Supplementary materials for "CADA: Communication-Adaptive Distributed Adam"

In this supplementary document, we first present some basic inequalities that will be used frequently in this document, and then present the missing derivations of some claims, as well as the proofs of all the lemmas and theorems in the paper, which is followed by details on our experiments. The content of this supplementary document is summarized as follows.

Table of Contents

6 Supporting Lemmas

Define the σ-algebra $\Theta^k = \{\theta^l, 1 \le l \le k\}$. For convenience, we also initialize parameters as $\theta^{-D}, \theta^{-D+1}, \ldots, \theta^{-1} = \theta^0$. Some basic facts used in the proof are reviewed as follows.

Fact 1. Assume that $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$ are independent random variables, and $EX_1 = \cdots = EX_n = 0$. Then

$$\mathbb{E}\left[\Big\| \sum_{i=1}^n X_i \Big\|^2\right] = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|^2\right]. \tag{18}$$

Fact 2. (Young's inequality) For any $\theta_1, \theta_2 \in \mathbb{R}^p, \varepsilon > 0$,

$$\left\langle \theta_1, \theta_2 \right\rangle \le \frac{\|\theta_1\|^2}{2\varepsilon} + \frac{\varepsilon\|\theta_2\|^2}{2}. \tag{19}$$

As a consequence, we have

$$\|\theta_1 + \theta_2\|^2 \le \Big(1 + \frac{1}{\varepsilon}\Big)\|\theta_1\|^2 + (1+\varepsilon)\|\theta_2\|^2. \tag{20}$$

Fact 3. (Cauchy-Schwarz inequality) For any $\theta_1, \theta_2, \ldots, \theta_n \in \mathbb{R}^p$, we have

$$\Big\| \sum_{i=1}^n \theta_i \Big\|^2 \le n \sum_{i=1}^n \|\theta_i\|^2. \tag{21}$$

Lemma 4. *For $k-\tau_{\max}\le l\le k-D$, if $\{\theta^k\}$ are the iterates generated by CADA, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\rangle\right]\\ \le&\frac{L\epsilon^{-\frac{1}{2}}}{12\alpha_k}\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right]+6DL\alpha_k\epsilon^{-\frac{1}{2}}\sigma_m^2\end{aligned}\tag{22}$$

and similarly, we have

$$\begin{aligned}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\mathcal{L}_m(\theta^l)-\nabla\ell(\theta^l;\theta^{k-\tau_m^k}\right)\right\rangle\right]\\ \le&\frac{L\epsilon^{-\frac{1}{2}}}{12\alpha_k}\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right]+3DL\alpha_k\epsilon^{-\frac{1}{2}}\sigma_m^2.\end{aligned}\tag{23}$$

Proof: We first show the following holds.

$$\begin{aligned}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^l),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\rangle\right]\\ \overset{(a)}{=}&\mathbb{E}\left[\mathbb{E}\Big[\left\langle\nabla\mathcal{L}(\theta^l),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\rangle\Big|\Theta^l\Big]\right]\\ \overset{(b)}{=}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^l),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\mathbb{E}\Big[\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\Big|\Theta^l\Big]\right\rangle\right]\\ =&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^l),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\mathcal{L}_m(\theta^l)-\nabla\mathcal{L}_m(\theta^l)\right)\right\rangle\right]=0\end{aligned}\tag{24}$$

where (a) follows from the law of total probability, and (b) holds because $\hat{V}^{k-D}$ is deterministic conditioned on Θ^l when $k-D\le l$.

We first prove (22) by decomposing it as

$$\begin{aligned}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\rangle\right]\\ \overset{(c)}{=}&\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\theta^k)-\nabla\mathcal{L}(\theta^l),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\rangle\right]\\ \overset{(d)}{\le}&L\mathbb{E}\left[\left\|(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{4}}\right\|\left\|\theta^k-\theta^l\right\|\left\|(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{4}}\left(\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\right)\right\|\right]\\ \overset{(e)}{\le}&\frac{L\epsilon^{-\frac{1}{2}}}{12D\alpha_k}\underbrace{\mathbb{E}\left[\|\theta^k-\theta^l\|^2\right]}_{I_1}+\frac{6DL\alpha_k\epsilon^{-\frac{1}{2}}}{2}\underbrace{\mathbb{E}\left[\|\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\|^2\right]}_{I_2}\end{aligned}\tag{25}$$

where (c) holds due to (24), (d) uses Assumption 1, and (e) applies the Young's inequality.

Applying the Cauchy-Schwarz inequality to I_1, we have

$$\begin{aligned}I_1=&\mathbb{E}\Big[\Big\|\sum_{d=1}^{k-l}(\theta^{k+1-d}-\theta^{k-d})\Big\|^2\Big]\\ \le&(k-l)\sum_{d=1}^{k-l}\mathbb{E}\Big[\|\theta^{k+1-d}-\theta^{k-d}\|^2\Big]\le D\sum_{d=1}^{D}\mathbb{E}\Big[\|\theta^{k+1-d}-\theta^{k-d}\|^2\Big].\end{aligned}\tag{26}$$

Applying Assumption 2 to I_2, we have

$$\begin{aligned}I_2=&\mathbb{E}\Big[\big\|\nabla\ell(\theta^l;\xi_m^k)-\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\big\|^2\Big]\\ =&\mathbb{E}\Big[\big\|\nabla\ell(\theta^l;\xi_m^k)\big\|^2\Big]+\mathbb{E}\Big[\big\|\nabla\ell(\theta^l;\xi_m^{k-\tau_m^k})\big\|^2\Big]\le 2\sigma_m^2\end{aligned}\tag{27}$$

where the last inequality uses Assumption 2. Plugging (26) and (27) into (25), it leads to (22).

Likewise, following the steps to (25), it can be verified that (23) also holds true.

Lemma 5. *Under Assumption 2, the parameters $\{h^k, \hat{v}^k\}$ of CADA in Algorithm 1 satisfy*

$$\|h^k\| \le \sigma, \quad \forall k; \quad \hat{v}_i^k \le \sigma^2, \quad \forall k, i \tag{28}$$

where $\sigma := \frac{1}{M}\sum_{m\in\mathcal{M}} \sigma_m$.

Proof: Using Assumption 2, it follows that

$$\|\boldsymbol{\nabla}^k\| = \left\| \frac{1}{M} \sum_{m\in\mathcal{M}} \nabla \ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) \right\| \le \frac{1}{M} \sum_{m\in\mathcal{M}} \left\| \nabla \ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) \right\| \le \frac{1}{M} \sum_{m\in\mathcal{M}} \sigma_m = \sigma. \tag{29}$$

Therefore, from the update (2a), we have

$$\|h^{k+1}\| \le \beta_1 \|h^k\| + (1-\beta_1)\|\boldsymbol{\nabla}^k\| \le \beta_1 \|h^k\| + (1-\beta_1)\sigma.$$

Since $\|h^1\| \le \sigma$, if follows by induction that $\|h^{k+1}\| \le \sigma, \ \forall k$.

Using Assumption 2, it follows that

$$\begin{aligned}
(\nabla_i^k)^2 &= \left(\frac{1}{M} \sum_{m\in\mathcal{M}} \nabla_i \ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) \right)^2 \\
&\le \frac{1}{M} \sum_{m\in\mathcal{M}} \left(\nabla_i \ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) \right)^2 \\
&\le \frac{1}{M} \sum_{m\in\mathcal{M}} \left\| \nabla \ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) \right\|^2 = \frac{1}{M} \sum_{m\in\mathcal{M}} \sigma_m^2 \le \sigma^2.
\end{aligned} \tag{30}$$

Similarly, from the update (2b), we have

$$\hat{v}_i^{k+1} \le \max\{\hat{v}_i^k, \beta_2 \hat{v}_i^k + (1-\beta_2)(\nabla_i^k)^2\} \le \max\{\hat{v}_i^k, \beta_2 \hat{v}_i^k + (1-\beta_2)\sigma^2\}.$$

Since $v_i^1 = \hat{v}_i^1 \le \sigma^2$, if follows by induction that $\hat{v}_i^{k+1} \le \sigma^2$.

Lemma 6. *Under Assumption 2, the iterates $\{\theta^k\}$ of CADA in Algorithm 1 satisfy*

$$\left\|\theta^{k+1} - \theta^k\right\|^2 \le \alpha_k^2 p (1-\beta_2)^{-1}(1-\beta_3)^{-1} \tag{31}$$

where p is the dimension of θ, $\beta_1 < \sqrt{\beta_2} < 1$, and $\beta_3 := \beta_1^2/\beta_2$.

Proof: Choosing $\beta_1 < 1$ and defining $\beta_3 := \beta_1^2/\beta_2$, it can be verified that

$$\begin{aligned}
|h_i^{k+1}| &= \left|\beta_1 h_i^k + (1-\beta_1)\nabla_i^k\right| \beta_1 |h_i^k| + |\nabla_i^k| \\
&\le \beta_1 \left(\beta_1 |h_i^{k-1}| + |\nabla_i^{k-1}|\right) + |\nabla_i^k| \\
&\le \sum_{l=0}^k \beta_1^{k-l} |\nabla_i^l| = \sum_{l=0}^k \sqrt{\beta_3}^{k-l} \sqrt{\beta_2}^{k-l} |\nabla_i^l| \\
&\overset{(a)}{\le} \left(\sum_{l=0}^k \beta_3^{k-l} \right)^{\frac{1}{2}} \left(\sum_{l=0}^k \beta_2^{k-l} (\nabla_i^l)^2 \right)^{\frac{1}{2}} \\
&\le (1-\beta_3)^{-\frac{1}{2}} \left(\sum_{l=0}^k \beta_2^{k-l} (\nabla_i^l)^2 \right)^{\frac{1}{2}}
\end{aligned} \tag{32}$$

where (a) follows from the Cauchy-Schwartz inequality.

For $\hat{v}_i^k$, first we have that $\hat{v}_i^1 \ge (1-\beta_2)(\nabla_i^1)^2$. Then since

$$\hat{v}_i^{k+1} \ge \beta_2 \hat{v}_i^k + (1-\beta_2)(\nabla_i^k)^2$$

by induction we have

$$\hat{v}_i^{k+1} \geq (1-\beta_2)\sum_{l=0}^{k}\beta_2^{k-l}(\nabla_i^l)^2. \tag{33}$$

Using (32) and (33), we have

$$\begin{aligned}|h_i^{k+1}|^2 \leq& (1-\beta_3)^{-1}\left(\sum_{l=0}^{k}\beta_2^{k-l}(\nabla_i^l)^2\right)\\ \leq& (1-\beta_2)^{-1}(1-\beta_3)^{-1}\hat{v}_i^{k+1}.\end{aligned}$$

From the update (2c), we have

$$\begin{aligned}\|\theta^{k+1}-\theta^k\|^2 &= \alpha_k^2\sum_{i=1}^{p}\left(\epsilon+\hat{v}_i^{k+1}\right)^{-1}|h_i^{k+1}|^2\\ &\leq \alpha_k^2 p(1-\beta_2)^{-1}(1-\beta_3)^{-1}\end{aligned} \tag{34}$$

which completes the proof.

7 Missing Derivations in Section 2.2

The analysis in this part is analogous to that in [12]. We define an auxiliary function as

$$\psi_m(\theta) = \mathcal{L}_m(\theta) - \mathcal{L}_m(\theta^\star) - \left\langle \nabla\mathcal{L}_m(\theta^\star), \theta-\theta^\star\right\rangle$$

where $\theta^\star$ is a minimizer of $\mathcal{L}$. Assume that $\nabla\ell(\theta;\xi_m)$ is $\bar{L}$-Lipschitz continuous for all ξ_m, we have

$$\|\nabla\ell(\theta;\xi_m)-\nabla\ell(\theta^\star;\xi_m)\|^2 \leq 2\bar{L}\left(\ell(\theta;\xi_m)-\ell(\theta^\star;\xi_m)-\left\langle\nabla\ell(\theta^\star;\xi_m),\theta-\theta^\star\right\rangle\right).$$

Taking expectation with respect to ξ_m, we can obtain

$$\mathbb{E}_{\xi_m}[\|\nabla\ell(\theta;\xi_m)-\nabla\ell(\theta^\star;\xi_m)\|^2] \leq 2\bar{L}\left(\mathcal{L}_m(\theta)-\mathcal{L}_m(\theta^\star)-\left\langle\nabla\mathcal{L}_m(\theta^\star),\theta-\theta^\star\right\rangle\right) = 2\bar{L}\psi_m(\theta).$$

Note that $\nabla\mathcal{L}_m$ is also $\bar{L}$-Lipschitz continuous and thus

$$\|\nabla\mathcal{L}_m(\theta)-\nabla\mathcal{L}_m(\theta^\star)\|^2 \leq 2\bar{L}(\mathcal{L}_m(\theta)-\mathcal{L}_m(\theta^\star)-\left\langle\nabla\mathcal{L}_m(\theta^\star),\theta-\theta^\star\right\rangle) = 2\bar{L}\psi_m(\theta).$$

7.1 Derivations of (6)

By (21), we can derive that

$$\|\theta_1+\theta_2\| \leq 2\|\theta_1\|^2+2\|\theta_2\|^2$$

which also implies $\|\theta_1\|^2 \geq \frac{1}{2}\|\theta_1+\theta_2\|^2-\|\theta_2\|^2$.

As a consequence, we can obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla\ell(\theta^k;\xi_m^k)-\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k})\right\|^2\right]\\ \geq&\frac{1}{2}\mathbb{E}\left[\left\|\left(\nabla\ell(\theta^k;\xi_m^k)-\nabla\mathcal{L}_m(\theta^k)\right)+\left(\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})-\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k})\right)\right\|^2\right]\\ &\qquad\qquad -\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(\theta^k)-\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})\right\|^2\right]\\ =&\frac{1}{2}\mathbb{E}\left[\left\|\nabla\ell(\theta^k;\xi_m^k)-\nabla\mathcal{L}_m(\theta^k)\right\|^2\right]+\frac{1}{2}\mathbb{E}\left[\left[\|\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k})-\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})\right\|^2\right]\right]\\ &\underbrace{+\mathbb{E}\left[\left\langle\nabla\ell(\theta^k;\xi_m^k)-\nabla\mathcal{L}_m(\theta^k),\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})-\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k})\right\rangle\right]}_{I_3}-\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(\theta^k)-\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})\right\|^2\right]\end{aligned}$$

where we used the fact that $I_3 = 0$ to obtain (6), that is

$$I_3 = \mathbb{E}\left[\left\langle\mathbb{E}\left[\nabla\ell(\theta^k;\xi_m^k)\middle|\Theta^k\right]-\nabla\mathcal{L}_m(\theta^k),\nabla\mathcal{L}_m(\theta^{k-\tau_m^k})-\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k})\right\rangle\right] = 0.$$

7.2 Derivations of (9)

Recall that

$$\begin{aligned}\tilde{\delta}_m^k - \tilde{\delta}_m^{k-\tau_m^k} =&\big(\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\tilde{\theta};\xi_m^k) + \nabla\mathcal{L}_m(\tilde{\theta})\big) - \big(\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k}) - \nabla\ell(\tilde{\theta};\xi_m^{k-\tau_m^k}) + \nabla\mathcal{L}_m(\tilde{\theta})\big)\\ =&\underbrace{\big(\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\tilde{\theta};\xi_m^k) + \nabla\psi_m(\tilde{\theta})\big)}_{g_m^k} - \underbrace{\big(\nabla\ell(\theta^{k-\tau_m^k};\xi_m^{k-\tau_m^k}) - \nabla\ell(\tilde{\theta};\xi_m^{k-\tau_m^k}) + \nabla\psi_m(\tilde{\theta})\big)}_{g_m^{k-\tau_m^k}}.\end{aligned}$$

And by (21), we have $\|\tilde{\delta}_m^k - \tilde{\delta}_m^{k-\tau_m^k}\|^2 \le 2\|g_m^k\|^2 + 2\|g_m^{k-\tau_m^k}\|^2$. We decompose the first term as

$$\begin{aligned}\mathbb{E}[\|g_m^k\|^2] \le&2\mathbb{E}[\|\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k)\|^2] + 2\mathbb{E}[\|\nabla\ell(\tilde{\theta};\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k) - \nabla\psi_m(\tilde{\theta})\|^2]\\ =&2\mathbb{E}[\mathbb{E}[\|\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k)\|^2|\Theta^k]]\\ &+ 2\mathbb{E}[\|\nabla\ell(\tilde{\theta};\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k) - \mathbb{E}[\nabla\ell(\tilde{\theta};\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k)|\Theta^k]\|^2]\\ \le&4\bar{L}\mathbb{E}\psi_m(\theta^k) + 2\mathbb{E}[\|\nabla\ell(\tilde{\theta};\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k)\|^2]\\ =&4\bar{L}\mathbb{E}\psi_m(\theta^k) + 2\mathbb{E}[\mathbb{E}[\|\nabla\ell(\tilde{\theta};\xi_m^k) - \nabla\ell(\theta^\star;\xi_m^k)\|^2|\Theta^k]]\\ \le&4\bar{L}\mathbb{E}\psi_m(\theta^k) + 4\bar{L}\mathbb{E}\psi_m(\tilde{\theta}).\end{aligned}$$

By nonnegativity of ψ_m, we have

$$\begin{aligned}\mathbb{E}[\|g_m^k\|^2] &\le 4\bar{L}\sum_{m\in\mathcal{M}}\mathbb{E}\psi_m(\theta^k) + 4\bar{L}\sum_{m\in\mathcal{M}}\mathbb{E}\psi_m(\tilde{\theta})\\ &= 4M\bar{L}(\mathbb{E}\mathcal{L}(\theta^k) - \mathcal{L}(\theta^\star)) + 4M\bar{L}(\mathbb{E}\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star)).\end{aligned} \tag{35}$$

Similarly, we can prove

$$\mathbb{E}[\|g_m^{k-\tau_m^k}\|^2] \le 4M\bar{L}(\mathbb{E}\mathcal{L}(\theta^{k-\tau_m^k}) - \mathcal{L}(\theta^\star)) + 4M\bar{L}(\mathbb{E}\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star)). \tag{36}$$

Therefore, it follows that

$$\begin{aligned}&\mathbb{E}[\|\tilde{\delta}_m^k - \tilde{\delta}_m^{k-\tau_m^k}\|^2]\\ &\quad\le 8M\bar{L}(\mathbb{E}\mathcal{L}(\theta^k) - \mathcal{L}(\theta^\star)) + 8M\bar{L}(\mathbb{E}\mathcal{L}(\theta^{k-\tau_m^k}) - \mathcal{L}(\theta^\star)) + 16M\bar{L}(\mathbb{E}\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star)).\end{aligned}$$

7.3 Derivations of (11)

The LHS of (10) can be written as

$$\begin{aligned}\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^{k-\tau_m^k};\xi_m^k) =&\big(\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^{k-\tau_m^k};\xi_m^k) + \nabla\mathcal{L}_m(\theta^{k-\tau_m^k})\big) - \nabla\mathcal{L}_m(\theta^{k-\tau_m^k})\\ =&\big(\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^{k-\tau_m^k};\xi_m^k) + \nabla\psi_m(\theta^{k-\tau_m^k})\big) - \nabla\psi_m(\theta^{k-\tau_m^k}).\end{aligned}$$

Similar to (35), we can obtain

$$\begin{aligned}&\mathbb{E}[\|\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^{k-\tau_m^k};\xi_m^k) + \nabla\psi_m(\theta^{k-\tau_m^k})\|^2]\\ &\qquad\le 4M\bar{L}(\mathbb{E}\mathcal{L}(\theta^k) - \mathcal{L}(\theta^\star)) + 4M\bar{L}(\mathbb{E}\mathcal{L}(\theta^{k-\tau_m^k}) - \mathcal{L}(\theta^\star)).\end{aligned}$$

Combined with the fact

$$\begin{aligned}\mathbb{E}[\|\nabla\psi_m(\theta^{k-\tau_m^k})\|^2] &= \mathbb{E}[\|\nabla\mathcal{L}_m(\theta^{k-\tau_m^k}) - \nabla\mathcal{L}_m(\theta^\star)\|^2]\\ &\le 2\bar{L}\mathbb{E}\psi_m(\theta^{k-\tau_m^k}) \le 2M\bar{L}(\mathbb{E}\mathcal{L}(\theta^{k-\tau_m^k}) - \mathcal{L}(\theta^\star))\end{aligned}$$

we have

$$\mathbb{E}[\|\nabla\ell(\theta^k;\xi_m^k) - \nabla\ell(\theta^{k-\tau_m^k};\xi_m^k)\|^2] \le 8M\bar{L}(\mathbb{E}\mathcal{L}(\theta^k) - \mathcal{L}(\theta^\star)) + 12M\bar{L}(\mathbb{E}\mathcal{L}(\theta^{k-\tau_m^k}) - \mathcal{L}(\theta^\star)).$$

8 Proof of Lemma 1

Using the smoothness of $\mathcal{L}(\theta)$ in Assumption 1, we have

$$\begin{aligned}\mathcal{L}(\theta^{k+1}) \le& \mathcal{L}(\theta^k) + \Big\langle \nabla\mathcal{L}(\theta^k), \theta^{k+1}-\theta^k\Big\rangle + \frac{L}{2}\big\|\theta^{k+1}-\theta^k\big\|^2\\ =& \mathcal{L}(\theta^k) - \alpha_k\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k+1})^{-\frac{1}{2}} h^{k+1}\Big\rangle + \frac{L}{2}\big\|\theta^{k+1}-\theta^k\big\|^2. \end{aligned} \tag{37}$$

We can further decompose the inner product as

$$\begin{aligned}&-\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k+1})^{-\frac{1}{2}} h^{k+1}\Big\rangle\\ =&-(1-\beta_1)\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} \boldsymbol{\nabla}^k\Big\rangle \underbrace{-\beta_1\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} h^k\Big\rangle}_{I_1^k}\\ &\underbrace{-\Big\langle \nabla\mathcal{L}(\theta^k), \Big((\epsilon I + \hat{V}^{k+1})^{-\frac{1}{2}} - (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}}\Big) h^{k+1}\Big\rangle}_{I_2^k}\end{aligned} \tag{38}$$

where we again decompose the first inner product as

$$\begin{aligned}-(1-\beta_1)\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} \boldsymbol{\nabla}^k\Big\rangle =& \underbrace{-(1-\beta_1)\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \boldsymbol{\nabla}^k\Big\rangle}_{I_3^k}\\ &\underbrace{-(1-\beta_1)\Big\langle \nabla\mathcal{L}(\theta^k), \Big((\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} - (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}\Big) \boldsymbol{\nabla}^k\Big\rangle}_{I_4^k}.\end{aligned} \tag{39}$$

Next, we bound the terms $I_1^k, I_2^k, I_3^k, I_4^k$ separately.

Taking expectation on I_1^k conditioned on Θ^k, we have

$$\begin{aligned}\mathbb{E}[I_1^k \mid \Theta^k] =& -\mathbb{E}\Big[\beta_1\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} h^k\Big\rangle \mid \Theta^k\Big]\\ =& -\beta_1\Big\langle \nabla\mathcal{L}(\theta^{k-1}), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} h^k\Big\rangle - \beta_1\Big\langle \nabla\mathcal{L}(\theta^k)-\nabla\mathcal{L}(\theta^{k-1}), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} h^k\Big\rangle\\ \overset{(a)}{\le}& -\beta_1\Big\langle \nabla\mathcal{L}(\theta^{k-1}), (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}} h^k\Big\rangle + \alpha_{k-1}^{-1}\beta_1 L\big\|\theta^k-\theta^{k-1}\big\|^2\\ \overset{(b)}{\le}& \beta_1\left(I_1^{k-1}+I_2^{k-1}+I_3^{k-1}+I_4^{k-1}\right) + \alpha_{k-1}^{-1}\beta_1 L\big\|\theta^k-\theta^{k-1}\big\|^2\end{aligned} \tag{40}$$

where follows from the L-smoothness of $\mathcal{L}(\boldsymbol{\theta})$ implied by Assumption 1; and (b) uses again the decomposition (38) and (39).

Taking expectation on I_2^k over all the randomness, we have

$$\begin{aligned}\mathbb{E}[I_2^k] =& \mathbb{E}\Big[-\Big\langle \nabla\mathcal{L}(\theta^k), \Big((\epsilon I + \hat{V}^{k+1})^{-\frac{1}{2}} - (\epsilon I + \hat{V}^{k})^{-\frac{1}{2}}\Big) h^{k+1}\Big\rangle\Big]\\ =& \mathbb{E}\Big[\sum_{i=1}^p \nabla_i\mathcal{L}(\theta^k) h_i^{k+1}\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}} - (\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\\ \overset{(d)}{\le}& \mathbb{E}\Big[\|\nabla\mathcal{L}(\theta^k)\|\|h^{k+1}\|\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}} - (\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\\ \overset{(e)}{\le}& \sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}} - (\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\end{aligned} \tag{41}$$

where (d) follows from the Cauchy-Schwarz inequality and (e) is due to Assumption 2.

Regarding I_3^k, we will bound separately in Lemma 2.

Taking expectation on I_4^k over all the randomness, we have

$$\begin{aligned}\mathbb{E}[I_4^k] =&\mathbb{E}\Big[-(1-\beta_1)\Big\langle\nabla\mathcal{L}(\theta^k),\Big((\epsilon I+\hat{V}^k)^{-\frac{1}{2}}-(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\Big)\boldsymbol{\nabla}^k\Big\rangle\Big]\\ =&-(1-\beta_1)\mathbb{E}\Big[\sum_{i=1}^p\nabla_i\mathcal{L}(\theta^k)\boldsymbol{\nabla}_i^k\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}\Big)\Big]\\ \le&(1-\beta_1)\mathbb{E}\Big[\|\nabla\mathcal{L}(\theta^k)\|\|\boldsymbol{\nabla}^k\|\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}\Big)\Big]\\ \le&(1-\beta_1)\sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}\Big)\Big].\end{aligned}\tag{42}$$

Taking expectation on (37) over all the randomness, and plugging (40), (41), and (42), we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta^{k+1})]-\mathbb{E}[\mathcal{L}(\theta^k)]\le&-\alpha_k\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k+1})^{-\frac{1}{2}}h^{k+1}\Big\rangle\right]+\frac{L}{2}\mathbb{E}\left[\left\|\theta^{k+1}-\theta^k\right\|^2\right]\\ =&\alpha_k\mathbb{E}\left[I_1^k+I_2^k+I_3^k+I_4^k\right]+\frac{L}{2}\mathbb{E}\left[\left\|\theta^{k+1}-\theta^k\right\|^2\right]\\ \le&-\alpha_k(1-\beta_1)\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\boldsymbol{\nabla}^k\Big\rangle\right]\\ &-\alpha_k\beta_1\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^{k-1}),(\epsilon I+\hat{V}^k)^{-\frac{1}{2}}h^k\Big\rangle\right]\\ &+\alpha_k\sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\\ &+\alpha_k(1-\beta_1)\sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}\Big)\Big]\\ &+\left(\frac{L}{2}+\alpha_k\alpha_{k-1}^{-1}\beta_1L\right)\mathbb{E}\left[\|\theta^{k+1}-\theta^k\|^2\right].\end{aligned}\tag{43}$$

Since $(\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}\le(\epsilon+\hat{v}_i^{k-1})^{-\frac{1}{2}}$, we have

$$\begin{aligned}&\sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)+(1-\beta_1)\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^k)^{-\frac{1}{2}}\Big)\Big]\\ \le&(2-\beta_1)\sigma^2\mathbb{E}\Big[\sum_{i=1}^p\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big].\end{aligned}\tag{44}$$

Plugging (44) into (43) leads to the statement of Lemma 1.

9 Proof of Lemma 2

We first analyze the inner produce under CADA2 and then CADA1.

First recall that $\bar{\boldsymbol{\nabla}}^k=\frac{1}{M}\sum_{m\in\mathcal{M}}\nabla\ell(\theta^k;\xi_m^k)$. Using the law of total probability implies that

$$\begin{aligned}\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\bar{\boldsymbol{\nabla}}^k\Big\rangle\right]&=\mathbb{E}\left[\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\bar{\boldsymbol{\nabla}}^k\Big\rangle\mid\Theta^k\right]\right]\\ &=\mathbb{E}\left[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\mathbb{E}\left[\bar{\boldsymbol{\nabla}}^k\mid\Theta^k\right]\Big\rangle\right]\\ &=\mathbb{E}\left[\left\|\nabla\mathcal{L}(\theta^k)\right\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\right].\end{aligned}\tag{45}$$

Taking expectation on $\left\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \boldsymbol{\nabla}^k \right\rangle$ over all randomness, we have

$$\begin{aligned}
&- \mathbb{E}\left[\left\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \boldsymbol{\nabla}^k \right\rangle\right] \\
=&- \mathbb{E}\left[\left\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \bar{\boldsymbol{\nabla}}^k \right\rangle\right] \\
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \frac{1}{M}\sum_{m\in\mathcal{M}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \\
\overset{(a)}{=}&- \mathbb{E}\left[\left\|\nabla\mathcal{L}(\theta^k)\right\|^2_{(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}}\right] \\
&- \frac{1}{M}\sum_{m\in\mathcal{M}} \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \qquad (46)
\end{aligned}$$

where (a) uses (45).

Decomposing the inner product, for the CADA2 rule (10), we have

$$\begin{aligned}
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \\
=&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k)\right) \Big\rangle\Big] \\
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \\
\overset{(b)}{\leq}&\frac{L\epsilon^{-\frac{1}{2}}}{12\alpha_k}\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] + 6DL\alpha_k\epsilon^{-\frac{1}{2}}\sigma_m^2 \\
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \qquad (47)
\end{aligned}$$

where (b) follows from Lemma 4.

Using the Young's inequality, we can bound the last inner product in (47) as

$$\begin{aligned}
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \\
\leq&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}}\Big] + \frac{1}{2}\mathbb{E}\Big[\Big\|(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}\Big\|\Big\|\left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k) - \nabla\ell(\theta^k; \xi_m^k)\right)\Big\|^2\Big] \\
\overset{(g)}{\leq}&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}}\Big] + \frac{1}{2}\mathbb{E}\Big[\Big\|(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}\Big\|\Big\|\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^k) - \nabla\ell(\theta^k; \xi_m^k)\Big\|^2\Big] \\
\overset{(h)}{\leq}&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}}\Big] + \frac{c}{2d_{\max}}\mathbb{E}\Big[\Big\|(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}\Big\|\sum_{d=1}^{d_{\max}}\left\|\theta^{k+1-d} - \theta^{k-d}\right\|^2\Big] \\
\overset{(i)}{\leq}&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}}}\Big] + \frac{c\epsilon^{-\frac{1}{2}}}{2d_{\max}}\sum_{d=1}^{D}\mathbb{E}\Big[\left\|\theta^{k+1-d} - \theta^{k-d}\right\|^2\Big] \qquad (48)
\end{aligned}$$

where (g) follows from the Cauchy-Schwarz inequality, and (h) uses the adaptive communication condition (10) in CADA2, and (i) follows since $\hat{V}^{k-D}$ is entry-wise nonnegative and $\left\|\theta^{k+1-d} - \theta^{k-d}\right\|^2$ is nonnegative.

Similarly for CADA1's condition (7), we have

$$\begin{aligned}
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\theta^{k-\tau_m^k}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\theta^k; \xi_m^k)\right) \Big\rangle\Big] \\
=&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\nabla\ell(\tilde{\theta}; \xi_m^{k-\tau_m^k}) - \nabla\ell(\tilde{\theta}; \xi_m^k)\right) \Big\rangle\Big] \\
&- \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\tilde{\delta}_m^{k-\tau_m^k} - \tilde{\delta}_m^k\right) \Big\rangle\Big] \\
\overset{(j)}{\leq}&\frac{L\epsilon^{-\frac{1}{2}}}{12\alpha_k}\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] + 6DL\alpha_k\epsilon^{-\frac{1}{2}}\sigma_m^2 - \mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I + \hat{V}^{k-D})^{-\frac{1}{2}} \left(\tilde{\delta}_m^{k-\tau_m^k} - \tilde{\delta}_m^k\right) \Big\rangle\Big] \qquad (49)
\end{aligned}$$

where (j) follows from Lemma 4 since $\tilde{\theta}$ is a snapshot among $\{\theta^k, \cdots, \theta^{k-D}\}$.

And the last product in (49) is bounded by

$$\begin{aligned}
&-\mathbb{E}\Big[\Big\langle \nabla\mathcal{L}(\theta^k), (\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}\Big(\tilde{\delta}_m^{k-\tau_m^k}-\tilde{\delta}_m^k\Big)\Big\rangle\Big]\\
\leq&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\Big]+\frac{c}{2}\mathbb{E}\Big[\Big\|(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2d_{\max}}}\Big\|\sum_{d=1}^{d_{\max}}\big\|\theta^{k+1-d}-\theta^{k-d}\big\|^2\Big]\\
\overset{(i)}{\leq}&\frac{1}{2}\mathbb{E}\Big[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\Big]+\frac{c\epsilon^{-\frac{1}{2}}}{2d_{\max}}\sum_{d=1}^{D}\mathbb{E}\Big[\big\|\theta^{k+1-d}-\theta^{k-d}\big\|^2\Big].
\end{aligned}\tag{50}$$

Combining (46)-(50) leads to the desired statement for CADA1 and CADA2.

10 Proof of Lemma 3

For notational brevity, we re-write the Lyapunov function (14) as

$$\begin{aligned}
\mathcal{V}^k:=\mathcal{L}(\theta^k)-\mathcal{L}(\theta^\star)-c_k\Big\langle\nabla\mathcal{L}(\theta^{k-1}),(\epsilon I+\hat{V}^k)^{-\frac{1}{2}}h^k\Big\rangle\\
+b_k\sum_{d=0}^{D}\sum_{i=1}^{p}(\epsilon+\hat{v}_i^{k-d})^{-\frac{1}{2}}+\sum_{d=1}^{D}\rho_d\|\theta^{k+1-d}-\theta^{k-d}\|^2
\end{aligned}\tag{51}$$

where $\{c_k\}$ are some positive constants.

Therefore, taking expectation on the difference of $\mathcal{V}^k$ and $\mathcal{V}^{k+1}$ in (51), we have (with $\rho_{D+1}=0$)

$$\begin{aligned}
\mathbb{E}[\mathcal{V}^{k+1}]-\mathbb{E}[\mathcal{V}^k]=&\mathbb{E}[\mathcal{L}(\theta^{k+1})]-\mathbb{E}[\mathcal{L}(\theta^k)]-c_{k+1}\mathbb{E}\Big[\Big\langle\nabla\mathcal{L}(\theta^k),(\epsilon I+\hat{V}^{k+1})^{-\frac{1}{2}}h^{k+1}\Big\rangle\Big]\\
&+c_k\mathbb{E}\Big[\Big\langle\nabla\mathcal{L}(\theta^{k-1}),(\epsilon I+\hat{V}^k)^{-\frac{1}{2}}h^k\Big\rangle\Big]\\
&+b_{k+1}\sum_{d=0}^{D}\sum_{i=1}^{p}(\epsilon+\hat{v}_i^{k+1-d})^{-\frac{1}{2}}-b_k\sum_{d=0}^{D}\sum_{i=1}^{p}(\epsilon+\hat{v}_i^{k-d})^{-\frac{1}{2}}\\
&+\rho_1\mathbb{E}\left[\|\theta^{k+1}-\theta^k\|^2\right]+\sum_{d=1}^{D}(\rho_{d+1}-\rho_d)\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right]\\
\overset{(a)}{\leq}&(\alpha_k+c_{k+1})\mathbb{E}\left[I_1^k+I_2^k+I_3^k+I_4^k\right]-c_k\mathbb{E}\left[I_1^{k-1}+I_2^{k-1}+I_3^{k-1}+I_4^{k-1}\right]\\
&+b_{k+1}\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big]-b_k\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}\Big]\\
&+\sum_{d=1}^{D}(b_{k+1}-b_k)\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k+1-d})^{-\frac{1}{2}}\Big]+\left(\frac{L}{2}+\rho_1\right)\mathbb{E}\left[\|\theta^{k+1}-\theta^k\|^2\right]\\
&+\sum_{d=1}^{D}(\rho_{d+1}-\rho_d)\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right]
\end{aligned}\tag{52}$$

where (a) uses the smoothness in Assumption 1 and the definition of $I_1^k, I_2^k, I_3^k, I_4^k$ in (38) and (39).

Note that we can bound $(\alpha_k+c_{k+1})\mathbb{E}\left[I_1^k+I_2^k+I_3^k+I_4^k\right]$ the same as (38) in the proof of Lemma 1. In addition, Lemma 2 implies that

$$\begin{aligned}
\mathbb{E}[I_3^k]\leq&-\frac{1-\beta_1}{2}\mathbb{E}\left[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\right]\\
&+(1-\beta_1)\epsilon^{-\frac{1}{2}}\left(\frac{L}{12\alpha_k}+\frac{c}{2d_{\max}}\right)\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right]+(1-\beta_1)\frac{6DL\alpha_k\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2.
\end{aligned}\tag{53}$$

Therefore, plugging Lemma 1 with α_k replaced by $\alpha_k + c_{k+1}$ into (52), together with (53), leads to

$$\begin{aligned}
\mathbb{E}[\mathcal{V}^{k+1}] - \mathbb{E}[\mathcal{V}^k] \leq & -(\alpha_k + c_{k+1})\left(\frac{1-\beta_1}{2}\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}(\theta^k)\right\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\right] \\
& + (\alpha_k + c_{k+1})(1-\beta_1)\epsilon^{-\frac{1}{2}}\left(\frac{L}{12\alpha_k} + \frac{c}{2d_{\max}}\right)\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] \\
& + (\alpha_k + c_{k+1})(1-\beta_1)\frac{6DL\alpha_k\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2 \\
& + ((\alpha_k + c_{k+1})\beta_1 - c_k)\mathbb{E}\left[I_1^{k-1} + I_2^{k-1} + I_3^{k-1} + I_4^{k-1}\right] \\
& + (\alpha_k + c_{k+1})(2-\beta_1)\sigma^2\mathbb{E}\Big[\sum_{i=1}^{p}\Big((\epsilon + \hat{v}_i^{k-D})^{-\frac{1}{2}} - (\epsilon + \hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big] \\
& + b_{k+1}\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon + \hat{v}_i^{k+1})^{-\frac{1}{2}}\Big] - b_k\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon + \hat{v}_i^{k-D})^{-\frac{1}{2}}\Big] \\
& + \sum_{d=1}^{D}(b_{k+1} - b_k)\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon + \hat{v}_i^{k+1-d})^{-\frac{1}{2}}\Big] \\
& + \sum_{d=1}^{D}(\rho_{d+1} - \rho_d)\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] \\
& + \left(\frac{L}{2} + \rho_1 + (\alpha_k + c_{k+1})\alpha_{k-1}^{-1}\beta_1 L\right)\mathbb{E}\left[\|\theta^{k+1} - \theta^k\|^2\right]. \qquad (54)
\end{aligned}$$

Select $\alpha_k \leq \alpha_{k-1}$ and $c_k := \sum_{j=k}^{\infty}\alpha_j\beta_1^{j-k+1} \leq (1-\beta_1)^{-1}\alpha_k$ so that $(\alpha_k + c_{k+1})\beta_1 = c_k$ and

$$\begin{aligned}
(\alpha_k + c_{k+1})(1-\beta_1) &\leq (\alpha_k + (1-\beta_1)^{-1}\alpha_{k+1})(1-\beta_1) \\
&\leq \alpha_k(1 + (1-\beta_1)^{-1})(1-\beta_1) = \alpha_k(2-\beta_1).
\end{aligned}$$

In addition, select b_k to ensure that $b_{k+1} \leq b_k$. Then it follows from (54) that

$$\begin{aligned}
\mathbb{E}[\mathcal{V}^{k+1}] - \mathbb{E}[\mathcal{V}^k] \leq & -\frac{\alpha_k(1-\beta_1)}{2}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\theta^k)\right\|^2_{(\epsilon I+\hat{V}^{k-D})^{-\frac{1}{2}}}\right] + (2-\beta_1)\alpha_k^2\frac{6DL\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2 \\
& + (2-\beta_1)\alpha_k\epsilon^{-\frac{1}{2}}\left(\frac{L}{12\alpha_k} + \frac{c}{2d_{\max}}\right)\sum_{d=1}^{D}\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] \\
& + \left(\frac{(2-\beta_1)^2}{(1-\beta_1)}\alpha_k\sigma^2 - b_k\right)\mathbb{E}\Big[\sum_{i=1}^{p}\Big((\epsilon + \hat{v}_i^{k-D})^{-\frac{1}{2}} - (\epsilon + \hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big] \\
& + \left(\frac{L}{2} + \rho_1 + (1-\beta_1)^{-1}L\right)\mathbb{E}\left[\|\theta^{k+1} - \theta^k\|^2\right] \\
& + \sum_{d=1}^{D}(\rho_{d+1} - \rho_d)\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right] \qquad (55)
\end{aligned}$$

where we have also used the fact that $-(\alpha_k + c_{k+1})\left(\frac{1-\beta_1}{2}\right) \leq -\frac{\alpha_k(1-\beta_1)}{2}$ since $c_{k+1} \geq 0$.

If we choose $\alpha_k \le \frac{1}{L}$ for $k = 1, 2 \ldots, K$, then it follows from (55) that

$$\begin{aligned}
&\mathbb{E}[\mathcal{V}^{k+1}] - \mathbb{E}[\mathcal{V}^k] \\
\le& -\frac{\alpha_k(1-\beta_1)}{2}\left(\epsilon + \frac{\sigma^2}{1-\beta_2}\right)^{-\frac{1}{2}} \mathbb{E}\left[\left\|\nabla \mathcal{L}(\theta^k)\right\|^2\right] + (2-\beta_1)\frac{6\alpha_k^2 DL\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2 \\
&+ \underbrace{\left(\frac{(2-\beta_1)^2}{(1-\beta_1)}\alpha_k\sigma^2 - b_k\right)}_{A^k}\mathbb{E}\Big[\sum_{i=1}^{p}\left((\epsilon + \hat{v}_i^{k-D})^{-\frac{1}{2}} - (\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\right)\Big] \\
&+ \left(\frac{L}{2} + \rho_1 + (1-\beta_1)^{-1}L\right)\mathbb{E}\left[\|\theta^{k+1} - \theta^k\|^2\right] \\
&+ \sum_{d=1}^{D}\underbrace{\left((2-\beta_1)\epsilon^{-\frac{1}{2}}\left(\frac{L}{12} + \frac{c\alpha_k}{2d_{\max}}\right) + \rho_{d+1} - \rho_d\right)}_{B_d^k}\mathbb{E}\left[\|\theta^{k+1-d} - \theta^{k-d}\|^2\right].
\end{aligned} \tag{56}$$

To ensure $A^k \le 0$ and $B_d^k \le 0$, it is sufficient to choose $\{b_k\}$ and $\{\rho_d\}$ satisfying (with $\rho_{D+1} = 0$)

$$\frac{(2-\beta_1)^2}{(1-\beta_1)}\alpha_k\sigma^2 - b_k \le 0, \quad k = 1, \cdots, K \tag{57}$$

$$(2-\beta_1)\epsilon^{-\frac{1}{2}}\left(\frac{L}{12} + \frac{c\alpha_k}{2d_{\max}}\right) + \rho_{d+1} - \rho_d \le 0, \quad d = 1, \cdots, D. \tag{58}$$

Solve this system of linear equations and get

$$b_k = \frac{(2-\beta_1)^2}{(1-\beta_1)L}\sigma^2, \quad k = 1, \cdots, K \tag{59}$$

$$\rho_d = (2-\beta_1)\epsilon^{-\frac{1}{2}}\left(\frac{L}{12} + \frac{c}{2Ld_{\max}}\right)(D-d+1), \quad d = 1, \cdots, D \tag{60}$$

plugging which into (56) leads to the conclusion of Lemma 3.

11 Proof of Theorem 1

From the definition of $\mathcal{V}^k$, we have for any k, that

$$\begin{aligned}
\mathbb{E}[\mathcal{V}^k] &\ge \mathcal{L}(\theta^k) - \mathcal{L}(\theta^*) - c_k\left\langle \nabla\mathcal{L}(\theta^{k-1}), (\epsilon I + \hat{V}^k)^{-\frac{1}{2}}h^k\right\rangle + \sum_{d=1}^{D}\rho_d\|\theta^{k+1-d} - \theta^{k-d}\|^2 \\
&\ge -|c_k|\left\|\nabla\mathcal{L}(\theta^{k-1})\right\|\left\|(\epsilon I + \hat{V}^k)^{-\frac{1}{2}}h^k\right\| \\
&\ge -(1-\beta_1)^{-1}\alpha_k\sigma^2\epsilon^{-\frac{1}{2}}
\end{aligned} \tag{61}$$

where we use Assumption 2 and Lemma 5.

By taking summation on (56) over $k = 0, \cdots, K-1$, it follows from that

$$\begin{aligned}
&\frac{\alpha(1-\beta_1)}{2}\left(\epsilon+\frac{\sigma^2}{1-\beta_2}\right)^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\theta^k)\right\|^2\right]\\
\leq&\frac{\mathbb{E}[\mathcal{V}^1]-\mathbb{E}[\mathcal{V}^{K+1}]}{K}+(2-\beta_1)\frac{6\alpha^2DL\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2+\frac{(2-\beta_1)^2}{(1-\beta_1)}\sigma^2pD\epsilon^{-\frac{1}{2}}\frac{\alpha}{K}\\
&+\left(\frac{L}{2}+\rho_1+(1-\beta_1)^{-1}L\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\theta^{k+1}-\theta^k\|^2\right]\\
\overset{(a)}{\leq}&\frac{\mathbb{E}[\mathcal{V}^1]}{K}+(2-\beta_1)\frac{6\alpha^2DL\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2+(1-\beta_1)^{-1}\sigma^2\epsilon^{-\frac{1}{2}}\frac{\alpha}{K}+\frac{(2-\beta_1)^2}{(1-\beta_1)}\sigma^2pD\epsilon^{-\frac{1}{2}}\frac{\alpha}{K}\\
&+\left(\frac{L}{2}+\rho_1+(1-\beta_1)^{-1}L\right)p(1-\beta_2)^{-1}(1-\beta_3)^{-1}\alpha^2
\end{aligned}\tag{62}$$

where (a) follows from (61) and Lemma 6.

Specifically, if we choose a constant stepsize $\alpha := \frac{\eta}{\sqrt{K}}$, where $\eta > 0$ is a constant, and define

$$\tilde{C}_1 := (2-\beta_1)6DL\epsilon^{-\frac{1}{2}}\tag{63}$$

and

$$\tilde{C}_2 := (1-\beta_1)^{-1}\epsilon^{-\frac{1}{2}}+\frac{(2-\beta_1)^2}{(1-\beta_1)}D\epsilon^{-\frac{1}{2}}\tag{64}$$

and

$$\tilde{C}_3 := \left(\frac{L}{2}+\rho_1+(1-\beta_1)^{-1}L\right)(1-\beta_2)^{-1}(1-\beta_3)^{-1}\tag{65}$$

and

$$\tilde{C}_4 := \frac{1}{2}(1-\beta_1)\left(\epsilon+\frac{\sigma^2}{1-\beta_2}\right)^{-\frac{1}{2}}\tag{66}$$

we can obtain from (62) that

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta^k)\|^2\right]\leq&\frac{\frac{\mathcal{L}(\theta^0)-\mathcal{L}(\theta^*)}{K}+\frac{\tilde{C}_1}{M}\sum_{m\in\mathcal{M}}\sigma_m^2\alpha^2+\tilde{C}_2p\sigma^2\frac{\alpha}{K}+\tilde{C}_3p\alpha^2}{\alpha\tilde{C}_4}\\
\leq&\frac{\mathcal{L}(\theta^0)-\mathcal{L}(\theta^*)}{K\alpha\tilde{C}_4}+\frac{\tilde{C}_1\alpha}{\tilde{C}_4M}\sum_{m\in\mathcal{M}}\sigma_m^2+\tilde{C}_2p\frac{\sigma^2}{K\tilde{C}_4}+\frac{\tilde{C}_3p\alpha}{\tilde{C}_4}\\
=&\frac{(\mathcal{L}(\theta^0)-\mathcal{L}(\theta^*))C_4}{\sqrt{K}\eta}+\frac{C_1\eta}{\sqrt{K}M}\sum_{m\in\mathcal{M}}\sigma_m^2+\frac{C_2p\sigma^2}{K}+\frac{C_3p\eta}{\sqrt{K}}
\end{aligned}$$

where we define $C_1 := \tilde{C}_1/\tilde{C}_4$, $C_2 := \tilde{C}_2/\tilde{C}_4$, $C_3 := \tilde{C}_3/\tilde{C}_4$, and $C_4 := 1/\tilde{C}_4$.

12 Proof of Theorem 2

By the PL-condition of $\mathcal{L}(\theta)$, we have

$$\begin{aligned}
&-\frac{\alpha_k(1-\beta_1)}{2}\Big(\epsilon+\frac{\sigma^2}{1-\beta_2}\Big)^{-\frac{1}{2}}\mathbb{E}\left[\Big\|\nabla\mathcal{L}(\theta^k)\Big\|^2\right]\\
\le&-\alpha_k\mu(1-\beta_1)\Big(\epsilon+\frac{\sigma^2}{1-\beta_2}\Big)^{-\frac{1}{2}}\mathbb{E}\Big[\mathcal{L}(\theta^k)-\mathcal{L}(\theta^\star)\Big]\\
\overset{(a)}{\le}&-2\alpha_k\mu\tilde{C}_4\Big(\mathbb{E}[\mathcal{V}^k]+c_k\left\langle\nabla\mathcal{L}(\theta^{k-1}),(\epsilon I+\hat{V}^k)^{-\frac{1}{2}}h^k\right\rangle-b_k\sum_{d=0}^{D}\sum_{i=1}^{p}(\epsilon+\hat{v}_i^{k-d})^{-\frac{1}{2}}-\sum_{d=1}^{D}\rho_d\|\theta^{k+1-d}-\theta^{k-d}\|^2\Big)\\
\overset{(b)}{\le}&-2\alpha_k\mu\tilde{C}_4\mathbb{E}[\mathcal{V}^k]+2\alpha_k^2\mu\tilde{C}_4(1-\beta_1)^{-1}\sigma^2\epsilon^{-\frac{1}{2}}+2\alpha_k\mu\tilde{C}_4b_k\sum_{d=0}^{D}\sum_{i=1}^{p}\mathbb{E}\left[(\epsilon+\hat{v}_i^{k-d})^{-\frac{1}{2}}\right]\\
&+2\alpha_k\mu\tilde{C}_4\sum_{d=1}^{D}\rho_d\mathbb{E}[\|\theta^{k+1-d}-\theta^{k-d}\|^2]
\end{aligned}\tag{67}$$

where (a) uses the definition of $\tilde{C}_4$ in (66), and (b) uses Assumption 2 and Lemma 5.

Plugging (67) into (55), we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{V}^{k+1}]-\mathbb{E}[\mathcal{V}^k]\\
\le&-2\alpha_k\mu\tilde{C}_4\mathbb{E}[\mathcal{V}^k]+(2-\beta_1)\frac{6\alpha_k^2DL\epsilon^{-\frac{1}{2}}}{M}\sum_{m\in\mathcal{M}}\sigma_m^2\\
&+\frac{(2-\beta_1)^2}{(1-\beta_1)}\alpha_k\sigma^2\mathbb{E}\Big[\sum_{i=1}^{p}\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\\
&+b_{k+1}\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big]-(b_k-2\alpha_k\mu\tilde{C}_4b_k)\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}\Big]\\
&+\sum_{d=1}^{D}(b_{k+1}-b_k+2\alpha_k\mu\tilde{C}_4b_k)\sum_{i=1}^{p}\mathbb{E}\Big[(\epsilon+\hat{v}_i^{k+1-d})^{-\frac{1}{2}}\Big]\\
&+\left(\frac{L}{2}+\rho_1+(1-\beta_1)^{-1}L\right)p(1-\beta_2)^{-1}(1-\beta_3)^{-1}\alpha_k^2+2\alpha_k^2\mu\tilde{C}_4(1-\beta_1)^{-1}\sigma^2\epsilon^{-\frac{1}{2}}\\
&+\sum_{d=1}^{D}\Big((2-\beta_1)\epsilon^{-\frac{1}{2}}\Big(\frac{L}{12}+\frac{c\alpha_k}{2d_{\max}}\Big)+\rho_{d+1}-\rho_d+2\alpha_k\mu\tilde{C}_4\rho_d\Big)\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right].
\end{aligned}\tag{68}$$

If we choose b_k to ensure that $b_{k+1}\le(1-2\alpha_k\mu\tilde{C}_4)b_k$, then we can obtain from (68) that

$$\begin{aligned}
&\mathbb{E}[\mathcal{V}^{k+1}]-\mathbb{E}[\mathcal{V}^k]\\
\le&-2\alpha_k\mu\tilde{C}_4\mathbb{E}[\mathcal{V}^k]+\frac{\tilde{C}_1}{M}\sum_{m\in\mathcal{M}}\sigma_m^2\alpha_k^2+\tilde{C}_3p\alpha_k^2+2\mu\tilde{C}_4(1-\beta_1)^{-1}\sigma^2\epsilon^{-\frac{1}{2}}\alpha_k^2\\
&+\left(\frac{(2-\beta_1)^2}{(1-\beta_1)}\alpha_k\sigma^2-(1-2\alpha_k\mu\tilde{C}_4)b_k\right)\mathbb{E}\Big[\sum_{i=1}^{p}\Big((\epsilon+\hat{v}_i^{k-D})^{-\frac{1}{2}}-(\epsilon+\hat{v}_i^{k+1})^{-\frac{1}{2}}\Big)\Big]\\
&+\sum_{d=1}^{D}\Big((2-\beta_1)\epsilon^{-\frac{1}{2}}\Big(\frac{L}{12}+\frac{c\alpha_k}{2d_{\max}}\Big)+\rho_{d+1}-\rho_d+2\alpha_k\mu\tilde{C}_4\rho_d\Big)\mathbb{E}\left[\|\theta^{k+1-d}-\theta^{k-d}\|^2\right].
\end{aligned}\tag{69}$$

If $\alpha_k\le\frac{1}{L}$, we choose parameters $\{b_k,\rho_d\}$ to guarantee that

$$\frac{(2-\beta_1)^2}{(1-\beta_1)L}\sigma^2-\Big(1-\frac{2\mu\tilde{C}_4}{L}\Big)b_k\le0,\quad\forall k\tag{70}$$

$$(2-\beta_1)\Big(\frac{L}{12}+\frac{c}{2Ld_{\max}}\Big)\epsilon^{-\frac{1}{2}}+\rho_{d+1}-\Big(1-\frac{2\mu\tilde{C}_4}{L}\Big)\rho_d\le0,\quad d=1,\cdots,D\tag{71}$$

and choose $\beta_1, \beta_2, \epsilon$ to ensure that $1 - \frac{2\mu\tilde{C}_4}{L} \geq 0$.

Then we have

$$\mathbb{E}[\mathcal{V}^{k+1}] \leq \left(1 - 2\alpha_k\mu\tilde{C}_4\right)\mathbb{E}[\mathcal{V}^k] + \underbrace{\left(\frac{\tilde{C}_1}{M}\sum_{m\in\mathcal{M}}\sigma_m^2 + \tilde{C}_3 p + 2\mu\tilde{C}_4(1-\beta_1)^{-1}\sigma^2\epsilon^{-\frac{1}{2}}\right)}_{\tilde{C}_5}\alpha_k^2$$

$$\leq \prod_{j=0}^{k}(1 - 2\alpha_j\mu\tilde{C}_4)\mathbb{E}[\mathcal{V}^0] + \sum_{j=0}^{k}\alpha_j^2\prod_{i=j+1}^{k}(1 - 2\alpha_i\mu\tilde{C}_4)\tilde{C}_5. \tag{72}$$

If we choose $\alpha_k = \frac{1}{\mu(k+K_0)\tilde{C}_4} \leq \frac{1}{L}$, where K_0 is a sufficiently large constant to ensure that α_k satisfies the aforementioned conditions, then we have

$$\begin{aligned}
\mathbb{E}[\mathcal{V}^K] \leq& \mathbb{E}[\mathcal{V}^0]\prod_{k=0}^{K-1}(1 - 2\alpha_k\mu\tilde{C}_4) + \tilde{C}_5\sum_{k=0}^{K-1}\alpha_k^2\prod_{j=k+1}^{K-1}(1 - 2\alpha_j\mu\tilde{C}_4)\\
\leq& \mathbb{E}[\mathcal{V}^0]\prod_{k=0}^{K-1}\frac{k+K_0-2}{k+K_0} + \frac{\tilde{C}_5}{\mu^2\tilde{C}_4^2}\sum_{k=0}^{K-1}\frac{1}{(k+K_0)^2}\prod_{j=k+1}^{K-1}\frac{j+K_0-2}{j+K_0}\\
\leq& \frac{(K_0-2)(K_0-1)}{(K+K_0-2)(K+K_0-1)}\mathbb{E}[\mathcal{V}^0] + \frac{\tilde{C}_5}{\mu^2\tilde{C}_4^2}\sum_{k=0}^{K-1}\frac{(k+K_0-1)}{(k+K_0)(K+K_0-2)(K+K_0-2)}\\
\leq& \frac{(K_0-1)^2}{(K+K_0-1)^2}\mathbb{E}[\mathcal{V}^0] + \frac{\tilde{C}_5 K}{\mu^2\tilde{C}_4^2(K+K_0-1)^2}\\
=& \frac{(K_0-1)^2}{(K+K_0-1)^2}(\mathcal{L}(\theta^0) - \mathcal{L}(\theta^\star)) + \frac{\tilde{C}_5 K}{\mu^2\tilde{C}_4^2(K+K_0-2)^2}
\end{aligned}$$

from which the proof is complete.

13 Additional Numerical Results

13.1 Logistic regression.

Data pre-processing. For *ijcnn1* and *covtype* datasets, they are imported from the popular library LIBSVM[1] without further preprocessing. For *MNIST*, we normalize the data and subtract the mean. We uniformly partition *ijcnn1* dataset with 91,701 samples and *MNIST* dataset with 60,000 samples into $M = 10$ workers. To simulate the heterogeneous setting, we partition *covtype* dataset with 581,012 samples randomly into $M = 20$ workers with different number of samples per worker.

For *covtype*, we fix the batch ratio to be 0.001 uniformly across all workers; and for *ijcnn1* and *MNIST*, we fix the batch ratio to be 0.01 uniformly across all workers.

Choice of hyperparameters. For the logistic regression task, the hyperparameters in each algorithm are chosen by hand to roughly optimize the training loss performance of each algorithm. We list the values of parameters used in each test in Tables 1-2.

Algorithm	**stepsize** α	**momentum weight** β	**averaging interval** H/D
FedAdam	$\alpha_l = 100$ $\alpha_s = 0.02$	0.9	$H = 10$
Local momentum	0.1	0.9	$H = 10$
ADAM	0.005	$\beta_1 = 0.9$ $\beta_2 = 0.999$	/
CADA1&2	0.005	$\beta_1 = 0.9$ $\beta_2 = 0.999$	$D = 100$, $d_{\max} = 10$
Stochastic LAG	0.1	/	$d_{\max} = 10$

Table 1: Choice of parameters in *covtype*.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Algorithm	stepsize α	momentum weight β	averaging interval H/D
FedAdam	$\alpha_l = 100$ $\alpha_s = 0.03$	0.9	$H = 10$
Local momentum	0.1	0.9	$H = 20$
ADAM	0.01	$\beta_1 = 0.9$ $\beta_2 = 0.999$	/
CADA	0.01	$\beta_1 = 0.9$ $\beta_2 = 0.999$	$D = 100$, $d_{\max} = 10$
Stochastic LAG	0.1	/	$d_{\max} = 10$

Table 2: Choice of parameters in *ijcnn1*.

13.2 Training neural networks.

For training neural networks, we use the cross-entropy loss but with different neural network models.

Neural network models. For *MNIST* dataset, we use a convolutional neural network with two convolution-ELUmaxpooling layers (ELU is a smoothed ReLU) followed by two fully-connected layers. The first convolution layer is $5 \times 5 \times 20$ with padding, and the second layer is $5 \times 5 \times 50$ with padding. The output of second layer is followed by two fully connected layers with one being 800×500 and the other being 500×10. The output goes through a softmax function. For *CIFAR10* dataset, we use the popular neural network architecture *ResNet20* [2] which has 20 and roughly 0.27 million parameters. We do not use a pre-trained model.

Data pre-processing. We uniformly partition *MNIST* and *CIFAR10* datasets into $M = 10$ workers. For *MNIST*, we use the raw data without preprocessing. The minibatch size per worker is 12. For *CIFAR10*, in addition to normalizing the data and subtracting the mean, we randomly flip and crop part of the original image every time it is used for training. This is a standard technique of data augmentation to avoid over-fitting. The minibatch size for *CIFAR10* is 50 per worker.

Choice of hyperparameters. For *MNIST* dataset which is relatively easy, the hyperparameters in each algorithm are chosen by hand to optimize the performance of each algorithm. We list the values of parameters used in each test in Table 3.

Algorithm	stepsize α	momentum weight β	averaging interval H/D
FedAdam	$\alpha_l = 0.1$ $\alpha_s = 0.001$	0.9	$H = 8$
Local momentum	0.001	0.9	$H = 8$
ADAM	0.0005	$\beta_1 = 0.9$ $\beta_2 = 0.999$	/
CADA1&2	0.0005	$\beta_1 = 0.9$ $\beta_2 = 0.999$	$D = 50, d_{\max} = 10$
Stochastic LAG	0.1	/	$d_{\max} = 10$

Table 3: Choice of parameters in multi-class *MNIST*.

For *CIFAR10* dataset, we search the best values of hyperparameters from the following search grid on a per-algorithm basis to optimize the testing accuracy versus the number of communication rounds. The chosen values of parameter are listed in Table 4.

FedAdam: $\alpha_s \in \{0.1, 0.01, 0.001\}$; $\alpha_l \in \{1, 0.5, 0.1\}$; $H \in \{1, 4, 6, 8, 16\}$.

Local momentum: $\alpha \in \{0.1, 0.01, 0.001\}$; $H \in \{1, 4, 6, 8, 16\}$.

CADA1: $\alpha \in \{0.1, 0.01, 0.001\}$; $c \in \{0.05, 0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$.

CADA2: $\alpha \in \{0.1, 0.01, 0.001\}$; $c \in \{0.05, 0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$.

LAG: $\alpha \in \{0.1, 0.01, 0.001\}$; $c \in \{0.05, 0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$.

[2] https://github.com/akamaster/pytorch_resnet_cifar10

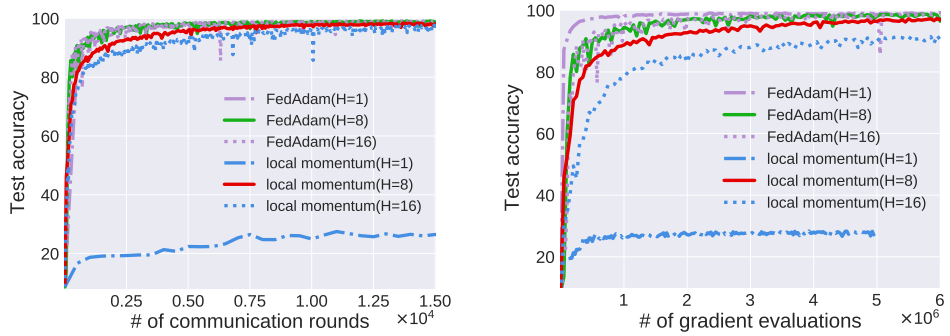

Figure 6: Performance of FedAdam and local momentum on *MNIST* under different averaging interval H.

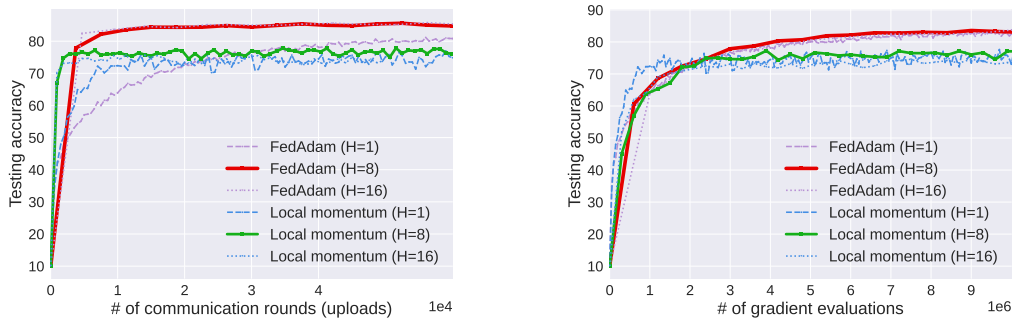

Figure 7: Performance of FedAdam and local momentum on *CIFAR10* under different averaging interval H.

Algorithm	**stepsize** α	**momentum weight** β	**averaging interval** H/D
FedAdam	$\alpha_l = 0.1$ $\alpha_s = 0.1$	0.9	$H = 8$
Local momentum	0.1	0.9	$H = 8$
CADA1	0.1	$\beta_1 = 0.9$ $\beta_2 = 0.99$	$D = 50$, $d_{\max} = 2$
CADA2	0.1	$\beta_1 = 0.9$ $\beta_2 = 0.99$	$D = 50$, $d_{\max} = 2$
Stochastic LAG	0.1	/	$d_{\max} = 2$

Table 4: Choice of parameters in *CIFAR10*.

Additional results. In addition to the results presented in the main paper, we report a new set of simulations on the performance of local update based algorithms under different averaging interval H. Since algorithms under $H = 4, 6$ do not perform as good as $H = 8$, we only plot $H = 1, 8, 16$ in Figures 6 and 7 to ease the comparison. Figure 6 compares the performance of FedAdam and local momentum on *MNIST* dataset under different averaging interval H. Figure 7 compares the performance of FedAdam and local momentum on *CIFAR10* dataset under different H.

Figure 7 compares the performance of FedAdam and local momentum on *CIFAR10* dataset under different averaging interval H. FedAdam and local momentum under a larger averaging interval H have faster convergence speed at the initial stage, but they reach slightly lower testing accuracy. This reduced test accuracy is common among local SGD-type methods, which has also been studied theoretically; see e.g., [14].