Supplementary Materials for

Stability-Based Generalization Analysis of the Asynchronous Decentralized SGD

The supplementary material contains the full experimental results and detailed proofs of our theoretical findings.

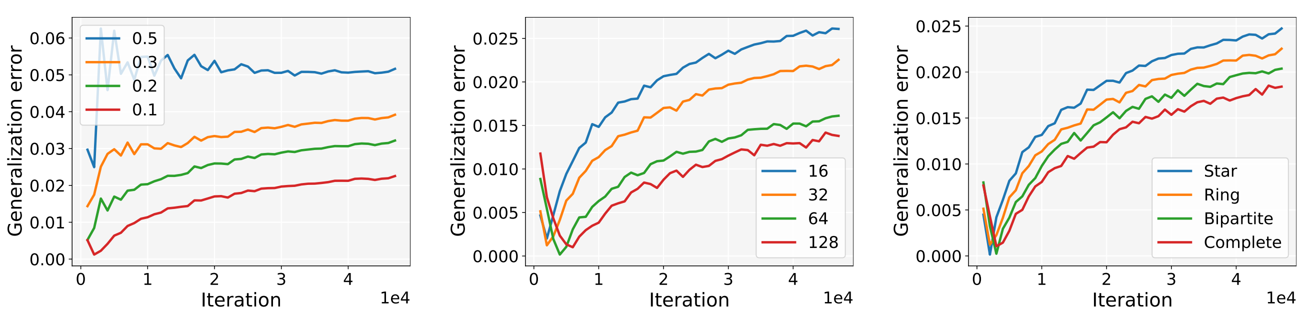
Contents

A More Experimental Results

In AD-SGD (Lian et al. 2018), the communication topology is designed as a bipartite graph in order to prevent the deadlock problem. The topologies that we have employed (as shown in Figure 3) all satisfy this property. Consider a distributed system with 16 computing workers, the corresponding doubly stochastic matrix of the four topologies are

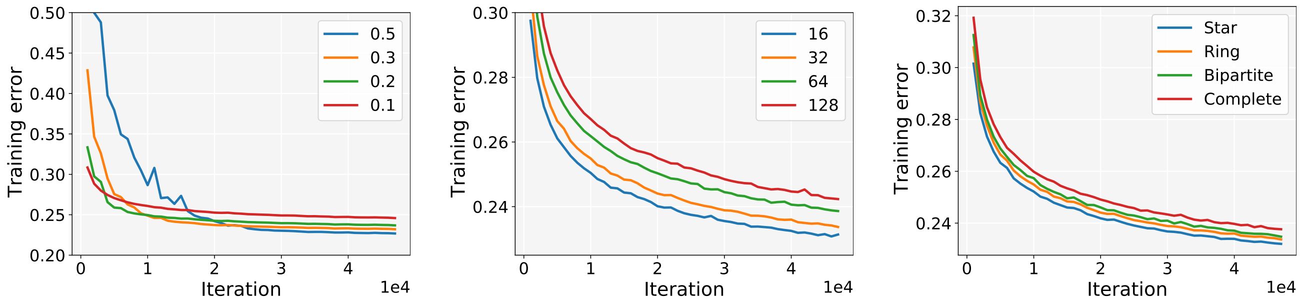
$$\mathbf{W}_{\text{comp}} = \begin{pmatrix} \frac{1}{16} & \cdots & \frac{1}{16} \\ \frac{1}{16} & \cdots & \frac{1}{16} \\ \vdots & \ddots & \vdots \\ \frac{1}{16} & \cdots & \frac{1}{16} \\ \frac{1}{16} & \cdots & \frac{1}{16} \end{pmatrix} \mathbf{W}_{\text{bipa}} = \begin{pmatrix} \frac{1}{9} & \frac{1}{9} & \cdots & 0 & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \cdots & \frac{1}{9} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \frac{1}{9} & \cdots & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & 0 & \cdots & \frac{1}{9} & \frac{1}{9} \end{pmatrix} \mathbf{W}_{\text{ring}} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \cdots & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \cdots & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \mathbf{W}_{\text{star}} = \begin{pmatrix} \frac{1}{16} & \frac{1}{16} & \cdots & \frac{1}{16} \\ \frac{1}{16} & \frac{15}{16} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{16} & 0 & \cdots & 0 \\ \frac{1}{16} & 0 & \cdots & \frac{15}{16} \end{pmatrix}$$

In the following, we will show more experimental results, including the performance of convex models with decreasing learning rate; non-convex ResNet-18 and VGG-16 on the CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. The experimental observations are consistent with the theoretical analysis and description of the experimental results in the main text.



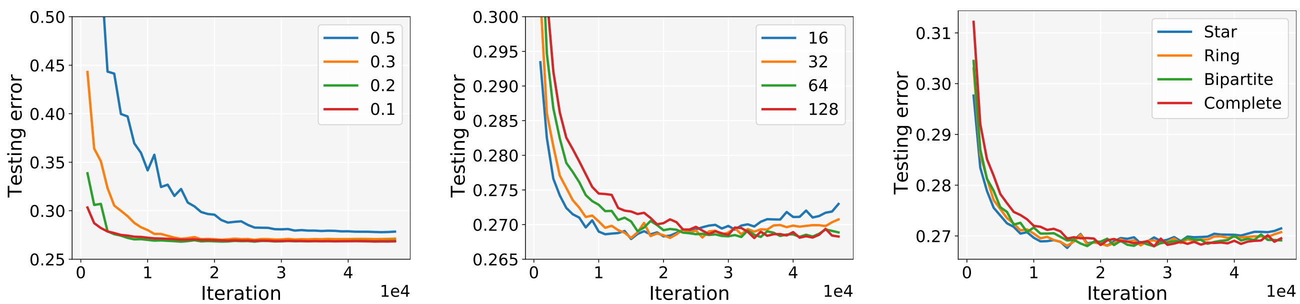
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 1: Convex model on the MNIST dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies with the decreasing learning rate. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology. Decreasing learning rate $\alpha_t = \frac{\alpha}{1+0.01t}$ with varying α; (b). Fixed $\alpha_t = \frac{0.1}{1+0.01t}$, ring topology; (c). Fixed $\alpha_t = \frac{0.1}{1+0.01t}$, $\overline{\tau} = 32$.



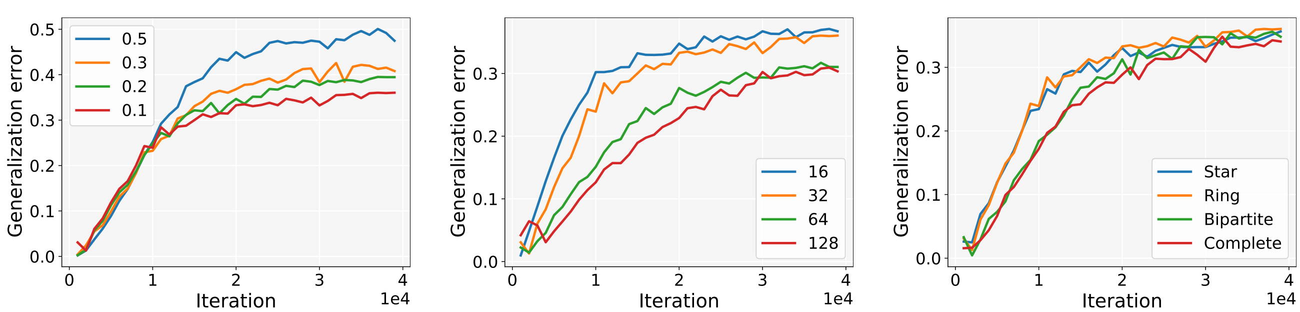
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 2: Convex model on the MNIST dataset. Training errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology. Decreasing learning rate $\alpha_t = \frac{\alpha}{1+0.01t}$ with varying α; (b). Fixed $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



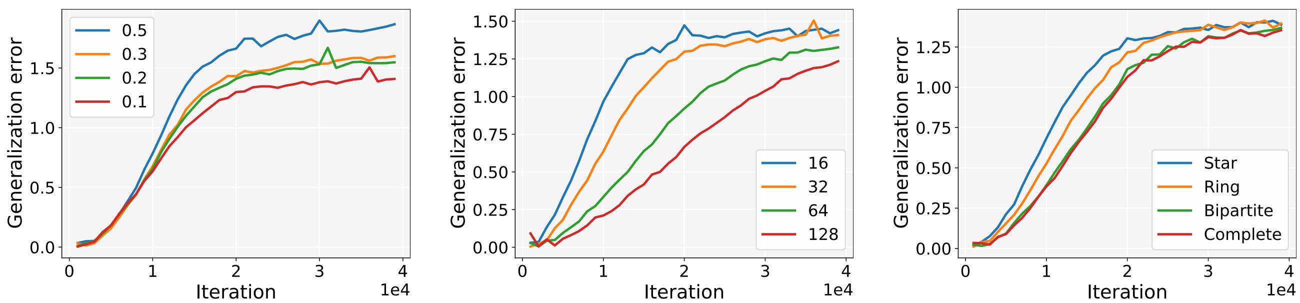
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 3: Convex model on the MNIST dataset. Testing errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology. Decreasing learning rate $\alpha_t = \frac{\alpha}{1+0.01t}$ with varying α; (b). Fixed $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



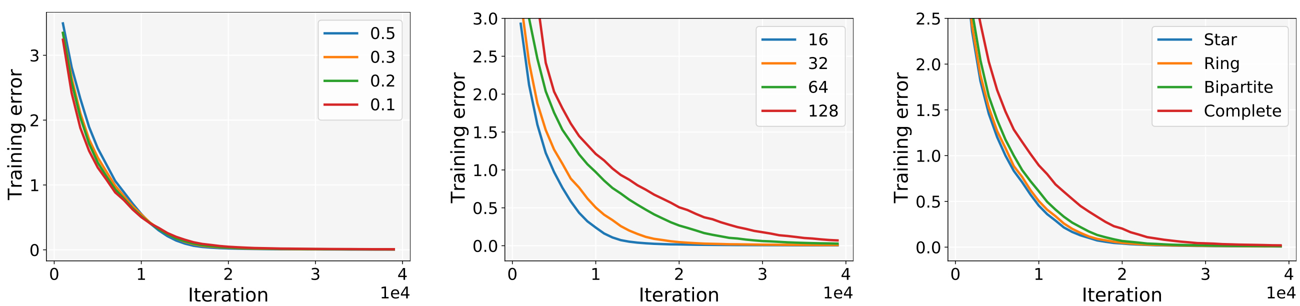
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 4: Non-convex ResNet-18 on the CIFAR-10 dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



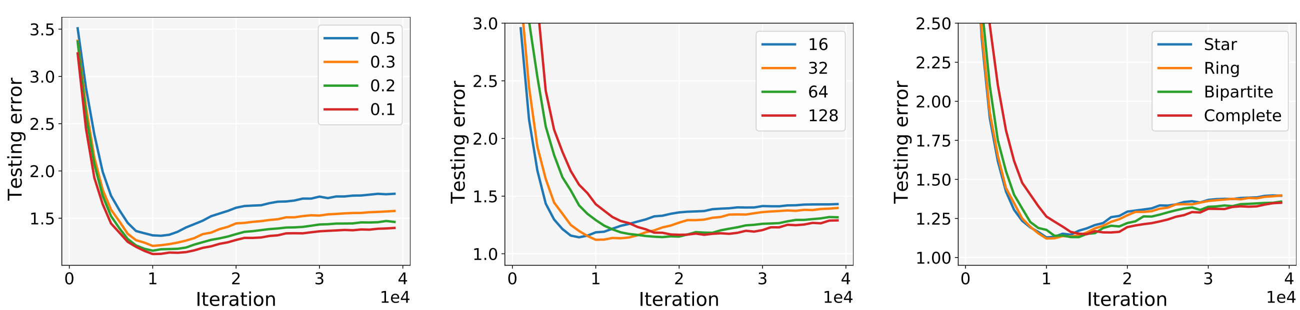
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 5: Non-convex ResNet-18 on the CIFAR-100 dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



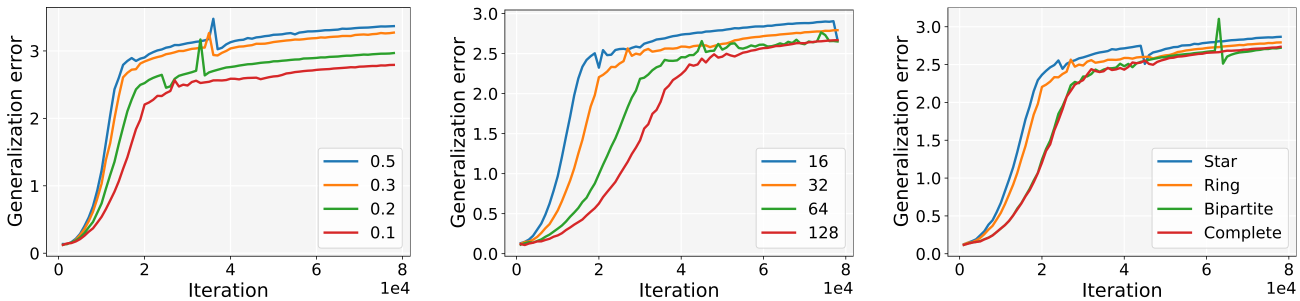
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 6: Non-convex ResNet-18 on the CIFAR-100 dataset. Training errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology. Decreasing learning rate $\alpha_t = \frac{\alpha}{1+0.01t}$ with varying α; (b). Fixed $\alpha_t = \frac{0.1}{1+0.01t}$, ring topology; (c). Fixed $\alpha_t = \frac{0.1}{1+0.01t}, \overline{\tau} = 32$.



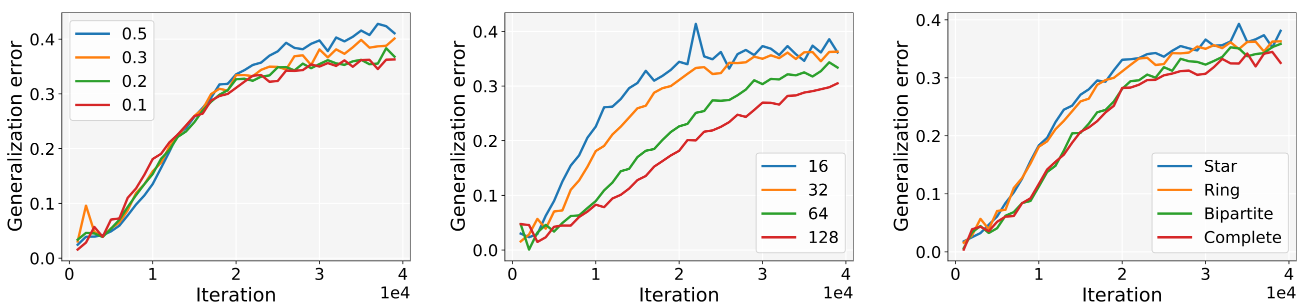
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 7: Non-convex ResNet-18 on the CIFAR-100 dataset. Testing errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology. Decreasing learning rate $\alpha_t = \frac{\alpha}{1+0.01t}$ with varying α; (b). Fixed $\alpha_t = \frac{0.1}{1+0.01t}$, ring topology; (c). Fixed $\alpha_t = \frac{0.1}{1+0.01t}, \overline{\tau} = 32$.



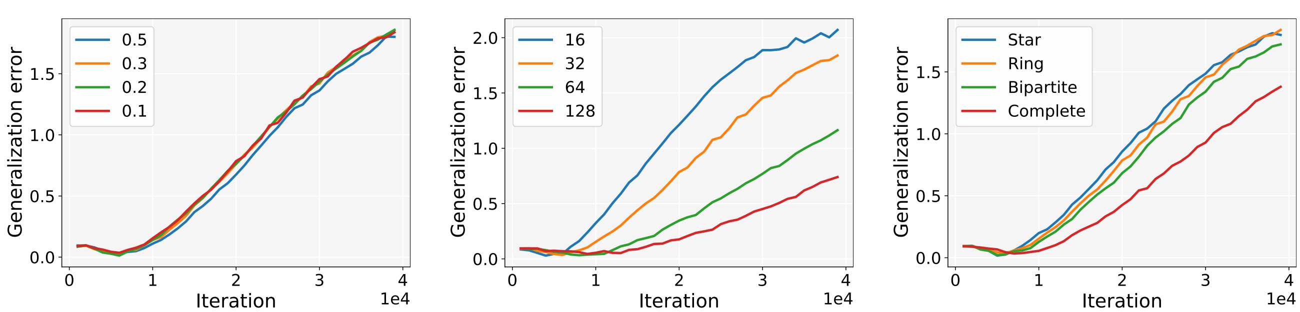
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 8: Non-convex ResNet-18 on the Tiny-ImageNet dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



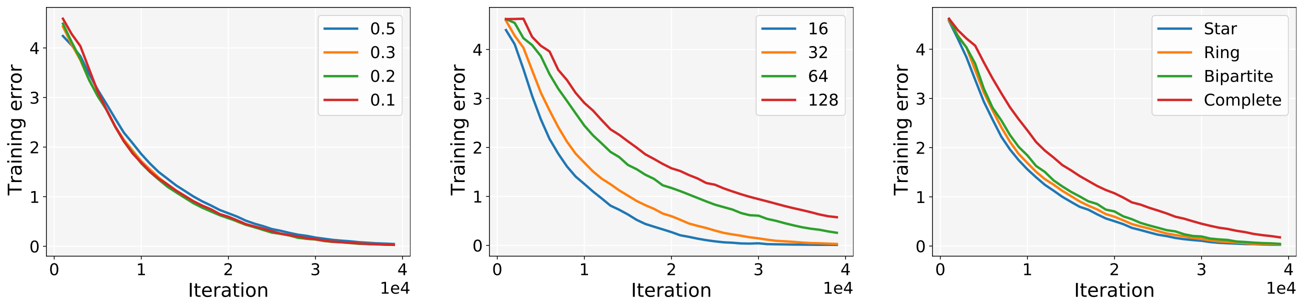
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 9: Non-convex VGG-16 on the CIFAR-10 dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



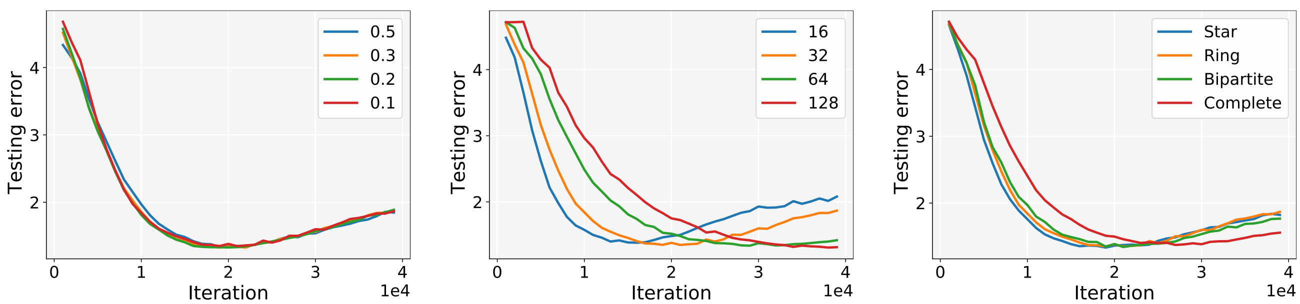
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 10: Non-convex VGG-16 on the CIFAR-100 dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



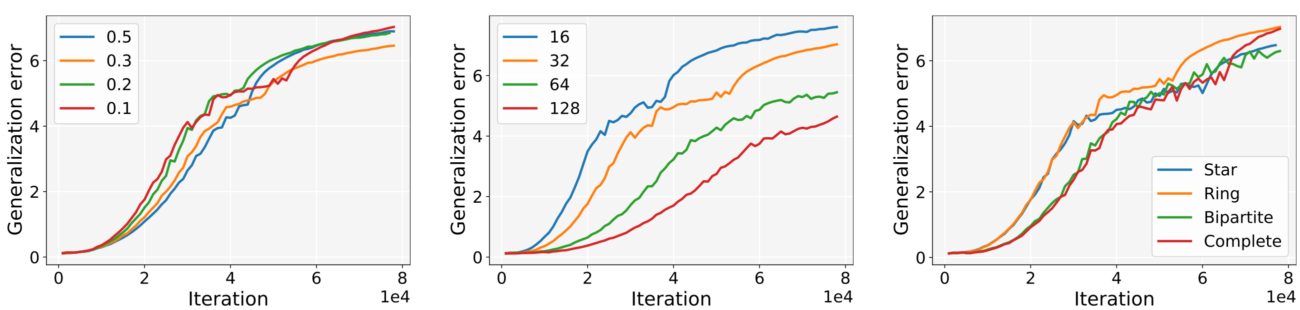
(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 11: Non-convex VGG-16 on the CIFAR-100 dataset. Training errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 12: Non-convex VGG-16 on the CIFAR-100 dataset. Testing errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.



(a) Learning rate (b) Asynchronous delay (c) Decentralized topology

Figure 13: Non-convex VGG-16 on the Tiny-ImageNet dataset. Generalization errors for varying learning rates, asynchronous delays, and decentralized topologies. (a). Fixed maximum delay $\overline{\tau} = 32$, ring topology; (b). Fixed learning rate $\alpha = 0.1$, ring topology; (c). Fixed $\alpha = 0.1, \overline{\tau} = 32$.

B Missing Theoretical Proofs

B.1 Properties and Technical Lemmas

From the iterative format of AD-SGD, i.e.,

$$\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{W} - \alpha_t \mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t}), \tag{B.1}$$

the consensus model has the following recursive property

$$\begin{aligned}
\mathbf{x}_{t+1} &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_{t+1}(i) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{m} w_{i,k}\mathbf{x}_t(k) - \alpha_t \frac{1}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_t(i) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) \\
&= \mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}).
\end{aligned} \tag{B.2}$$

Lemma 4 (Lemma 3.7, (Hardt, Recht, and Singer 2016)) *The following properties hold for every* $\mathbf{z}$.

1. *Assume that f is β-smooth. Then*

$$\left\|\mathbf{x} - \frac{\alpha}{m}\nabla f(\mathbf{x};\mathbf{z}) - \mathbf{x}' + \frac{\alpha}{m}\nabla f(\mathbf{x}';\mathbf{z})\right\| \le (1 + \frac{\beta\alpha}{m})\|\mathbf{x} - \mathbf{x}'\|. \tag{B.3}$$

2. *Assume that f is β-smooth, convex. Then for any $\alpha \le 2m/\beta$*

$$\left\|\mathbf{x} - \frac{\alpha}{m}\nabla f(\mathbf{x};\mathbf{z}) - \mathbf{x}' + \frac{\alpha}{m}\nabla f(\mathbf{x}';\mathbf{z})\right\| \le \|\mathbf{x} - \mathbf{x}'\|. \tag{B.4}$$

3. *Assume that f is β-smooth, μ-strongly convex. Then for any $\alpha \le m/\beta$*

$$\left\|\mathbf{x} - \frac{\alpha}{m}\nabla f(\mathbf{x};\mathbf{z}) - \mathbf{x}' + \frac{\alpha}{m}\nabla f(\mathbf{x}';\mathbf{z})\right\| \le (1 - \frac{\mu\alpha}{m})\|\mathbf{x} - \mathbf{x}'\|. \tag{B.5}$$

Lemma 5 *For any $0 < \lambda < 1$ and $t \in \mathbb{Z}^+$, it holds*

$$\sum_{s=1}^{t-1} \frac{\lambda^{t-1-s}}{s+1} \le \frac{C_\lambda}{t}, \tag{B.6}$$

where $C_\lambda = \frac{8}{\lambda e^2 \ln^2 \frac{1}{\lambda}} + \frac{2}{\lambda \ln \frac{1}{\lambda}}$ is a constant.

Proof. The proof is very similar to [Lemma 5, (Sun, Li, and Wang 2021)], and we include a proof for completeness. For any $0 < \lambda < 1, x \in [s, s+1]$, we have that $\frac{\lambda^{t-1-s}}{s+1} \le \frac{\lambda^{t-1-x}}{x}$. Then

$$\begin{aligned}
\sum_{s=1}^{t-1} \frac{\lambda^{t-1-s}}{s+1} &\le \sum_{s=1}^{t-1}\int_s^{s+1} \frac{\lambda^{t-1-x}}{x}dx \le \lambda^{t-1}\int_1^t \frac{\lambda^{-x}}{x}dx \le \lambda^{t-1}\int_1^{\frac{t}{2}} \frac{\lambda^{-x}}{x}dx + \lambda^{t-1}\int_{\frac{t}{2}}^t \frac{\lambda^{-x}}{x}dx \\
&\le \lambda^{\frac{t}{2}-1}\int_1^{\frac{t}{2}} \frac{1}{x}dx + \frac{2\lambda^{t-1}}{t}\int_{\frac{t}{2}}^t \lambda^{-x}dx \le \lambda^{\frac{t}{2}-1}\ln(\frac{t}{2}) + \frac{2}{t\lambda \ln \frac{1}{\lambda}} \\
&\le \frac{t\lambda^{\frac{t}{2}-1}}{2} + \frac{2}{t\lambda \ln \frac{1}{\lambda}}.
\end{aligned}$$

Now, we provide the bound for $\sup_{t\ge 1}\{t^2\lambda^{\frac{t}{2}-1}\}$. It is easy to check that $t = 4/\ln\frac{1}{\lambda}$ achieves the maximum, which indicates

$$\sup_{t\ge 1}\{t^2\lambda^{\frac{t}{2}-1}\} \le \frac{16}{\lambda e^2 \ln^2 \frac{1}{\lambda}}.$$

In conclude, for $0 < \lambda < 1$

$$\sum_{s=1}^{t-1} \frac{\lambda^{t-1-s}}{s+1} \le \left[\frac{8}{\lambda e^2 \ln^2 \frac{1}{\lambda}} + \frac{2}{\lambda \ln \frac{1}{\lambda}}\right]\frac{1}{t}.$$

We then competed the proof.

■

B.2 Proof of Lemma 2

From the iterative format (B.1) of AD-SGD and the following notation

$$\mathbf{X}_t = [\mathbf{x}_t(1) \quad \mathbf{x}_t(2) \quad \cdots \quad \mathbf{x}_t(m)];$$
$$\mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t}) = [\mathbf{0} \quad \cdots \quad \mathbf{0} \quad \cdots \quad \nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) \quad \mathbf{0} \quad \cdots \quad \mathbf{0}],$$

we have that $\mathbf{x}_t = \frac{\mathbf{X}_t \mathbf{1}_m}{m}$, $\mathbf{x}_t(i) = \mathbf{X}_t \mathbf{e}_i$, where $\mathbf{e}_i$ is the column vector in $\mathbb{R}^m$ whose i-th element is 1. Then we can derive

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_{t+1}(i)\| &= \left\| \frac{\mathbf{X}_{t+1}\mathbf{1}_m}{m} - \mathbf{X}_{t+1}\mathbf{e}_i \right\| \\
&= \left\| \frac{\mathbf{X}_t \mathbf{W} \mathbf{1}_m - \alpha_t \mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t})\mathbf{1}_m}{m} - (\mathbf{X}_t \mathbf{W} \mathbf{e}_i - \alpha_t \mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t})\mathbf{e}_i) \right\| \\
&= \left\| \frac{\mathbf{X}_t \mathbf{1}_m - \alpha_t \mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t})\mathbf{1}_m}{m} - (\mathbf{X}_t \mathbf{W} \mathbf{e}_i - \alpha_t \mathbf{G}(\hat{\mathbf{X}}_t; \mathbf{z}_{j_t})\mathbf{e}_i) \right\| \\
&= \left\| \frac{\mathbf{X}_1 \mathbf{1}_m - \sum_{s=1}^t \alpha_s \mathbf{G}(\hat{\mathbf{X}}_s; \mathbf{z}_{j_s})\mathbf{1}_m}{m} - \left(\mathbf{X}_1 \mathbf{W}^t \mathbf{e}_i - \sum_{s=1}^t \alpha_s \mathbf{G}(\hat{\mathbf{X}}_s; \mathbf{z}_{j_s}) \mathbf{W}^{t-s} \mathbf{e}_i \right) \right\| \\
&\overset{(a)}{=} \left\| \sum_{s=1}^t \alpha_s \mathbf{G}(\hat{\mathbf{X}}_s; \mathbf{z}_{j_s}) \left(\frac{\mathbf{1}_m}{m} - \mathbf{W}^{t-s}\mathbf{e}_i \right) \right\| \\
&\overset{(b)}{\leq} L \sum_{s=1}^t \alpha_s \left\| \frac{\mathbf{1}_m}{m} - \mathbf{W}^{t-s}\mathbf{e}_i \right\| \\
&\overset{(c)}{\leq} L \sum_{s=1}^t \alpha_s \lambda^{t-s},
\end{aligned}$$

where (a) uses $\mathbf{x}_1(1) = \mathbf{x}_1(2) = \cdots = \mathbf{x}_1(m)$, which indicates $\mathbf{X}_1\mathbf{W} = \mathbf{X}_1 \quad \frac{\mathbf{X}_1 \mathbf{1}_m}{m} - \mathbf{X}_1 \mathbf{e}_i = 0, \forall\, i$. (b) uses the bounded gradient assumption, and (c) uses the properties of the doubly random matrix $\mathbf{W}$ ([Lemma 3, (Lian et al. 2018)]). Thus

$$\|\mathbf{x}_{t+1} - \mathbf{x}_{t+1}(i)\| \leq L \sum_{s=1}^t \alpha_s \lambda^{t-s}. \tag{B.7}$$

Remark 1 *If $t = 1$, we have that $\|\mathbf{x}_1 - \mathbf{x}_1(i)\| = 0$, then we define $\sum_{s=1}^{t-1} \alpha_s \lambda^{t-s}|_{t=1} = 0$.*

∎

B.3 Proof of Lemma 3

$$\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq \sum_{s=t-\tau_t}^{t-1} \|\mathbf{x}_{s+1} - \mathbf{x}_s\| \leq \sum_{s=t-\tau_t}^{t-1} \frac{\alpha_s}{m} \|\nabla f(\mathbf{x}_{s-\tau_s}(i_s); \mathbf{z}_{j_t(i_s)})\| \leq \frac{L}{m} \sum_{s=t-\tau_t}^{t-1} \alpha_s. \tag{B.8}$$

Remark 2 *If $\tau_t = 0$, we have that $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| = 0$, then we define $\sum_{s=t-\tau_t}^{t-1} \alpha_s|_{\tau_t=0} = 0$.*

∎

B.4 Proof of Theorem 1 (generalization error in the convex case)

Let $\mathcal{S} = \{\mathbf{z}_1, \cdots, \mathbf{z}_{j_*}, \cdots, \mathbf{z}_n\}$ and $\mathcal{S}' = \{\mathbf{z}_1, \cdots, \mathbf{z}'_{j_*}, \cdots, \mathbf{z}_n\}$ be two training dataset of size n differing in only a single example $\mathbf{z}_{j_*}$. $\mathbf{x}_T$ and $\mathbf{x}'_T$ denote the output model of running AD-SGD on $\mathcal{S}$ and $\mathcal{S}'$ for T iterations, respectively. For the two data dividing methods, the probability of AD-SGD selecting the same sample in both $\mathcal{S}$ and $\mathcal{S}'$ at the t-th iteration is $1 - \frac{1}{n}$, i.e., $j_t(i_t) \neq j_*$. Then we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| = \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\le \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_t(i_t)}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_t(i_t)})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\overset{(a)}{\le} \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}(i_t)\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_t - \mathbf{x}'_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}(i_t)\| \\
&\overset{(b)}{\le} \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2\beta\alpha_t}{m}\frac{L}{m}\sum_{s=t-\tau_t}^{t-1}\alpha_s + \frac{2\beta\alpha_t}{m}L\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} \\
&\le \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big),
\end{aligned} \tag{B.9}$$

where (a) uses the convexity (B.4) and the β-smoothness assumption; (b) uses inequalities (B.7), (B.8). With probability $\frac{1}{n}$ the selected example is different, i.e., $j_t(i_t) = j_*$. With the bounded gradient assumption, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| = \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_*}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}'_{j_*})\| \le \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2L\alpha_t}{m}. \tag{B.10}$$

Denote $\delta_t = \|\mathbf{x}_t - \mathbf{x}'_t\|$, then $\delta_1 = \|\mathbf{x}_1 - \mathbf{x}'_1\| = 0$. With inequalities (B.9) and (B.10), taking expectation of δ_{t+1} with respect to the randomness of the algorithm, we have

$$\begin{aligned}
\mathbb{E}[\delta_{t+1}] &\le (1 - \frac{1}{n})\mathbb{E}[\delta_t] + (1 - \frac{1}{n})\frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big) + \frac{1}{n}\mathbb{E}[\delta_t] + \frac{1}{n}\frac{2L\alpha_t}{m} \\
&\le \mathbb{E}[\delta_t] + \frac{2L\alpha_t}{nm} + \frac{2(n-1)\beta L\alpha_t}{nm}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big).
\end{aligned} \tag{B.11}$$

We then have

$$\begin{aligned}
\mathbb{E}[\delta_T] &\le \frac{2L}{nm}\sum_{t=1}^{T-1}\alpha_t + \frac{2(n-1)\beta L}{nm}\sum_{t=1}^{T-1}\alpha_t\Big[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big] \\
&\le \frac{2L}{n}\sum_{t=1}^{T-1}\frac{\alpha_t}{m} + 2\beta L\sum_{t=1}^{T-1}\frac{\alpha_t}{m}\Big[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big].
\end{aligned} \tag{B.12}$$

For every $\mathbf{z}$, the L-Lipschitz condition indicate that

$$\mathbb{E}|f(\mathbf{x}_T; \mathbf{z}) - f(\mathbf{x}'_T; \mathbf{z})| \le L\mathbb{E}[\delta_T] \le \frac{2L^2}{n}\sum_{t=1}^{T-1}\frac{\alpha_t}{m} + 2\beta L^2\sum_{t=1}^{T-1}\frac{\alpha_t}{m}\Big[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big],$$

which means that the uniform stability satisfies

$$\epsilon_{\text{stab}} \le \sum_{t=1}^{T-1}\Big[\frac{2L^2\alpha_t}{nm} + \frac{2\beta L^2\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big)\Big]. \tag{B.13}$$

■

B.5 Proof of Corollary 1 (generalization error for different learning rate in the convex case)

According to (B.13), for the constant learning rate $\alpha_t = \alpha$, we have

$$\begin{aligned}
\epsilon_{\text{stab}} &\le \frac{2L^2}{nm}\sum_{t=1}^{T-1}\alpha + 2\beta L^2\sum_{t=1}^{T-1}\frac{\alpha}{m}\Big[\sum_{s=1}^{t-\tau_t-1}\alpha\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha}{m}\Big] \\
&\le \frac{2L^2\alpha(T-1)}{nm} + \frac{2\beta L^2\alpha^2}{m}\sum_{t=1}^{T-1}\Big(\frac{1}{1-\lambda} + \frac{\tau_t}{m}\Big) \\
&\le \frac{2L^2\alpha(T-1)}{nm} + \frac{2\beta L^2\alpha^2(T-1)}{m}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big).
\end{aligned}$$

For the decreasing learning rate $\alpha_t = \frac{1}{t+1}$, it follows that

$$
\begin{aligned}
\epsilon_{\text{stab}} &\leq \frac{2L^2}{nm}\sum_{t=1}^{T-1}\alpha_t + \frac{2\beta L^2}{m}\sum_{t=1}^{T-1}\alpha_t\Big[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big] \\
&\leq \frac{2L^2}{nm}\sum_{t=1}^{T-1}\frac{1}{t+1} + \frac{2\beta L^2}{m}\sum_{t=1}^{T-1}\frac{1}{t+1}\Big[\frac{1}{\lambda^{\overline{\tau}}}\sum_{s=1}^{t-1}\frac{\lambda^{t-1-s}}{s+1} + \sum_{s=t-\tau_t}^{t-1}\frac{1}{m(s+1)}\Big] \\
&\overset{(a)}{\leq} \frac{2L^2}{nm}\ln T + \frac{2\beta L^2}{m}\sum_{t=1}^{T-1}\frac{1}{t+1}\Big[\frac{C_\lambda}{t\lambda^{\overline{\tau}}} + \frac{\tau_t}{m(t-\tau_t+1)}\Big] \\
&\leq \frac{2L^2}{nm}\ln T + \frac{2\beta L^2}{m}\Big[\frac{C_\lambda}{\lambda^{\overline{\tau}}}\sum_{t=1}^{T-1}(\frac{1}{t} - \frac{1}{t+1}) + \frac{1}{m}\sum_{t=1}^{T-1}(\frac{1}{t-\tau_t+1} - \frac{1}{t+1})\Big] \\
&\overset{(b)}{\leq} \frac{2L^2}{nm}\ln T + \frac{2\beta L^2}{m}\Big(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{\overline{\tau} + \ln(\overline{\tau}+1)}{m}\Big) \\
&\leq \frac{2L^2}{nm}\ln T + \frac{2\beta L^2}{m}\Big(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}\Big),
\end{aligned}
$$

where (a) uses the inequality (B.6) and

$$
\sum_{t=1}^{T-1}\frac{1}{t+1} \leq \sum_{t=1}^{T-1}\int_t^{t+1}\frac{1}{x}dx \leq \int_1^T \frac{1}{x}dx \leq \ln T, \tag{B.14}
$$

and (b) uses

$$
\begin{aligned}
\sum_{t=1}^{T-1}(\frac{1}{t-\tau_t+1} - \frac{1}{t+1}) &\leq \sum_{t=1}^{\overline{\tau}}(1 - \frac{1}{t+1}) + \sum_{t=\overline{\tau}+1}^{T-1}(\frac{1}{t-\overline{\tau}+1} - \frac{1}{t+1}) \\
&\leq \overline{\tau} + \sum_{t=1}^{\overline{\tau}}\frac{1}{t+1} - \sum_{t=T-\overline{\tau}}^{T-1}\frac{1}{t+1} \leq \overline{\tau} + \sum_{t=1}^{\overline{\tau}}\int_t^{t+1}\frac{1}{x}dx \leq \overline{\tau} + \ln(\overline{\tau}+1).
\end{aligned} \tag{B.15}
$$

■

B.6 Proof of Theorem 2 (generalization error for different learning rate in the strongly convex case)

$\mathbf{x}_T$ and $\mathbf{x}'_T$ denote the output model of running AD-SGD on $\mathcal{S}$ and $\mathcal{S}'$ for T iterations, respectively. With probability $1 - \frac{1}{n}$, the example selected in $\mathcal{S}$ and $\mathcal{S}'$ is the same at the t-th iteration, i.e., $j_t(i_t) \neq j_*$. Then we have

$$
\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| = \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\leq \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_t(i_t)}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_t(i_t)})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\overset{(a)}{\leq} (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}(i_t)\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_t - \mathbf{x}'_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}(i_t)\| \\
&\overset{(b)}{\leq} (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2\beta\alpha_t}{m}\frac{L}{m}\sum_{s=t-\tau_t}^{t-1}\alpha_s + \frac{2\beta\alpha_t}{m}L\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} \\
&\leq (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big),
\end{aligned} \tag{B.16}
$$

where (a) uses the strong convexity (B.5) and the β-smoothness assumption; (b) uses inequalities (B.7), (B.8). With probability $\frac{1}{n}$ the selected example is different, i.e., $j_t(i_t) = j_*$. With the bounded gradient assumption, we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| = \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_*}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}'_{j_*})\| \\
&\leq \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_*}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_*})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}_t; \mathbf{z}_{j_*}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_*}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_*}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_*})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}'_t; \mathbf{z}_{j_*}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_*}) + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_*}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_*})\| \\
&\quad + \|\frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_*}) - \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}'_{j_*})\| \\
&\leq (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}(i_t)\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_t - \mathbf{x}'_{t-\tau_t}\| + \frac{\beta\alpha_t}{m}\|\mathbf{x}'_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}(i_t)\| \\
&\quad + \frac{\alpha_t}{m}\|\nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_*}) - \nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}'_{j_*})\| \\
&\leq (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2\beta\alpha_t}{m}\frac{L}{m}\sum_{s=t-\tau_t}^{t-1}\alpha_s + \frac{2\beta\alpha_t}{m}L\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \frac{2L\alpha_t}{m} \\
&\leq (1 - \frac{\mu\alpha_t}{m})\|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2L\alpha_t}{m} + \frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big).
\end{aligned} \tag{B.17}$$

Combining the inequalities (B.16) and (B.17), we have

$$\begin{aligned}
\mathbb{E}[\delta_{t+1}] &\leq (1 - \frac{1}{n})(1 - \frac{\mu\alpha_t}{m})\mathbb{E}[\delta_t] + (1 - \frac{1}{n})\frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big) \\
&\quad + \frac{1}{n}(1 - \frac{\mu\alpha_t}{m})\mathbb{E}[\delta_t] + \frac{1}{n}\frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big) + \frac{1}{n}\frac{2L\alpha_t}{m} \\
&\leq (1 - \frac{\mu\alpha_t}{m})\mathbb{E}[\delta_t] + \frac{2L\alpha_t}{nm} + \frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big).
\end{aligned}$$

We then derive

$$\mathbb{E}[\delta_T] \leq \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1 - \frac{\mu\alpha_k}{m})\Big)\Big[\frac{2L\alpha_t}{nm} + \frac{2\beta L\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big)\Big]. \tag{B.18}$$

For every $\mathbf{z}$, the L-Lipschitz condition indicate that

$$\mathbb{E}|f(\mathbf{x}_T; \mathbf{z}) - f(\mathbf{x}'_T; \mathbf{z})| \leq L\mathbb{E}[\delta_T] \leq \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1 - \frac{\mu\alpha_k}{m})\Big)\cdot\Big[\frac{2L^2\alpha_t}{nm} + \frac{2\beta L^2\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big)\Big],$$

which means the uniform stability satisfies

$$\epsilon_{\text{stab}} \leq \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1 - \frac{\mu\alpha_k}{m})\Big)\Big[\frac{2L^2\alpha_t}{nm} + \frac{2\beta L^2\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big)\Big].$$

For the constant learning rate $\alpha_t = \alpha$, we have

$$\begin{aligned}
\epsilon_{\text{stab}} &\leq \sum_{t=1}^{T-1} \Big((1 - \frac{\mu\alpha}{m})^{T-1-t} \Big) \Big[\frac{2L^2\alpha}{nm} + \frac{2\beta L^2 \alpha^2}{m} \Big(\sum_{s=1}^{t-\tau_t-1} \lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1} \frac{1}{m} \Big) \Big] \\
&\leq \Big[\frac{2L^2\alpha}{nm} + \frac{2\beta L^2\alpha^2}{m} \Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m} \Big) \Big] \cdot \sum_{t=1}^{T-1} (1 - \frac{\mu\alpha}{m})^{T-1-t} \\
&\leq \Big[\frac{2L^2\alpha}{nm} + \frac{2\beta L^2\alpha^2}{m} \Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m} \Big) \Big] \cdot \frac{m}{\mu\alpha} \\
&\leq \frac{2L^2}{\mu n} + \frac{2\beta L^2 \alpha}{\mu} \Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m} \Big).
\end{aligned}$$

For the decreasing learning rate $\alpha_t = \frac{m}{\mu(t+1)}$, the stability turns to

$$\begin{aligned}
\epsilon_{\text{stab}} &\leq \sum_{t=1}^{T-1} \Big(\prod_{k=t+1}^{T-1} (1 - \frac{1}{k+1}) \Big) \Big[\frac{2L^2}{\mu n(t+1)} + \frac{2\beta L^2}{\mu(t+1)} \Big(\frac{m}{\mu} \sum_{s=1}^{t-\tau_t-1} \frac{\lambda^{t-\tau_t-1-s}}{s+1} + \frac{1}{\mu} \sum_{s=t-\tau_t}^{t-1} \frac{1}{s+1} \Big) \Big] \\
&\leq \sum_{t=1}^{T-1} \frac{t+1}{T} \Big[\frac{2L^2}{\mu n(t+1)} + \frac{2\beta L^2}{\mu(t+1)} \Big(\frac{m}{\mu\lambda^{\overline{\tau}}} \sum_{s=1}^{t-1} \frac{\lambda^{t-1-s}}{s+1} + \frac{1}{\mu} \sum_{s=t-\tau_t}^{t-1} \frac{1}{s+1} \Big) \Big] \\
&\overset{(a)}{\leq} \sum_{t=1}^{T-1} \frac{t+1}{T} \Big[\frac{2L^2}{\mu n(t+1)} + \frac{2\beta L^2}{\mu(t+1)} \Big(\frac{mC_\lambda}{\mu t \lambda^{\overline{\tau}}} + \frac{\tau_t}{\mu(t-\tau_t+1)} \Big) \Big] \\
&\leq \sum_{t=1}^{T-1} \Big[\frac{2L^2}{\mu n T} + \frac{2\beta L^2}{\mu T} \Big(\frac{mC_\lambda}{\mu t \lambda^{\overline{\tau}}} + \frac{\overline{\tau}}{\mu(t-\tau_t+1)} \Big) \Big] \\
&\overset{(b)}{\leq} \frac{2L^2}{\mu n} + \frac{2m\beta L^2 C_\lambda}{\mu^2 \lambda^{\overline{\tau}}} \frac{\ln T + 1}{T} + \frac{2\beta L^2}{\mu^2} \frac{\overline{\tau}^2 + \overline{\tau} \ln T}{T} \\
&\leq \frac{2L^2}{\mu n} + \frac{2\beta L^2 (mC_\lambda + \overline{\tau}^2 \lambda^{\overline{\tau}})}{\mu^2 \lambda^{\overline{\tau}}} \frac{\ln T + 1}{T},
\end{aligned}$$

where (a) uses the inequality (B.6), and (b) uses the following inequalities

$$\sum_{t=1}^{T-1} \frac{1}{t} = 1 + \sum_{t=1}^{T-2} \frac{1}{t+1} \leq 1 + \sum_{t=1}^{T-2} \int_t^{t+1} \frac{1}{x} dx \leq 1 + \int_1^{T-1} \frac{1}{x} dx \leq \ln T + 1; \tag{B.19}$$

$$\sum_{t=1}^{T-1} \frac{1}{t-\tau_t+1} \leq \sum_{t=1}^{\overline{\tau}} \frac{1}{t-\tau_t+1} + \sum_{t=\overline{\tau}+1}^{T-1} \frac{1}{t-\overline{\tau}+1} \leq \overline{\tau} + \sum_{t=1}^{T-\overline{\tau}-1} \frac{1}{t+1} \leq \overline{\tau} + \ln(T-\overline{\tau}) \leq \overline{\tau} + \ln T. \tag{B.20}$$

■

B.7 Proof of Theorem 3 (generalization error in the non-convex case)

$\mathbf{x}_T$ and $\mathbf{x}'_T$ denote the output model of running AD-SGD on $\mathcal{S}$ and $\mathcal{S}'$ for T iterations, respectively. With probability $1 - \frac{1}{n}$, the example selected in $\mathcal{S}$ and $\mathcal{S}'$ is the same at the t-th iteration, i.e., $j_t(i_t) \neq j_*$. Then we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| = \|\mathbf{x}_t - \frac{\alpha_t}{m} \nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) - \mathbf{x}'_t + \frac{\alpha_t}{m} \nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\leq \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\alpha_t}{m} \|\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \\
&\leq \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\alpha_t}{m} \Big[\|\nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)})\| \\
&\quad + \|\nabla f(\mathbf{x}_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}; \mathbf{z}_{j_t(i_t)})\| + \|\nabla f(\mathbf{x}'_{t-\tau_t}; \mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}'_{t-\tau_t}(i_t); \mathbf{z}_{j_t(i_t)})\| \Big] \\
&\leq \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\beta\alpha_t}{m} \|\mathbf{x}_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}\| + \frac{\beta\alpha_t}{m} \|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}(i_t)\| + \frac{\beta\alpha_t}{m} \|\mathbf{x}'_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}(i_t)\| \\
&\leq \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{\beta\alpha_t}{m} \|\mathbf{x}_{t-\tau_t} - \mathbf{x}'_{t-\tau_t}\| + \frac{2\beta L \alpha_t}{m} \sum_{s=1}^{t-\tau_t-1} \alpha_s \lambda^{t-\tau_t-1-s}.
\end{aligned} \tag{B.21}$$

With probability $\frac{1}{n}$, $j_t = j_*$, we can get

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| &= \|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_*}) - \mathbf{x}'_t + \frac{\alpha_t}{m}\nabla f(\mathbf{x}'_{t-\tau_t}(i_t);\mathbf{z}'_{j_*})\| \\ &\le \|\mathbf{x}_t - \mathbf{x}'_t\| + \frac{2L\alpha_t}{m}. \end{aligned} \tag{B.22}$$

Combining inequalities (B.21) and (B.22), we have

$$\begin{aligned} \mathbb{E}[\delta_{t+1}] &\le (1-\frac{1}{n})\mathbb{E}[\delta_t] + (1-\frac{1}{n})\frac{\beta\alpha_t}{m}\mathbb{E}[\delta_{t-\tau_t}] + (1-\frac{1}{n})\frac{2\beta L\alpha_t}{m}\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \frac{1}{n}\mathbb{E}[\delta_t] + \frac{1}{n}\frac{2L\alpha_t}{m} \\ &\le \mathbb{E}[\delta_t] + \frac{(n-1)\beta\alpha_t}{nm}\mathbb{E}[\delta_{t-\tau_t}] + \frac{2L\alpha_t}{nm} + \frac{2(n-1)\beta L\alpha_t}{nm}\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} \\ &\le \mathbb{E}[\delta_t] + \frac{\beta\alpha_t}{m}\max_{t-\tau_t\le k\le t}\mathbb{E}[\delta_k] + \frac{2L\alpha_t}{m}\Big(\frac{1}{n} + \sum_{s=1}^{t-\tau_t-1}\beta\alpha_s.\lambda^{t-\tau_t-1-s}\Big). \end{aligned} \tag{B.23}$$

Following [Proposition 2, (Regatti et al. 2019)] and we define $\prod_{k=t'+1}^{t'}(1+\frac{\beta\alpha_k}{m}) = 1$. Then we have

$$\mathbb{E}[\delta_T] \le \sum_{t=1}^{T-1}\left(\prod_{k=t+1}^{T-1}(1+\frac{\beta\alpha_k}{m})\right)\frac{2L\alpha_t}{m}\left(\frac{1}{n} + \sum_{s=1}^{t-\tau_t-1}\beta\alpha_s\lambda^{t-\tau_t-1-s}\right). \tag{B.24}$$

For every $\mathbf{z}$, the L-Lipschitz condition indicate that

$$\mathbb{E}|f(\mathbf{x}_T;\mathbf{z}) - f(\mathbf{x}'_T;\mathbf{z})| \le L\mathbb{E}[\delta_T] \le \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1+\frac{\beta\alpha_k}{m})\Big)\frac{2L^2\alpha_t}{m}\left(\frac{1}{n} + \sum_{s=1}^{t-\tau_t-1}\beta\alpha_s\lambda^{t-\tau_t-1-s}\right).$$

which means the uniform stability in the non-convex case satisfies

$$\epsilon_{\text{stab}} \le \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1+\frac{\beta\alpha_k}{m})\Big)\frac{2L^2\alpha_t}{m}\left(\frac{1}{n} + \sum_{s=1}^{t-\tau_t-1}\beta\alpha_s\lambda^{t-\tau_t-1-s}\right). \tag{B.25}$$

■

B.8 Proof of Corollary 2 (generalization error for different learning rate in the non-convex case)

According to (B.25), for the constant learning rate $\alpha_t = \alpha$, we have

$$\begin{aligned} \epsilon_{\text{stab}} &\le \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1+\frac{\beta\alpha}{m})\Big)\frac{2L^2\alpha}{m}\left(\frac{1}{n} + \sum_{s=1}^{t-\tau_t-1}\beta\alpha\lambda^{t-\tau_t-1-s}\right) \\ &\le \Big(\frac{2L^2\alpha}{nm} + \frac{2\beta L^2\alpha^2}{m(1-\lambda)}\Big)\sum_{t=1}^{T-1}(1+\frac{\beta\alpha}{m})^{T-1-t} \\ &\le \Big(\frac{2L^2\alpha}{nm} + \frac{2\beta L^2\alpha^2}{m(1-\lambda)}\Big)\frac{m}{\beta\alpha}\left[(1+\frac{\beta\alpha}{m})^{T-1} - 1\right] \\ &\le \frac{2L^2(1+\beta n\alpha-\lambda)}{\beta n(1-\lambda)}(1+\frac{\beta\alpha}{m})^{T-1}. \end{aligned} \tag{B.26}$$

For the decreasing learning rate $\alpha_t = \frac{mc}{t+1}$, it follows that

$$\begin{aligned}
\epsilon_{\text{stab}} &\le \sum_{t=1}^{T-1} \left\{ \prod_{k=t+1}^{T-1} (1 + \frac{\beta c}{k+1}) \right\} \left(\frac{2L^2 c}{n(t+1)} + \frac{2\beta L^2 mc^2}{t+1} \sum_{s=1}^{t-\tau_t-1} \frac{\lambda^{t-\tau_t-1-s}}{s+1} \right) \\
&\overset{(a)}{\le} \sum_{t=1}^{T-1} \left\{ \prod_{k=t+1}^{T-1} \exp\left(\frac{\beta c}{k+1} \right) \right\} \left(\frac{2L^2 c}{n(t+1)} + \frac{2\beta L^2 mc^2}{t+1} \sum_{s=1}^{t-\tau_t-1} \lambda^{t-\tau_t-1-s} \right) \\
&\le \sum_{t=1}^{T-1} \exp\left(\beta c \sum_{k=t+1}^{T-1} \frac{1}{k+1} \right) \left[\frac{2L^2 c}{n(t+1)} + \frac{2\beta L^2 mc^2}{(1-\lambda)(t+1)} \right] \\
&\overset{(b)}{\le} \sum_{t=1}^{T-1} \exp\left(\beta c \ln(\frac{T}{t+1}) \right) \left[\frac{2L^2 c}{n(t+1)} + \frac{2\beta L^2 mc^2}{(1-\lambda)(t+1)} \right] \\
&\le \left[\frac{2L^2 c}{n} + \frac{2\beta L^2 mc^2}{1-\lambda} \right] T^{\beta c} \sum_{t=1}^{T-1} (t+1)^{-\beta c - 1} \\
&\overset{(c)}{\le} \left[\frac{2L^2 c}{n} + \frac{2\beta L^2 mc^2}{1-\lambda} \right] T^{\beta c} \frac{1}{\beta c} (1 - \frac{1}{T^{\beta c}}) \\
&\le \frac{2L^2 (1 + \beta nmc - \lambda)}{\beta n (1-\lambda)} T^{\beta c},
\end{aligned} \tag{B.27}$$

where (a) uses $1 + x \le e^x$. (b) and (c) respectively use the following inequalities

$$\sum_{k=t+1}^{T-1} \frac{1}{k+1} \le \sum_{k=t+1}^{T-1} \int_k^{k+1} \frac{1}{x} dx \le \int_{t+1}^{T} \frac{1}{x} dx = \ln(\frac{T}{t+1});$$

$$\sum_{t=1}^{T-1} (t+1)^{-\beta c - 1} \le \sum_{t=1}^{T-1} \int_t^{t+1} x^{-\beta c - 1} dx \le \int_1^T x^{-\beta c - 1} dx = \frac{1}{\beta c}(1 - T^{-\beta c}).$$

With $c = 1/\beta$, we have

$$\epsilon_{\text{stab}} \le \frac{2L^2 (1 + nm - \lambda)}{\beta n (1-\lambda)} T.$$

■

B.9 Proof of Theorem 4 (generalization error for decreasing learning rate in the non-convex case)

Following [Lemma 3.11, (Hardt, Recht, and Singer 2016)], let $\delta_{t_0 = 0}$ and we have

$$\epsilon_{\text{stab}} \le \frac{t_0}{n} + L\mathbb{E}[\delta_T | \delta_{t_0=0}].$$

Similar to the derivation in (B.27), we have

$$\mathbb{E}[\delta_T | \delta_{t_0=0}] \le \frac{2L(1 + \beta nmc - \lambda)}{\beta n (1-\lambda)} \left(\frac{T}{t_0} \right)^{\beta c}.$$

Then we get

$$\epsilon_{\text{stab}} \le \frac{t_0}{n} + \frac{2L^2(1 + \beta nmc - \lambda)}{\beta n (1-\lambda)} \left(\frac{T}{t_0} \right)^{\beta c}.$$

Assume c is small enough, minimizing this bound with respect to t_0, i.e., let

$$t_0 = \left[2L^2 c (1 + \frac{\beta nmc}{1-\lambda}) \right]^{\frac{1}{\beta c + 1}} T^{\frac{\beta c}{\beta c + 1}},$$

then the uniform stability satisfies

$$\epsilon_{\text{stab}} \le \frac{1 + 1/\beta c}{n} \left[2L^2 c (1 + \frac{\beta nmc}{1-\lambda}) \right]^{\frac{1}{\beta c + 1}} T^{\frac{\beta c}{\beta c + 1}}.$$

■

B.10 Proof of Theorem 5 (optimization error and excess generalization error in the strongly convex case)

Recall that $\mathbf{x}_t$ is the output model after minimizing the empirical risk $F_{\mathcal{S}}$ for t AD-SGD iterations, and $\mathbf{x}^*_{\mathcal{S}}$ denotes the minimizer of $F_{\mathcal{S}}$. From the iterative relation (B.2), we can derive

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}^*_{\mathcal{S}}\|^2 = \mathbb{E}\|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}) - \mathbf{x}^*_{\mathcal{S}}\|^2 \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{\alpha_t^2}{m^2}\mathbb{E}\|\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\|^2 \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{2\alpha_t}{m}\mathbb{E}\langle \nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{L^2\alpha_t^2}{m^2} \\
&\overset{(a)}{\le} \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{4r\alpha_t}{m}\mathbb{E}\|\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\| + \frac{L^2\alpha_t^2}{m^2} \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{L^2\alpha_t^2}{m^2} \\
&\quad + \frac{4r\alpha_t}{m}\left[\mathbb{E}\|\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t};\mathbf{z}_{j_t(i_t)})\| + \mathbb{E}\|\nabla f(\mathbf{x}_{t-\tau_t};\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\|\right] \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{4\beta r\alpha_t}{m}\left[\|\mathbf{x}_t-\mathbf{x}_{t-\tau_t}\| + \|\mathbf{x}_{t-\tau_t}-\mathbf{x}_{t-\tau_t}(i_t)\|\right] + \frac{L^2\alpha_t^2}{m^2} \\
&\overset{(b)}{\le} \mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{4\beta rL\alpha_t}{m}\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{L^2\alpha_t^2}{m^2} \\
&\overset{(c)}{\le} (1-\frac{2\mu\alpha_t}{m})\mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2 + \frac{4\beta rL\alpha_t}{m}\left(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right) + \frac{L^2\alpha_t^2}{m^2},
\end{aligned} \tag{B.28}$$

where (a) uses the inequality $\langle \mathbf{a},\mathbf{b}\rangle \le \|\mathbf{a}\|\|\mathbf{b}\|$ and Assumption 4 (r is the radius of the close ball). (b) uses inequalities (B.7) and (B.8). (c) employs the following μ-strongly convexity

$$\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\rangle \ge \mu\|\mathbf{x}_t-\mathbf{x}^*_{\mathcal{S}}\|^2.$$

We then have

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_T-\mathbf{x}^*_{\mathcal{S}}\|^2 \le &\sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1-\frac{2\mu\alpha_k}{m})\Big)\Big[\frac{L^2\alpha_t^2}{m^2} + \frac{4\beta rL\alpha_t}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\Big)\Big] \\
&+ \prod_{t=1}^{T-1}(1-\frac{2\mu\alpha_t}{m})\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2.
\end{aligned}$$

For the constant learning rate $\alpha_t = \alpha$

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_T-\mathbf{x}^*_{\mathcal{S}}\|^2 &\le \sum_{t=1}^{T-1}\Big((1-\frac{2\mu\alpha}{m})^{T-1-t}\Big)\Big[\frac{L^2\alpha^2}{m^2} + \frac{4\beta rL\alpha^2}{m}\Big(\sum_{s=1}^{t-\tau_t-1}\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{1}{m}\Big)\Big] + (1-\frac{2\mu\alpha}{m})^{T-1}\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2 \\
&\le \Big[\frac{L^2\alpha^2}{m^2} + \frac{4\beta rL\alpha^2}{m}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big)\Big]\cdot\sum_{t=1}^{T-1}(1-\frac{2\mu\alpha}{m})^{T-1-t} + (1-\frac{2\mu\alpha}{m})^{T-1}\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2 \\
&\le \Big[\frac{L^2\alpha^2}{m^2} + \frac{4\beta rL\alpha^2}{m}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big)\Big]\cdot\frac{m}{2\mu\alpha} + (1-\frac{2\mu\alpha}{m})^{T-1}\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2 \\
&\le \frac{L^2\alpha}{2\mu m} + \frac{2\beta rL\alpha}{\mu}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big) + (1-\frac{2\mu\alpha}{m})^{T-1}\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2.
\end{aligned}$$

With β-smooth property, the optimization error satisfies

$$\begin{aligned}
\epsilon_{\text{opt}} = \mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_T) - F_{\mathcal{S}}(\mathbf{x}^*_{\mathcal{S}})] &\le \mathbb{E}\langle \nabla F_{\mathcal{S}}(\mathbf{x}^*_{\mathcal{S}}), \mathbf{x}_T-\mathbf{x}^*_{\mathcal{S}}\rangle + \frac{\beta}{2}\mathbb{E}\|\mathbf{x}_T-\mathbf{x}^*_{\mathcal{S}}\|^2 \le \frac{\beta}{2}\mathbb{E}\|\mathbf{x}_T-\mathbf{x}^*_{\mathcal{S}}\|^2 \\
&\le \frac{\beta L^2\alpha}{4\mu m} + \frac{\beta^2 rL\alpha}{\mu}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big) + (1-\frac{2\mu\alpha}{m})^{T-1}\frac{\beta\|\mathbf{x}_1-\mathbf{x}^*_{\mathcal{S}}\|^2}{2}.
\end{aligned}$$

Following the decomposition (1), the excess generalization error satisfies

$$\begin{aligned}\epsilon_{\text{exc}} &\le \epsilon_{\text{stab}} + \epsilon_{\text{opt}} \\ &\le \frac{2L^2}{\mu n} + \frac{2\beta L^2\alpha}{\mu}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big) + \frac{\beta L^2\alpha}{4\mu m} + \frac{\beta^2 r L\alpha}{\mu}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big) + (1 - \frac{2\mu\alpha}{m})^{T-1}\frac{\beta\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2} \\ &\le \frac{L^2(8m + \beta n\alpha)}{4\mu n m} + \frac{\beta L\alpha(2L + \beta r)}{\mu}\Big(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\Big) + (1 - \frac{2\mu\alpha}{m})^{T-1}\frac{\beta\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2}.\end{aligned}$$

For the decreasing learning rate $\alpha_t = \frac{m}{2\mu(t+1)}$, we have

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}_T - \mathbf{x}_{\mathcal{S}}^*\|^2 \\ &\le \sum_{t=1}^{T-1}\Big(\prod_{k=t+1}^{T-1}(1 - \frac{1}{k+1})\Big)\Big[\frac{L^2}{4\mu^2(t+1)^2} + \frac{2\beta r L}{\mu(t+1)}\Big(\frac{m}{2\mu}\sum_{s=1}^{t-\tau_t-1}\frac{\lambda^{t-\tau_t-1-s}}{s+1} + \frac{1}{2\mu}\sum_{s=t-\tau_t}^{t-1}\frac{1}{s+1}\Big)\Big] \\ &\quad + \prod_{t=1}^{T-1}(1 - \frac{1}{t+1})\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2 \\ &\le \sum_{t=1}^{T-1}\frac{t+1}{T}\Big[\frac{L^2}{4\mu^2(t+1)^2} + \frac{2\beta r L}{\mu(t+1)}\Big(\frac{m}{2\mu\lambda^{\overline{\tau}}}\sum_{s=1}^{t-1}\frac{\lambda^{t-1-s}}{s+1} + \frac{1}{2\mu}\sum_{s=t-\tau_t}^{t-1}\frac{1}{s+1}\Big)\Big] + \frac{\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{T} \\ &\overset{(a)}{\le} \sum_{t=1}^{T-1}\frac{t+1}{T}\Big[\frac{L^2}{4\mu^2(t+1)^2} + \frac{2\beta r L}{\mu(t+1)}\Big(\frac{mC_\lambda}{2\mu t\lambda^{\overline{\tau}}} + \frac{\tau_t}{2\mu(t-\tau_t+1)}\Big)\Big] + \frac{\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{T} \\ &\le \sum_{t=1}^{T-1}\Big[\frac{L^2}{4\mu^2 T(t+1)} + \frac{2\beta r L}{\mu T}\Big(\frac{mC_\lambda}{2\mu t\lambda^{\overline{\tau}}} + \frac{\overline{\tau}}{2\mu(t-\tau_t+1)}\Big)\Big] + \frac{\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{T} \\ &\overset{(b)}{\le} \frac{L^2\ln T}{4\mu^2 T} + \frac{\beta r L m C_\lambda}{\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\beta r L}{\mu^2}\frac{\overline{\tau}^2 + \overline{\tau}\ln T}{T} + \frac{\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{T} \\ &\le \frac{L^2\ln T}{4\mu^2 T} + \frac{\beta r L(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{T},\end{aligned}$$

where (a) uses inequality (B.6), and (b) uses (B.14), (B.19) and (B.20). With β-smooth property, the optimization error satisfies

$$\begin{aligned}\epsilon_{\text{opt}} &= \mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_T) - F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)] \le \mathbb{E}\langle\nabla F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*), \mathbf{x}_T - \mathbf{x}_{\mathcal{S}}^*\rangle + \frac{\beta}{2}\mathbb{E}\|\mathbf{x}_T - \mathbf{x}_{\mathcal{S}}^*\|^2 \le \frac{\beta}{2}\mathbb{E}\|\mathbf{x}_T - \mathbf{x}_{\mathcal{S}}^*\|^2 \\ &\le \frac{\beta L^2\ln T}{8\mu^2 T} + \frac{\beta^2 r L(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{2\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\beta\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2T} \\ &\le \frac{\beta L^2\ln T}{8\mu^2 T} + \frac{\beta^2 r L(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{2\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{2\beta r^2}{T}.\end{aligned}$$

Following the decomposition (1), the excess generalization risk satisfies

$$\begin{aligned}\epsilon_{\text{exc}} &\le \epsilon_{\text{stab}} + \epsilon_{\text{opt}} \\ &\le \frac{2L^2}{\mu n} + \frac{2\beta L^2(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\beta L^2\ln T}{8\mu^2 T} + \frac{\beta^2 r L(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{2\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\beta\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2T} \\ &\le \frac{2L^2}{\mu n} + \frac{\beta L(4L + \beta r)(mC_\lambda + \overline{\tau}^2\lambda^{\overline{\tau}})}{2\mu^2\lambda^{\overline{\tau}}}\frac{\ln T + 1}{T} + \frac{\beta L^2\ln T}{8\mu^2 T} + \frac{\beta\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2T}.\end{aligned}$$

■

B.11 Proof of Theorem 6 and 7 (optimization error and excess generalization error in the convex case)

Similar to the analysis in (B.28), we have the following relationship

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_{\mathcal{S}}^*\|^2 = \mathbb{E}\|\mathbf{x}_t - \frac{\alpha_t}{m}\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}) - \mathbf{x}_{\mathcal{S}}^*\|^2 \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{\alpha_t^2}{m^2}\mathbb{E}\|\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\|^2 \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{2\alpha_t}{m}\mathbb{E}\langle \nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{L^2\alpha_t^2}{m^2} \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{4r\alpha_t}{m}\mathbb{E}\|\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\| + \frac{L^2\alpha_t^2}{m^2} \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{L^2\alpha_t^2}{m^2} \\
&\quad + \frac{4r\alpha_t}{m}\left[\mathbb{E}\|\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t};\mathbf{z}_{j_t(i_t)})\| + \mathbb{E}\|\nabla f(\mathbf{x}_{t-\tau_t};\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\|\right] \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{4\beta r\alpha_t}{m}\left[\|\mathbf{x}_t-\mathbf{x}_{t-\tau_t}\| + \|\mathbf{x}_{t-\tau_t}-\mathbf{x}_{t-\tau_t}(i_t)\|\right] + \frac{L^2\alpha_t^2}{m^2} \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 + \frac{2\alpha_t}{m}\mathbb{E}\langle -\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle + \frac{4\beta rL\alpha_t}{m}\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{L^2\alpha_t^2}{m^2} \\
&\le \mathbb{E}\|\mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\|^2 - \frac{2\alpha_t}{m}\mathbb{E}\left[F_{\mathcal{S}}(\mathbf{x}_t)-F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)\right] + \frac{4\beta rL\alpha_t}{m}\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{L^2\alpha_t^2}{m^2}.
\end{aligned}
$$

The last inequality uses the unbiased property of the stochastic gradient and the convexity of the loss function, i.e.,

$$\langle \nabla F_{\mathcal{S}}(\mathbf{x}_t), \mathbf{x}_t-\mathbf{x}_{\mathcal{S}}^*\rangle \ge F_{\mathcal{S}}(\mathbf{x}_t)-F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*).$$

Then we have

$$\sum_{t=1}^{T}\alpha_t\mathbb{E}\left[F_{\mathcal{S}}(\mathbf{x}_t)-F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)\right] \le \frac{m}{2}\|\mathbf{x}_1-\mathbf{x}_{\mathcal{S}}^*\|^2 + 2\beta rL\sum_{t=1}^{T}\alpha_t\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{L^2}{2m}\sum_{t=1}^{T}\alpha_t^2.$$

Devote the average model as

$$\overline{\mathbf{x}}_T = \frac{\sum_{t=1}^{T}\alpha_t\mathbf{x}_t}{\sum_{t=1}^{T}\alpha_t}.$$

It follows that

$$
\begin{aligned}
\epsilon_{\text{opt}} = \mathbb{E}[F_{\mathcal{S}}(\overline{\mathbf{x}}_T)-F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)] &\le \frac{\sum_{t=1}^{T}\alpha_t\mathbb{E}\left[F_{\mathcal{S}}(\mathbf{x}_t)-F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)\right]}{\sum_{t=1}^{T}\alpha_t} \\
&\le \frac{m\|\mathbf{x}_1-\mathbf{x}_{\mathcal{S}}^*\|^2}{2\sum_{t=1}^{T}\alpha_t} + \frac{2\beta rL}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{L^2\sum_{t=1}^{T}\alpha_t^2}{2m\sum_{t=1}^{T}\alpha_t}.
\end{aligned}
$$

For the constant learning rate $\alpha_t = \alpha$

$$
\begin{aligned}
\epsilon_{\text{opt}} &\le \frac{m\|\mathbf{x}_1-\mathbf{x}_{\mathcal{S}}^*\|^2}{2\sum_{t=1}^{T}\alpha} + \frac{2\beta rL}{\sum_{t=1}^{T}\alpha}\sum_{t=1}^{T}\alpha\left[\sum_{s=1}^{t-\tau_t-1}\alpha\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha}{m}\right] + \frac{L^2\sum_{t=1}^{T}\alpha^2}{2m\sum_{t=1}^{T}\alpha} \\
&\le \frac{m\|\mathbf{x}_1-\mathbf{x}_{\mathcal{S}}^*\|^2}{2T\alpha} + \frac{2\beta rL}{T\alpha}\sum_{t=1}^{T}\alpha^2\left[\frac{1}{1-\lambda}+\frac{\overline{\tau}}{m}\right] + \frac{L^2T\alpha^2}{2mT\alpha} \\
&\le \frac{m\|\mathbf{x}_1-\mathbf{x}_{\mathcal{S}}^*\|^2}{2T\alpha} + 2\beta rL\alpha\left[\frac{1}{1-\lambda}+\frac{\overline{\tau}}{m}\right] + \frac{L^2\alpha}{2m}.
\end{aligned}
$$

For the decreasing learning rate $\alpha_t = \frac{1}{t+1}$, we have

$$\begin{aligned}
\epsilon_{\text{opt}} &\leq \frac{m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2\sum_{t=1}^{T}\frac{1}{t+1}} + \frac{2\beta r L}{\sum_{t=1}^{T}\frac{1}{t+1}} \sum_{t=1}^{T} \frac{1}{t+1}\left[\sum_{s=1}^{t-\tau_t-1} \frac{\lambda^{t-\tau_t-1-s}}{s+1} + \sum_{s=t-\tau_t}^{t-1} \frac{1}{m(s+1)}\right] + \frac{L^2\sum_{t=1}^{T}\frac{1}{(t+1)^2}}{2m\sum_{t=1}^{T}\frac{1}{t+1}} \\
&\overset{(a)}{\leq} \frac{m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{\ln(T+1)} + \frac{4\beta r L}{\ln(T+1)} \sum_{t=1}^{T} \frac{1}{t+1}\left[\frac{C_\lambda}{t\lambda^{\overline{\tau}}} + \frac{\tau_t}{m(t-\tau_t+1)}\right] + \frac{L^2}{m\ln(T+1)} \\
&\overset{(b)}{\leq} \frac{m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{\ln(T+1)} + \frac{4\beta r L}{\ln(T+1)}\left[\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{\overline{\tau} + \ln(\overline{\tau}+1)}{m}\right] + \frac{L^2}{m\ln(T+1)} \\
&\leq \left[m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2 + 4\beta r L(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}) + \frac{L^2}{m}\right]\frac{1}{\ln(T+1)} \\
&\leq \left[4mr^2 + 4\beta r L(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}) + \frac{L^2}{m}\right]\frac{1}{\ln(T+1)},
\end{aligned}$$

where (a) uses (B.6) and the following inequalities

$$\sum_{t=1}^{T} \frac{1}{t+1} \geq \frac{1}{2}\sum_{t=1}^{T}\frac{1}{t} \geq \frac{1}{2}\sum_{t=1}^{T}\int_t^{t+1}\frac{1}{x}dx \geq \frac{1}{2}\ln(T+1); \tag{B.29}$$

$$\sum_{t=1}^{T} \frac{1}{(t+1)^2} \leq \sum_{t=1}^{T}\int_t^{t+1}\frac{1}{x^2}dx \leq \int_1^{T+1}\frac{1}{x^2}dx \leq 1 - \frac{1}{T+1} \leq 1. \tag{B.30}$$

(b) uses inequality (B.15). In the following, we fist derive the uniform stability bound for the average model $\overline{\mathbf{x}}_T$. From the analysis in (B.12), we have

$$\mathbb{E}\|\mathbf{x}_t - \mathbf{x}_t'\| \leq \frac{2L}{n}\sum_{k=1}^{t-1}\frac{\alpha_k}{m} + 2\beta L\sum_{k=1}^{t-1}\frac{\alpha_k}{m}\left[\sum_{s=1}^{k-\tau_k-1}\alpha_s\lambda^{k-\tau_k-1-s} + \sum_{s=k-\tau_k}^{k-1}\frac{\alpha_s}{m}\right].$$

Then we can derive

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{x}}_T - \overline{\mathbf{x}}_T'\| &= \mathbb{E}\left\|\frac{\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{x}_t')}{\sum_{t=1}^{T}\alpha_t}\right\| \leq \frac{\sum_{t=1}^{T}\alpha_t\mathbb{E}\|\mathbf{x}_t - \mathbf{x}_t'\|}{\sum_{t=1}^{T}\alpha_t} \\
&\leq \frac{\frac{2L}{nm}\sum_{t=1}^{T}\alpha_t\sum_{k=1}^{t-1}\alpha_k + \frac{2\beta L}{m}\sum_{t=1}^{T}\alpha_t\sum_{k=1}^{t-1}\alpha_k\left[\sum_{s=1}^{k-\tau_k-1}\alpha_s\lambda^{k-\tau_k-1-s} + \sum_{s=k-\tau_k}^{k-1}\frac{\alpha_s}{m}\right]}{\sum_{t=1}^{T}\alpha_t}.
\end{aligned}$$

For the constant learning rate $\alpha_t = \alpha$

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{x}}_T - \overline{\mathbf{x}}_T'\| &\leq \frac{\frac{2L\alpha^2}{nm}\sum_{t=1}^{T}(t-1) + \frac{2\beta L\alpha^3}{m}\sum_{t=1}^{T}(t-1)\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right]}{T\alpha} \\
&\leq \frac{\frac{2L\alpha^2}{nm}\frac{T(T-1)}{2} + \frac{2\beta L\alpha^3}{m}\frac{T(T-1)}{2}\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right]}{T\alpha} \\
&\leq \frac{L\alpha(T-1)}{nm} + \frac{\beta L\alpha^2(T-1)}{m}\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right].
\end{aligned}$$

Combine with the L-Lipschitz condition, the uniform stability bound of $\overline{\mathbf{x}}_T$ satisfies

$$\epsilon_{\text{ave-stab}} \leq \frac{L^2\alpha(T-1)}{nm} + \frac{\beta L^2\alpha^2(T-1)}{m}\left(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right).$$

Then the excess generalization risk follows

$$\begin{aligned}
\epsilon_{\text{exc}} &\leq \epsilon_{\text{ave-stab}} + \epsilon_{\text{opt}} \\
&\leq \frac{L^2\alpha(T-1)}{nm} + \frac{\beta L^2\alpha^2(T-1)}{m}\left(\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right) + \frac{m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2}{2T\alpha} + 2\beta r L\alpha\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right] + \frac{L^2\alpha}{2m}.
\end{aligned}$$

For the decreasing learning rate $\alpha_t = \frac{1}{t+1}$

$$
\begin{aligned}
\mathbb{E}\|\overline{\mathbf{x}}_T - \overline{\mathbf{x}}_T'\| &\le \frac{\frac{2L}{nm}\sum_{t=1}^T \frac{1}{t+1}\sum_{k=1}^{t-1}\frac{1}{k+1} + \frac{2\beta L}{m}\sum_{t=1}^T\frac{1}{t+1}\sum_{k=1}^{t-1}\frac{1}{k+1}\left[\sum_{s=1}^{k-\tau_k-1}\frac{1}{s+1}\lambda^{k-\tau_k-1-s} + \sum_{s=k-\tau_k}^{k-1}\frac{1}{m(s+1)}\right]}{\sum_{t=1}^T\frac{1}{t+1}} \\
&\overset{(a)}{\le} \frac{\frac{4L}{nm}\sum_{t=1}^T\frac{\ln t}{t+1} + \frac{4\beta L}{m}\sum_{t=1}^T\frac{1}{t+1}\sum_{k=1}^{t-1}\frac{1}{k+1}\left[\frac{C_\lambda}{\lambda^{\overline{\tau}}}\frac{1}{k} + \frac{\tau_k}{m(k-\tau_k-1)}\right]}{\ln(T+1)} \\
&\overset{(b)}{\le} \frac{\frac{2L}{nm}\ln^2(T+1) + \frac{4\beta L}{m}\sum_{t=1}^T\frac{1}{t+1}\left[\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{\overline{\tau}+\ln(\overline{\tau}+1)}{m}\right]}{\ln(T+1)} \\
&\le \frac{2L}{nm}\ln(T+1) + \frac{4\beta L}{m}\left(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}\right),
\end{aligned}
$$

where (a) uses inequalities (B.6), (B.14) and (B.29). (b) uses inequality (B.20) and

$$
\sum_{t=1}^T\frac{\ln t}{t+1} \le \sum_{t=1}^T\int_t^{t+1}\frac{\ln x}{x}dx \le \int_1^{T+1}\frac{\ln x}{x}dx = \frac{\ln^2(T+1)}{2}.
$$

Then the uniform stability bound of $\overline{\mathbf{x}}_T$ satisfies

$$
\epsilon_{\text{ave-stab}} \le \frac{2L^2}{nm}\ln(T+1) + \frac{4\beta L^2}{m}\left(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}\right).
$$

The excess generalization risk in the decreasing learning rate follows

$$
\begin{aligned}
&\epsilon_{\text{exc}} \le \epsilon_{\text{ave-stab}} + \epsilon_{\text{opt}} \\
&\quad\le \frac{2L^2}{nm}\ln(T+1) + \frac{4\beta L^2}{m}\left(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}\right) + \left[m\|\mathbf{x}_1 - \mathbf{x}_{\mathcal{S}}^*\|^2 + 4\beta r L(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}) + \frac{L^2}{m}\right]\frac{1}{\ln(T+1)}.
\end{aligned}
$$

■

B.12 Proof of Theorem 8 and 9 (optimization error and excess generalization error in the non-convex case)

With the β-smooth property, we have

$$
\begin{aligned}
&\mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_{t+1}) - F_{\mathcal{S}}(\mathbf{x}_t)] \le \mathbb{E}\langle\nabla F_{\mathcal{S}}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{\beta}{2}\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\le \mathbb{E}\langle\nabla F_{\mathcal{S}}(\mathbf{x}_t), -\frac{\alpha_t}{m}\nabla F_{\mathcal{S}}(\mathbf{x}_t)\rangle + \mathbb{E}\left\langle\nabla f(\mathbf{x}_t(i_t);\mathbf{z}_{j_t}), \frac{\alpha_t}{m}\left(\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\right)\right\rangle \\
&\quad + \frac{\beta\alpha_t^2}{2m^2}\mathbb{E}\|\nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\|^2 \\
&\le -\frac{\alpha_t}{m}\mathbb{E}\|\nabla F_{\mathcal{S}}(\mathbf{x}_t)\|^2 + \frac{\alpha_t L}{m}\mathbb{E}\|\nabla f(\mathbf{x}_t;\mathbf{z}_{j_t(i_t)}) - \nabla f(\mathbf{x}_{t-\tau_t}(i_t);\mathbf{z}_{j_t(i_t)})\| + \frac{\beta L^2\alpha_t^2}{2m^2} \\
&\overset{(a)}{\le} -\frac{2\gamma\alpha_t}{m}\mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_t) - F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)] + \frac{\beta\alpha_t L}{m}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| + \|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}(i_t)\|\right] + \frac{\beta L^2\alpha_t^2}{2m^2} \\
&\le -\frac{2\gamma\alpha_t}{m}\mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_t) - F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)] + \frac{\beta\alpha_t L^2}{m}\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{\beta L^2\alpha_t^2}{2m^2},
\end{aligned}
$$

where (a) uses the following γ-PŁ condition

$$
2\gamma[F_{\mathcal{S}}(\mathbf{x}_t) - F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)] \le \|\nabla F_{\mathcal{S}}(\mathbf{x}_t)\|^2. \tag{B.31}
$$

Then we have

$$
\sum_{t=1}^T\alpha_t\mathbb{E}\left[F_{\mathcal{S}}(\mathbf{x}_t) - F_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}^*)\right] \le \frac{m}{2\gamma}\mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_1) - F_{\mathcal{S}}(\mathbf{x}_{T+1})] + \frac{\beta L^2}{2\gamma}\sum_{t=1}^T\alpha_t\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{\beta L^2}{4\gamma m}\sum_{t=1}^T\alpha_t^2.
$$

The optimization error satisfies

$$\begin{aligned}
\epsilon_{\rm opt} &= \mathbb{E}[F_{\mathcal{S}}(\overline{\mathbf{x}}_T) - F_{\mathcal{S}}(\mathbf{x}^*_{\mathcal{S}})] \le \frac{\sum_{t=1}^T \alpha_t \mathbb{E}\left[F_{\mathcal{S}}(\mathbf{x}_t) - F_{\mathcal{S}}(\mathbf{x}^*_{\mathcal{S}})\right]}{\sum_{t=1}^T \alpha_t} \\
&\le \frac{m\mathbb{E}[F_{\mathcal{S}}(\mathbf{x}_1) - F_{\mathcal{S}}(\mathbf{x}_{T+1})]}{2\gamma\sum_{t=1}^T \alpha_t} + \frac{\beta L^2}{2\gamma\sum_{t=1}^T \alpha_t}\sum_{t=1}^T \alpha_t\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{\beta L^2\sum_{t=1}^T \alpha_t^2}{4\gamma m\sum_{t=1}^T\alpha_t} \\
&\le \frac{Lmr}{\gamma\sum_{t=1}^T \alpha_t} + \frac{\beta L^2}{2\gamma\sum_{t=1}^T \alpha_t}\sum_{t=1}^T \alpha_t\left[\sum_{s=1}^{t-\tau_t-1}\alpha_s\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha_s}{m}\right] + \frac{\beta L^2\sum_{t=1}^T \alpha_t^2}{4\gamma m\sum_{t=1}^T\alpha_t}.
\end{aligned}$$

For the constant learning rate $\alpha_t = \alpha$

$$\begin{aligned}
\epsilon_{\rm opt} &\le \frac{Lmr}{\gamma\sum_{t=1}^T \alpha} + \frac{\beta L^2}{2\gamma\sum_{t=1}^T \alpha}\sum_{t=1}^T \alpha\left[\sum_{s=1}^{t-\tau_t-1}\alpha\lambda^{t-\tau_t-1-s} + \sum_{s=t-\tau_t}^{t-1}\frac{\alpha}{m}\right] + \frac{\beta L^2\sum_{t=1}^T \alpha^2}{4\gamma m\sum_{t=1}^T\alpha} \\
&\le \frac{Lmr}{T\gamma\alpha} + \frac{\beta L^2}{2T\gamma\alpha}\sum_{t=1}^T \alpha^2\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right] + \frac{\beta L^2 T\alpha^2}{4\gamma m T\alpha} \\
&\le \frac{Lmr}{T\gamma\alpha} + \frac{\beta L^2\alpha}{2\gamma}\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right] + \frac{\beta L^2\alpha}{4\gamma m}.
\end{aligned} \tag{B.32}$$

For the decreasing learning rate $\alpha_t = \frac{mc}{t+1}$, we have

$$\begin{aligned}
\epsilon_{\rm opt} &\le \frac{Lmr}{\gamma\sum_{t=1}^T \frac{mc}{t+1}} + \frac{\beta L^2}{2\gamma\sum_{t=1}^T \frac{mc}{t+1}}\sum_{t=1}^T \frac{mc}{t+1}\left[mc\sum_{s=1}^{t-\tau_t-1}\frac{\lambda^{t-\tau_t-1-s}}{s+1} + c\sum_{s=t-\tau_t}^{t-1}\frac{1}{s+1}\right] + \frac{\beta L^2\sum_{t=1}^T (\frac{mc}{t+1})^2}{4\gamma m\sum_{t=1}^T \frac{mc}{t+1}} \\
&\overset{(a)}{\le} \frac{2Lr}{\gamma c\ln(T+1)} + \frac{\beta L^2 c}{\gamma\ln(T+1)}\sum_{t=1}^T\frac{1}{t+1}\left[\frac{mC_\lambda}{t\lambda^{\overline{\tau}}} + \frac{\tau_t}{t-\tau_t+1}\right] + \frac{\beta L^2 c}{2\gamma\ln(T+1)} \\
&\overset{(b)}{\le} \frac{2Lr}{\gamma c\ln(T+1)} + \frac{\beta L^2 c}{\gamma\ln(T+1)}\left[\frac{mC_\lambda}{\lambda^{\overline{\tau}}} + \overline{\tau} + \ln(\overline{\tau}+1)\right] + \frac{\beta L^2 c}{2\gamma\ln(T+1)} \\
&\le \left[2Lr + \beta m L^2 c^2(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}) + \frac{\beta L^2 c^2}{2}\right]\frac{1}{\gamma c\ln(T+1)},
\end{aligned} \tag{B.33}$$

where (a) uses inequalities (B.6), (B.29) and (B.30). With $c = \frac{1}{\gamma}$, we then get

$$\epsilon_{\rm opt} \le \left[2Lr + \frac{\beta m L^2}{\gamma^2}(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}) + \frac{\beta L^2}{2\gamma^2}\right]\frac{1}{\ln(T+1)}.$$

For the constant learning rate $\alpha_t = \alpha$, it follows from (B.26) that

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{x}}_T - \overline{\mathbf{x}}'_T\| &\le \frac{\frac{2L\alpha(1+\beta n\alpha-\lambda)}{\beta n(1-\lambda)}\sum_{t=1}^T(1+\frac{\beta\alpha}{m})^{t-1}}{T\alpha} \\
&\le \frac{\frac{2L\alpha(1+\beta n\alpha-\lambda)}{\beta n(1-\lambda)}\frac{m}{\beta\alpha}(1+\frac{\beta\alpha}{m})^T}{T\alpha} \\
&\le \frac{2Lm(1+\beta n\alpha-\lambda)}{\beta^2 n\alpha(1-\lambda)}\frac{(1+\frac{\beta\alpha}{m})^T}{T}.
\end{aligned}$$

Then the uniform stability bound of $\overline{\mathbf{x}}_T$ satisfies

$$\epsilon_{\rm ave-stab} \le \frac{2L^2 m(1+\beta n\alpha-\lambda)}{\beta^2 n\alpha(1-\lambda)}\frac{(1+\frac{\beta\alpha}{m})^T}{T}.$$

Combined with the optimization error (B.32), we have

$$\begin{aligned}
\epsilon_{\rm exc} &\le \epsilon_{\rm ave-stab} + \epsilon_{\rm opt} \\
&\le \frac{2L^2 m(1+\beta n\alpha-\lambda)}{\beta^2 n\alpha(1-\lambda)}\frac{(1+\frac{\beta\alpha}{m})^T}{T} + \frac{Lmr}{T\gamma\alpha} + \frac{\beta L^2\alpha}{2\gamma}\left[\frac{1}{1-\lambda} + \frac{\overline{\tau}}{m}\right] + \frac{\beta L^2\alpha}{4\gamma m}.
\end{aligned}$$

For the decreasing learning rate $\alpha_t = \frac{mc}{t+1}$, it follows from (B.27)

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{x}}_T - \overline{\mathbf{x}}'_T\| &\le \frac{\sum_{t=1}^T \frac{mc}{t+1}\left[\frac{2L(1+\beta nmc-\lambda)}{\beta n(1-\lambda)}\right] t^{\beta c}}{\sum_{t=1}^T \frac{mc}{t+1}} \\ &\overset{(a)}{\le} \left[\frac{4L(1+\beta nmc-\lambda)}{\beta n(1-\lambda)}\right] \frac{\sum_{t=1}^T (t+1)^{\beta c-1}}{\ln(T+1)} \\ &\overset{(b)}{\le} \left[\frac{4L(1+\beta nmc-\lambda)}{\beta^2 nc(1-\lambda)}\right] \frac{(T+1)^{\beta c}}{\ln(T+1)},\end{aligned}$$

where (a) uses inequality (B.29). With $c < \frac{1}{\beta}$, (b) follows from

$$\sum_{t=1}^T (t+1)^{\beta c-1} \le \sum_{t=1}^T \int_t^{t+1} x^{\beta c-1} dx \le \int_1^{T+1} x^{\beta c-1} dx = \frac{1}{\beta c}(T+1)^{\beta c}.$$

Then the uniform stability of $\overline{\mathbf{x}}_T$ satisfies

$$\epsilon_{\text{ave}-\text{stab}} \le \frac{4L^2}{\beta c}\left(\frac{1}{\beta n} + \frac{mc}{1-\lambda}\right)\frac{(T+1)^{\beta c}}{\ln(T+1)}.$$

Combined with the optimization error (B.33), we have

$$\begin{aligned}\epsilon_{\text{exc}} &\le \epsilon_{\text{ave}-\text{stab}} + \epsilon_{\text{opt}} \\ &\le \frac{4L^2}{\beta c}\left(\frac{1}{\beta n} + \frac{mc}{1-\lambda}\right)\frac{(T+1)^{\beta c}}{\ln(T+1)} + \Big[2Lr + \beta mL^2c^2\big(\frac{C_\lambda}{\lambda^{\overline{\tau}}} + \frac{2\overline{\tau}}{m}\big) + \frac{\beta L^2 c^2}{2}\Big]\frac{1}{\gamma c \ln(T+1)}.\end{aligned}$$

■

References

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 1225–1234. PMLR.

Sun, T.; Li, D.; and Wang, B. 2021. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9756–9764.

Lian, X.; Zhang, W.; Zhang, C.; and Liu, J. 2018. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, 3043–3052. PMLR.

Regatti, J.; Tendolkar, G.; Zhou, Y.; Gupta, A.; and Liang, Y. 2019. Distributed SGD generalizes well under asynchrony. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 863–870. IEEE.