figure 3, we observe that Q-SADDLe and Comp Q-SADDLe have much smoother loss landscapes than QGM and Comp QGM. The compressed counterparts of QGM and Q-SADDLe are relatively sharper than their respective full communication versions. This is intuitively expected since communication compression leads agents to receive less information from their neighbors, resulting in more reliance on local updates. This can exacerbate over-fitting in the presence of data heterogeneity. We observe similar trends for NGM, N-SADDLe, and their compressed versions as shown in Figure 4.



(a) Compression Error (b) Gradient Norm

Figure 1. Impact of flatness on (a) Compression Error and (b) Gradient Norm for ResNet-20 trained on CIFAR-10 distributed in a non-IID manner across a 10 agent ring topology.



(a) $\alpha = 0.01$ (b) $\alpha = 0.001$

Figure 2. Visualization of the number of samples from each class allocated to each agent for different Dirichlet distribution α values on the CIFAR-10 dataset.

3.6. Communication Cost

This section presents the exact amount of data transmitted (in Gigabytes) during training (Tables 4, 5, 6, 7 and 8).

Table 4. Communication costs per agent (in GBs) for experiments in Table 1 (main paper) for QGM and *Q-SADDLe* with a stochastic quantization-based compression scheme with 8 bits, leading to a 4× reduction in communication cost.

Agents	Comp	CIFAR-10	CIFAR-100
5		136.45	111.32
	✓	34.11	27.83
10		68.44	55.66
	✓	17.11	13.91
20		34.43	27.83
	✓	8.60	6.95
40		17.43	14.02
	✓	4.35	3.50

Table 5. Communication costs per agent (in GBs) for experiments in Table 2 for QGM and *Q-SADDLe* with a top-30% compression scheme, leading to a 2.2× reduction in communication cost.

Agents	Comp	CIFAR-10
5	✓	61.38
10	✓	30.78
20	✓	15.49

Table 6. Communication costs per agent (in GBs) for experiments in Table 2 (main paper) for QGM and *Q-SADDLe* with a stochastic quantization-based compression scheme with 10 bits , leading to a 3.2× reduction in communication cost.

Agents	Comp	Imagenette
5		110.23
	✓	34.44
10		55.10
	✓	17.21

4. Decentralized Learning Setup

All our experiments were conducted on a system with 4 NVIDIA A40 GPUs, each with 48GB GDDR6. We report the test accuracy of the consensus model averaged over three randomly chosen seeds.

4.1. Visualization of Non-IID Data

Figure 2 illustrates the number of samples from each class allocated to each agent for the 2 different Dirichlet distribution α values used in our work. $\alpha = 0.001$ corresponds to the most extreme form of data heterogeneity, i.e. samples from only 1 class per agent. Note that this level of non-IIDness has been used in CGA [3] and NGM [1]

Table 7. Communication costs per agent (in GBs) for experiments in Table 3 (main paper) for NGM and *N-SADDLe* with 1-bit Sign SGD, leading to a 32× reduction in the cost for the second round and a total of 1.94× reduction in the entire communication cost.

Agents	Comp	CIFAR-10	CIFAR-100
5		272.91	222.65
	✓	140.67	114.76
10		136.89	111.32
	✓	70.56	57.38
20		68.88	55.66
	✓	35.50	28.69

Table 8. Communication costs per agent (in GBs) for experiments in Table 4 (main paper) for NGM and *N-SADDLe* with 1-bit Sign SGD, leading to a 32× reduction in the cost for the second round and a total of 1.94× reduction in the entire communication cost.

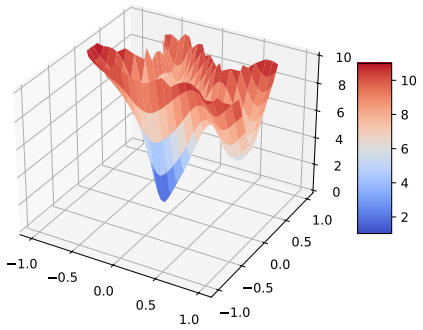
Agents	Comp	Imagenette	ImageNet
10		110.25	22466.30
	✓	56.82	11580.56

to evaluate the performance. $\alpha = 0.01$ has been used in QGM [8] and is relatively mild, with most agents accessing samples from 2 different classes (some even from 4 classes).

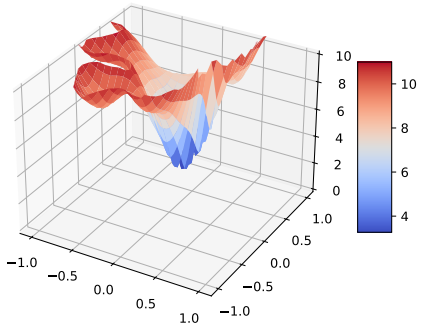
4.2. Hyper-parameters

This section presents the hyper-parameters for results presented in Section 6 (main paper) and Section 3. All our experiments were run for three randomly chosen seeds. We decay the learning rate by 10× after 50% and 75% of the training for all experiments except for ImageNet results in Table 4 and Figure 2. For ImageNet, we decay the learning rate by 10× after 33%, 67%, and 90% of the training. For Figure 2, we use the StepLR scheduler, where the learning rate decays by 0.981 after every epoch. We use a Nesterov momentum of 0.9 for all our experiments, and keep $\mu = \beta$, similar to QGM [8]. We also use a weight decay of 1e-4 for all the presented experiments. Please refer to Table 9 for the learning rate, perturbation radius, number of epochs, and batch size per agent for all the experiments in this paper. For a fair comparison, we ensure that all the techniques utilize the same set of hyper-parameters.

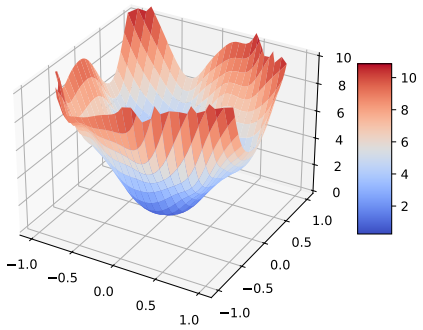
We tune the global averaging rate γ through a grid search over $\gamma = \{0.01, 0.1, 0.2, ..., 1.0\}$ and present the fine-tuned γ used for experiments in Tables 3, 4 from the main paper and Table 3 in Table 10. For results in Tables 1, 2 (main paper), and 1, we use $\gamma = 1.0$ for all the experiments. For Top-30% Sparsification results shown in Table 2, we use $\gamma = 0.4$. For our experiments on torus topology in Table 5 (main paper), we use an averaging rate of 0.5.



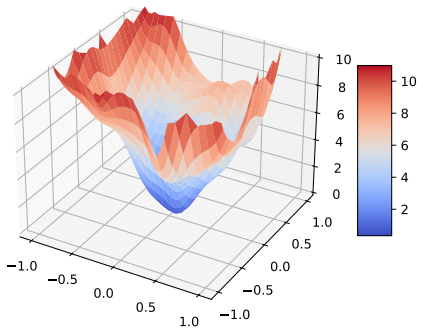
(a) QGM



(b) Comp QGM

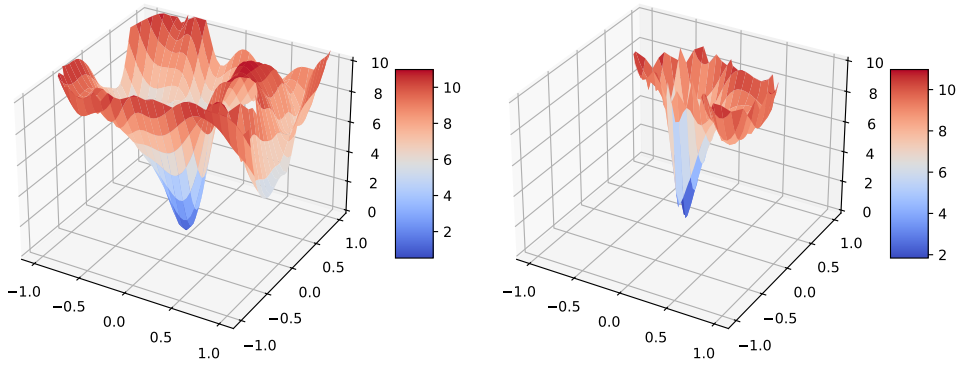


(c) Q-SADDLe

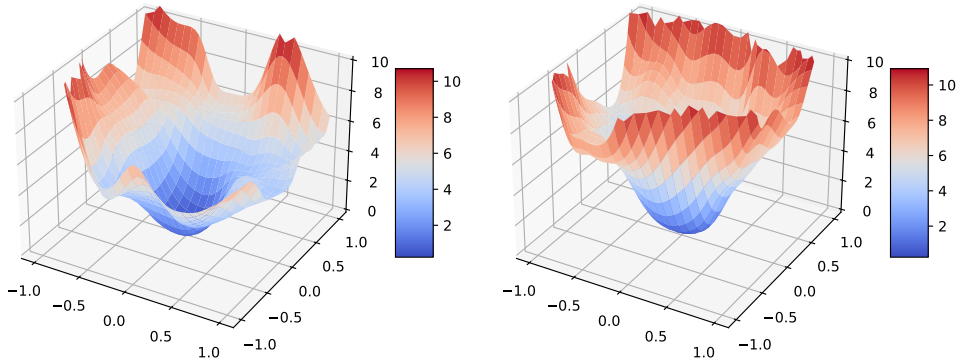


(d) Comp Q-SADDLe

Figure 3. Visualization of the loss landscape for ResNet-20 trained on the CIFAR-10 dataset distributed across a 10 agent ring topology with $\alpha = 0.001$.



(a) NGM (b) Comp NGM



(c) N-SADDLe (d) Comp N-SADDLe

Figure 4. Visualization of the loss landscape for ResNet-20 trained on the CIFAR-10 dataset distributed across a 10 agent ring topology with $\alpha = 0.001$.

Table 9. Learning rate (η), the perturbation radius (ρ) (where applicable), batch size per agent, and the number of epochs for all the experiments for QGM, Q-SADDLe, NGM, N-SADDLe, and their compressed versions across various datasets.

Dataset	CIFAR-10	CIFAR-100	Imagenette	ImageNet
Learning Rate (η)	0.1	0.1	0.01	0.01
Perturbation Radius (ρ)	0.1	0.05	0.01	0.05
Epochs	200	100	100	60
Batch-Size/Agent	32	20	32	64

Table 10. Global averaging rate (γ) for our experiments in Table 3, 4 (main paper) and 3.

Method	Non-IID Level (α)	CIFAR-10	CIFAR-100	Imagenette	ImageNet
NGM	0.01	1.0	1.0	0.5	1.0
	0.001	1.0	1.0	0.5	1.0
Comp NGM	0.01	0.5	0.5	0.1	0.5
	0.001	0.5	0.5	0.5	0.5
N-SADDLe	0.01	1.0	1.0	0.5	1.0
	0.001	1.0	1.0	0.5	1.0
Comp N-SADDLe	0.01	0.5	0.5	0.1	1.0
	0.001	0.5	0.5	0.5	1.0

References

[1] Sai Aparna Aketi, Sangamesh Kodge, and Kaushik Roy. Neighborhood gradient mean: An efficient decentralized learning method for non-IID data. *Transactions on Machine Learning Research*, 2023. 7, 10

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 9

[3] Yasaman Esfandiari, Sin Yong Tan, Zhanhong Jiang, Aditya Balu, Ethan Herron, Chinmay Hegde, and Soumik Sarkar. Cross-gradient aggregation for decentralized learning from non-iid data. In *International Conference on Machine Learning*, pages 3036–3046. PMLR, 2021. 10

[4] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019. 9

[5] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019. 9

[6] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 9

[7] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017. 7, 8

[8] Tao Lin, Sai Praneeth Karimireddy, Sebastian Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6654–6665. PMLR, 18–24 Jul 2021. 2, 3, 7, 8, 11