

## A Additional Experiments

### A.1 Additional Results

**Additional experiments on multilingual settings** We conducted multilingual experiments by fine-tuning both the Spanish and the original English Alpaca-GPT4 models, and evaluated them on the ARC challenge. Due to time and space constraints, we only compare ParaBlock with FedBCD here to demonstrate that ParaBlock can still achieve performance comparable to FedBCD.

**GPU memory consumption** We first present the peak GPU memory consumption for all baselines in Table 8 when fine-tuning on the Alpaca-GPT4 dataset (Peng et al., 2023). The results indicate that ParaBlock and FedBCD incur the lowest GPU memory costs among the fine-tuning methods. FedCyBGD exhibits a relatively higher memory cost compared to FedBCD and ParaBlock due to differences in block assignment. Furthermore, the proposed ParaBlock consistently consumes less memory than LoRA-based methods.

Table 8: GPU peak consumption when fine-tuning the Alpaca-GPT4 dataset.

Methods	Llama 3-8B	Llama 3.2-3B
FedLoRA	27.0G	14.3G
FFA-LoRA	26.8G	14.2G
FLoRA	27.0G	14.3G
FedCyBGD	29.6G	13.9G
FedBCD	23.3G	10.0G
ParaBlock	23.3G	10.0G

**Orthogonal to existing communication efficient FL methods** Previous studies in FL have explored various communication-efficient techniques, such as model/update compression, quantization, and pruning (Reisizadeh et al., 2020; Haddadpour et al., 2019; Wang et al., 2022; Jiang et al., 2022). These approaches primarily aim to reduce the number of communication bits, thereby improving the communication efficiency in FL systems. In contrast, our proposed ParaBlock targets the latency inherent in the communication process, making ParaBlock orthogonal to existing communication-reduction methods. In Table 9, 1) we compare ParaBlock with existing communication-reduction method applied to FedBCD, and 2) we show that for those computation time cannot overlap the communication time, we can further apply existing communication-reduction methods to improve the communication efficiency.

The top part of Table 9, it shows that ParaBlock achieves superior performance compared to directly applying top- $k$  compression (as utilized in (Wang et al., 2022; Li et al., 2024)) with top 20% ratio to the standard FedBCD baseline, while also requiring less runtime. This is because top- $k$  compression, despite reducing communication bits, still necessitates transmitting compressed model at each global round. In contrast, ParaBlock directly reduces the communication latency, resulting in better overall performance than applying compression to the vanilla FedBCD baseline. Moreover, ParaBlock is compatible with existing communication reduction techniques, as shown in the bottom lines in Table 9. By further integrating top- $k$  compression, ParaBlock can effectively reduce the extra communication time with still achieving reasonable performance in both tasks.

**Hyper-parameter details** We conduct learning rate searches to find the best learning rate for each baseline. We grid the learning rate  $\eta_l$  from  $\{3e-7, 1e-6, 3e-6, 1e-5, 3e-5\}$ , and the global learning  $\eta = 1$  for all experiments. The extra hyper-parameters for AdamW optimizer is following the default parameter in **Trainer**, i.e.,  $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-6}$ . Table 10 summarize the learning rates in our experiments.

Table 9: Comparison with Top- $k$  compression methods, where MT-B is the abbreviation for MT-Bench.

Method	MT-B $\uparrow$	Runtime(m) $\downarrow$	GSM8K $\uparrow$	Runtime(m) $\downarrow$
100M/s, ebs=4				
FedBCD	<b>5.14</b>	30.2	54.74	24.9
+Top-20%	5.00	26.4	54.12	21.1
ParaBlock	<b>5.14</b>	<u>21.1</u>	<b>55.88</b>	<u>15.8</u>
+Top-20%	5.11	<b>21.0</b>	<u>55.27</u>	<b>15.6</b>
50M/s, ebs=2				
FedBCD	<b>5.03</b>	30.0	<b>54.66</b>	26.8
+Top-20%	<u>4.99</u>	22.5	53.90	<u>19.2</u>
ParaBlock	<u>4.99</u>	<u>19.4</u>	53.53	19.4
+Top-20%	4.98	<b>11.6</b>	<u>54.14</u>	<b>11.6</b>

Table 10: Learning rates in our experiments

	Alpaca-GPT4		Math Instruct	
	Llama 3-8B	Llama 3.2-3B	Llama 3-8B	Llama 3.2-3B
FFT	3e-7	1e-7	1e-7	1e-6
FedIT	3e-6	3e-7	3e-6	3e-5
FFA-LoRA	3e-6	3e-7	3e-6	3e-5
FLoRA	3e-6	3e-7	3e-6	3e-5
FedCyBGD	1e-5	1e-6	3e-6	3e-5
FedBCD	1e-5	1e-6	3e-6	3e-5
ParaBlock	1e-5	1e-6	3e-6	3e-5

## B Theoretical Analysis

### B.1 Additional Discussions about Assumptions

**Discussion about Assumption 5.1.** The block-wise smoothness property could be naturally implied by the general smoothness of objective function. Given to the non-negativity of the norm operation, there is  $\|\nabla_b f(\theta_1) - \nabla_b f(\theta_2)\| \leq \|\nabla f(\theta_1) - \nabla f(\theta_2)\| = L\|\theta_1 - \theta_2\|$ . We adopt the general smoothness for convenience and notational clarity. Alternatively, if we assume block-wise smoothness: for each block  $b$ , there is  $\|\nabla_b f(\theta_1) - \nabla_b f(\theta_2)\| \leq L_b\|\theta_1 - \theta_2\|$ , and with  $\bar{L} = \max_b L_b$ , the convergence analysis can be modified accordingly. The convergence rate will maintain  $O(1/\sqrt{T})$  but depend on  $\bar{L}$  instead of  $L$ .

**Discussion about Assumption 5.2.** The block-wise heterogeneity naturally follows from the original bounded heterogeneity in Assumption 5.2 as well. Using the property of partial derivatives,  $\nabla_b f(\theta_t) = [\nabla f(\theta_t)]_b$ , and following the argument in Lemma C.3, we have

$$\frac{1}{N} \sum_{i=1}^N \sum_{b=1}^B \|\nabla f(\theta) - [\nabla f_i(\theta)]_b\|^2 = \frac{1}{N} \sum_{i=1}^N \|\nabla f(\theta) - \nabla f_i(\theta)\|^2 \leq \sigma_g^2 \quad (7)$$

Therefore, for simplicity, we adopts a general bounded variance assumption on the full gradient. Moreover, if we instead assume bounded variances for each block, the convergence rate of  $O(1/\sqrt{T})$  remains valid. We will discuss this in the revision.

## B.2 Convergence Analysis

For the global model of two consecutive steps, there is

$$[\boldsymbol{\theta}_{t+1}]_{b_t} - [\boldsymbol{\theta}_t]_{b_t} = \eta \boldsymbol{\Delta}_t. \quad (8)$$

For  $\bar{\boldsymbol{\Delta}}_t = [\mathbf{0}, \dots, \mathbf{0}, \boldsymbol{\Delta}_t, \mathbf{0}, \dots, \mathbf{0}]$ , where  $[\bar{\boldsymbol{\Delta}}_t]_{b_t} = \boldsymbol{\Delta}_t$ . Given the fact that  $[\nabla f(\boldsymbol{\theta}_t)]_b = \nabla_b f(\boldsymbol{\theta}_t)$ , for each time step  $t$ ,

$$\begin{aligned} & \mathbb{E}[f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t)] \\ &= \mathbb{E}[f(\boldsymbol{\theta}_{t+1})] - f(\boldsymbol{\theta}_t) \\ &\leq \mathbb{E}[\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle] + \frac{L}{2} \mathbb{E}[\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2] \\ &= \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \eta \boldsymbol{\Delta}_t \rangle] + \frac{L}{2} \mathbb{E}[\|\eta \boldsymbol{\Delta}_t\|^2] \\ &= \underbrace{\eta \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t \rangle]}_{I_1} + \underbrace{\frac{\eta^2 L}{2} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2]}_{I_2}. \end{aligned} \quad (9)$$

where the first inequality follows Assumption 5.1 and the second equation holds by  $[\nabla f(\boldsymbol{\theta}_t)]_b = \nabla_b f(\boldsymbol{\theta}_t)$  and Eq. (8). For the first term  $I_1$ , there is

$$\begin{aligned} I_1 &= \eta \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t \rangle] \\ &= \eta \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t + \eta_l K \nabla_{b_t} f(\boldsymbol{\theta}_t) - \eta_l K \nabla_{b_t} f(\boldsymbol{\theta}_t) \rangle] \\ &= -\eta \eta_l K \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \eta \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t + \eta_l K \nabla_{b_t} f(\boldsymbol{\theta}_t) \rangle]. \end{aligned} \quad (10)$$

Then

$$\begin{aligned} & \eta \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t + \eta_l K \nabla_{b_t} f(\boldsymbol{\theta}_t) \rangle] \\ &= \eta \mathbb{E} \left[ \left\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Delta}_t^i + \frac{\eta_l K}{N} \sum_{i=1}^N \nabla_{b_t} f(\boldsymbol{\theta}_t) \right\rangle \right] \\ &= \eta \mathbb{E} \left[ \left\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), -\frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_{t,k}^i + \frac{\eta_l K}{N} \sum_{i=1}^N \nabla_{b_t} f(\boldsymbol{\theta}_t) \right\rangle \right] \\ &= \eta \mathbb{E} \left[ \left\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), -\frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) + \frac{\eta_l K}{N} \sum_{i=1}^N \nabla_{b_t} f_i(\boldsymbol{\theta}_t) \right\rangle \right] \\ &= \eta \mathbb{E} \left[ \left\langle \sqrt{\eta_l K} \cdot \nabla_{b_t} f(\boldsymbol{\theta}_t), -\frac{\sqrt{\eta_l K}}{NK} \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) + \frac{\sqrt{\eta_l K}}{N} \sum_{i=1}^N \nabla_{b_t} f_i(\boldsymbol{\theta}_t) \right\rangle \right] \\ &= \frac{\eta \eta_l K}{2} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{\eta \eta_l}{2N^2 K} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} [\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_t)] \right\|^2 \right] \\ &\quad - \frac{\eta \eta_l}{2N^2 K} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right], \end{aligned} \quad (11)$$

where the third equation holds by the unbiased-ness of stochastic gradient, and the last one holds by the fact of  $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2} [\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{a} - \mathbf{b}\|^2]$ . For the second term in Eq. (11), there is

$$\begin{aligned}
& \frac{\eta\eta_l}{2N^2K} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} [\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_t)] \right\|^2 \right] \\
& \leq \frac{\eta\eta_l}{2N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E} [\|\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_t)\|^2] \\
& \leq \frac{\eta\eta_l L^2}{2N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2].
\end{aligned} \tag{12}$$

Therefore, for the whole  $I_1$  term, we have

$$\begin{aligned}
I_1 & \leq -\eta\eta_l K \mathbb{E} [\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{\eta\eta_l K}{2} \mathbb{E} [\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{\eta\eta_l L^2}{2N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2] \\
& \quad - \frac{\eta\eta_l}{2N^2K} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right] \\
& = -\frac{\eta\eta_l K}{2} \mathbb{E} [\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{\eta\eta_l L^2}{2N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2] - \frac{\eta\eta_l}{2N^2K} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right].
\end{aligned} \tag{13}$$

where the equation holds by Lemma C.3

Note that the model  $\boldsymbol{\theta}_{t,k}^{i,b}$  is the  $k$ -th local step model for update block  $b_t$  at time step  $t$ , thus the previous  $b_{t-1}$  block has been updated yet, i.e.,

$$\begin{aligned}
\boldsymbol{\theta}_{t,k}^i &= \text{LocalBlockTraining}(\boldsymbol{\theta}_{t,0}^i, \eta_l, k), \\
\boldsymbol{\theta}_{t,0}^i &= \boldsymbol{\theta}_t^i = \boldsymbol{\theta}_{t-1}^i + \eta \bar{\boldsymbol{\Delta}}_{t-1}^i + \eta \bar{\boldsymbol{\Delta}}_{t-2}^i - \eta \bar{\boldsymbol{\Delta}}_{t-2}^i \\
&= \boldsymbol{\theta}_{t-1}^i + \eta \bar{\boldsymbol{\Delta}}_{t-1}^i, \\
\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \eta \bar{\boldsymbol{\Delta}}_{t-1},
\end{aligned} \tag{14}$$

then

$$\begin{aligned}
\mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2] &= \mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i + \boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\
&\leq 2\mathbb{E} [\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] + 2\mathbb{E} [\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2],
\end{aligned} \tag{15}$$

where the first term consists of  $k$  steps of local updates, while the second term includes the updates difference when updating the  $b_{t-1}$  block. For  $k = 0, \dots, K-1$ , we obtain

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] &= \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|_{b_t}^2] \\
&= \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i - \boldsymbol{\theta}_{t,0}^i - \eta_l \mathbf{g}_{t,k}^i\|_{b_t}^2] \\
&\leq \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i\|_{b_t} - \|\boldsymbol{\theta}_{t,0}^i\|_{b_t} - \eta_l (\|\mathbf{g}_{t,k}^i\|_{b_t} - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k-1}^i) + \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k-1}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) + \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) \\
&\quad - \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i) + \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i))\|^2] \\
&\leq \left(1 + \frac{1}{2K-1}\right) \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i\|_{b_t} - \|\boldsymbol{\theta}_{t,0}^i\|_{b_t}\|^2] + \mathbb{E}[\|\eta_l (\|\mathbf{g}_{t,k}^i\|_{b_t} - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k-1}^i))\|^2] \\
&\quad + 6K \mathbb{E}[\|\eta_l (\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k-1}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i))\|^2] + 6K \mathbb{E}[\|\eta_l (\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i))\|^2] \\
&\quad + 6K \mathbb{E}[\|\eta_l \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] \\
&\leq \left(1 + \frac{1}{2K-1} + 6K \eta_l^2 L^2\right) \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i\|_{b_t} - \|\boldsymbol{\theta}_{t,0}^i\|_{b_t}\|^2] + \eta_l^2 \sigma^2 \\
&\quad + 6K \mathbb{E}[\|\eta_l (\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i))\|^2] + 6K \mathbb{E}[\|\eta_l \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] \\
&= \left(1 + \frac{1}{2K-1} + 6K \eta_l^2 L^2\right) \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i - \boldsymbol{\theta}_{t,0}^i\|^2] + \eta_l^2 \sigma^2 \\
&\quad + 6K \mathbb{E}[\|\eta_l (\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i))\|^2] + 6K \mathbb{E}[\|\eta_l \nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2], \tag{16}
\end{aligned}$$

then by taking average over all clients  $i \in [N]$ ,

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] &\leq \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{1}{2K-1} + 6K \eta_l^2 L^2\right) \mathbb{E}[\|\boldsymbol{\theta}_{t,k-1}^i - \boldsymbol{\theta}_{t,0}^i\|^2] + \eta_l^2 \sigma^2 \\
&\quad + 6K \eta_l^2 \sigma_g^2 + \frac{6K \eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2]. \tag{17}
\end{aligned}$$

Since  $\eta_l \leq \frac{1}{8KL}$ , unrolling the recursion, then we have

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] \\
&\leq \sum_{p=0}^{k-1} \left(1 + \frac{1}{K-1}\right)^p \left[\eta_l^2 \sigma^2 + 6K \eta_l^2 \sigma_g^2 + \frac{6K \eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2]\right] \\
&\leq (K-1) \left[\left(1 + \frac{1}{K-1}\right)^K - 1\right] \left[\eta_l^2 \sigma^2 + 6K \eta_l^2 \sigma_g^2 + \frac{6K \eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2]\right] \\
&\leq 5K \eta_l^2 \sigma^2 + 30K^2 \eta_l^2 \sigma_g^2 + \frac{30K^2 \eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2], \tag{18}
\end{aligned}$$

for the last item, we have

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f(\boldsymbol{\theta}_t) + \nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\leq \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\leq \frac{2L^2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] + 2\mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2], \tag{19}
\end{aligned}$$

where there is

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] &= \mathbb{E}[\|\boldsymbol{\theta}_{t-1} + \eta \bar{\boldsymbol{\Delta}}_{t-1}^i - (\boldsymbol{\theta}_{t-1} + \eta \bar{\boldsymbol{\Delta}}_{t-1})\|^2] \\
&= \eta^2 \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_{t-1}^i - \bar{\boldsymbol{\Delta}}_{t-1}\|^2] \\
&\leq 2\eta^2 \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_{t-1}^i\|^2] + 2\eta^2 \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_{t-1}\|^2] \\
&= 2\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] + 2\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2].
\end{aligned} \tag{20}$$

Merging items together, then we obtain

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2] &\leq \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] + \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\
&\leq 10K\eta_l^2\sigma^2 + 60K^2\eta_l^2\sigma_g^2 + \frac{60K^2\eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] + \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\
&\leq 10K\eta_l^2\sigma^2 + 60K^2\eta_l^2\sigma_g^2 + 120K^2\eta_l^2 \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + \frac{120K^2\eta_l^2L^2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] + \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\
&\leq 10K\eta_l^2\sigma^2 + 60K^2\eta_l^2\sigma_g^2 + 120K^2\eta_l^2 \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{4}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\
&\leq 10K\eta_l^2\sigma^2 + 60K^2\eta_l^2\sigma_g^2 + 120K^2\eta_l^2 \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + 8\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2] + \frac{8\eta^2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2].
\end{aligned} \tag{21}$$

Therefore, reorganizing the  $I_1$  term, we obtain

$$\begin{aligned}
I_1 &\leq -\frac{\eta\eta_l K}{2} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \frac{\eta\eta_l L^2}{2N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_t\|^2] - \frac{\eta\eta_l}{2N^2 K} \mathbb{E}\left[\left\|\sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)\right\|^2\right] \\
&\leq -\frac{\eta\eta_l K}{2} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \eta\eta_l L^2 K \left[5K\eta_l^2\sigma^2 + 30K^2\eta_l^2\sigma_g^2 + 60K^2\eta_l^2 \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2]\right] \\
&\quad + 4\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2] + \frac{4\eta^2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] - \frac{\eta\eta_l}{2N^2 K} \mathbb{E}\left[\left\|\sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)\right\|^2\right]
\end{aligned} \tag{22}$$

Summing up Eq. (9),

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{\theta}_T)] - f(\boldsymbol{\theta}_0) &\leq \eta \sum_{t=0}^{T-1} \mathbb{E}[\langle \nabla_{b_t} f(\boldsymbol{\theta}_t), \boldsymbol{\Delta}_t \rangle] + \frac{\eta^2 L}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\leq -\frac{\eta \eta_l K}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \eta \eta_l L^2 K \left[ 5TK\eta_l^2 \sigma^2 + 30TK^2\eta_l^2 \sigma_g^2 \right. \\
&\quad \left. + 60K^2\eta_l^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + 4\eta^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2] + \frac{4\eta^2}{N} \sum_{t=0}^{T-1} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] \right] \\
&\quad - \frac{\eta \eta_l}{2N^2 K} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right] + \frac{\eta^2 L}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\leq -\frac{\eta \eta_l K}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + 60K^3\eta\eta_l^3 L^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + \eta \eta_l L^2 K (5TK\eta_l^2 \sigma^2 + 30TK^2\eta_l^2 \sigma_g^2) + \left( 4\eta^3 \eta_l L^2 K + \frac{\eta^2 L}{2} \right) \sum_{t=0}^T \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\quad + \frac{4\eta^3 \eta_l L^2 K}{N} \sum_{t=0}^{T-1} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] - \frac{\eta \eta_l}{2N^2 K} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right]. \tag{23}
\end{aligned}$$

By Lemma C.2, the inequality becomes

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{\theta}_T)] - f(\boldsymbol{\theta}_0) &\leq -\frac{\eta \eta_l K}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + 60K^3\eta\eta_l^3 L^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + \eta \eta_l L^2 K (5TK\eta_l^2 \sigma^2 + 30TK^2\eta_l^2 \sigma_g^2) + \left( 4\eta^3 \eta_l L^2 K + \frac{\eta^2 L}{2} \right) \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] \\
&\quad + 4\eta^3 \eta_l L^2 K \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] + \frac{\eta \eta_l K}{4} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + 4\eta^3 \eta_l L^2 K (2TK\eta_l^2 \sigma^2 + 20TK^2\eta_l^4 L^2 \sigma^2 + 120TK^3\eta_l^4 L^2 \sigma_g^2) \\
&\quad - \frac{\eta \eta_l}{2N^2 K} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2 \right]. \tag{24}
\end{aligned}$$

By condition on learning rates, i.e.,  $\eta_l \leq \frac{1}{22KL}$  and  $\eta_l \leq \frac{1}{4KL\eta}$ ,

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{\theta}_T)] - f(\boldsymbol{\theta}_0) &\leq -\frac{\eta\eta_l K}{8} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \eta\eta_l L^2 K (5TK\eta_l^2 \sigma^2 + 30TK^2\eta_l^2 \sigma_g^2) \\
&\quad + \left(8\eta^3\eta_l L^2 K + \frac{\eta^2 L}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \\
&\quad + 4\eta^3\eta_l L^2 K (2TK\eta_l^2 \sigma^2 + 20TK^2\eta_l^4 L^2 \sigma^2 + 120TK^3\eta_l^4 L^2 \sigma_g^2) \\
&\quad - \frac{\eta\eta_l}{2N^2 K} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)\right\|^2\right] \\
&\leq -\frac{\eta\eta_l K}{8} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] + \eta\eta_l L^2 K (5TK\eta_l^2 \sigma^2 + 30TK^2\eta_l^2 \sigma_g^2) \\
&\quad + \left(8\eta^3\eta_l L^2 K + \frac{\eta^2 L}{2}\right) \frac{TK\eta_l^2}{N} \sigma^2 \\
&\quad + 4\eta^3\eta_l L^2 K (2TK\eta_l^2 \sigma^2 + 20TK^2\eta_l^4 L^2 \sigma^2 + 120TK^3\eta_l^4 L^2 \sigma_g^2). \tag{25}
\end{aligned}$$

Then,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] &\leq \frac{8}{\eta\eta_l K} [f(\boldsymbol{\theta}_0) - \mathbb{E}[f(\boldsymbol{\theta}_T)]] + 40TK\eta_l^2 L^2 \sigma^2 + 240TK^2\eta_l^2 L^2 \sigma_g^2 \\
&\quad + \left(8\eta^2\eta_l L^2 K + \frac{\eta L}{2}\right) \frac{T\eta_l}{N} \sigma^2 \\
&\quad + 32\eta^2 L^2 (2TK\eta_l^2 \sigma^2 + 20TK^2\eta_l^4 L^2 \sigma^2 + 120TK^3\eta_l^4 L^2 \sigma_g^2). \tag{26}
\end{aligned}$$

Dividing by  $T$ , there is

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] &\leq \frac{8}{\eta\eta_l TK} [f(\boldsymbol{\theta}_0) - \mathbb{E}[f(\boldsymbol{\theta}_T)]] + 40K\eta_l^2 L^2 \sigma^2 + 240K^2\eta_l^2 L^2 \sigma_g^2 \\
&\quad + \left(8\eta^2\eta_l L^2 K + \frac{\eta L}{2}\right) \frac{\eta_l}{N} \sigma^2 \\
&\quad + 32\eta^2 L^2 (2K\eta_l^2 \sigma^2 + 20K^2\eta_l^4 L^2 \sigma^2 + 120K^3\eta_l^4 L^2 \sigma_g^2). \tag{27}
\end{aligned}$$

### B.3 Extension to Local Adaptive Optimizer

The theoretical analysis of the proposed ParaBlock is not limited to the local SGD setting. Essentially, the main differences between the convergence analysis under SGD and adaptive optimizers can be summarized as follows:

- The local updates  $\Delta_t^i$  are aggregated to  $\Delta_t$  on the server. Hence, the most crucial part of modifying to AdamW is to deal with these  $\Delta$  terms.



- For  $\Delta_t^i$  in Adam, there is  $\Delta_t^i = \theta_{t,K}^i - \theta_{t,0}^i = \sum_{k=1}^K (\theta_{t,k}^i - \theta_{t,k-1}^i)$ . Thus there is

$$\begin{aligned}
\Delta_t &= \frac{1}{N} \sum_{i=1}^N \Delta_t^i = \frac{1}{N} \sum_{i=1}^N [\theta_{t,K}^i - \theta_{t,0}^i] = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (\theta_{t,k}^i - \theta_{t,k-1}^i) \\
&= \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \eta_l \frac{\mathbf{m}_{t,k}^i}{\sqrt{\mathbf{v}_{t,k}^i + \epsilon}}, \\
\Rightarrow \|\Delta_t\|^2 &= \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \eta_l \frac{\mathbf{m}_{t,k}^i}{\sqrt{\mathbf{v}_{t,k}^i + \epsilon}} \right\|^2 \\
&\leq \frac{\eta_l^2}{\epsilon} \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \mathbf{m}_{t,k}^i \right\|^2 \\
&= \frac{\eta_l^2}{\epsilon} \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \sum_{j=1}^k (1 - \beta_1) \beta_1^{k-j} \mathbf{g}_{t,j}^i \right\|^2 \\
&= \frac{\eta_l^2}{\epsilon} \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (1 - \beta_1^{K-k+1}) \mathbf{g}_{t,k}^i \right\|^2, \tag{28}
\end{aligned}$$

therefore, by similar theoretical analysis in Lemma [C.1](#), we have

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \frac{\eta_l^2}{\epsilon} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} (1 - \beta_1^{K-k+1}) [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\theta_{t,k}^i) + \nabla_{b_t} f_i(\theta_{t,k}^i)] \right\|^2 \right] \\
&= \frac{\eta_l^2}{\epsilon} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} (1 - \beta_1^{K-k+1}) [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\theta_{t,k}^i)] \right\|^2 \right] \\
&\quad + \frac{\eta_l^2}{\epsilon} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} (1 - \beta_1^{K-k+1}) \nabla_{b_t} f_i(\theta_{t,k}^i) \right\|^2 \right] \\
&\leq \frac{K\eta_l^2}{N\epsilon} \sigma^2 + \frac{\eta_l^2}{N^2\epsilon^2} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\theta_{t,k}^i) \right\|^2 \right]. \tag{29}
\end{aligned}$$

- The properties about bounding  $\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\Delta_t^i\|^2]$  would be also similar to the analysis in Lemma [C.2](#).
- In a nutshell, adopting local Adam achieves the same convergence rate of  $O(1/\sqrt{T})$  as SGD.

## C Supporting Lemmas

**Lemma C.1.** *The global update parameter  $\Delta_t = \frac{1}{N} \sum_{i=1}^N \Delta_t^i$  satisfies*

$$\mathbb{E}[\|\Delta_t\|^2] \leq \frac{K\eta_l^2}{N} \sigma^2 + \frac{\eta_l^2}{N^2} \mathbb{E} \left[ \left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\theta_{t,k}^i) \right\|^2 \right]. \tag{30}$$

*Proof.* By definition,

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\left[\left\| -\frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_{t,k}^i \right\|^2\right] \\
&= \mathbb{E}\left[\left\| -\frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) + \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)] \right\|^2\right] \\
&= \mathbb{E}\left[\left\| \frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)] \right\|^2\right] + \mathbb{E}\left[\left\| \frac{\eta_l}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2\right] \\
&\leq \frac{K\eta_l^2}{N} \sigma^2 + \frac{\eta_l^2}{N^2} \mathbb{E}\left[\left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2\right], \tag{31}
\end{aligned}$$

where the third equation holds by the unbiased-ness of the stochastic gradient, and the inequality holds by Assumption 5.2.  $\square$

**Lemma C.2.** *The global update parameter  $\Delta_t^i$  satisfies*

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\Delta_t^i\|^2] &\leq \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] + \frac{1}{2\eta^2 L^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + 2TK\eta_l^2 \sigma^2 + 20TK^2\eta_l^4 L^2 \sigma^2 + 120TK^3\eta_l^4 L^2 \sigma_g^2. \tag{32}
\end{aligned}$$

*Proof.* By definition,

$$\begin{aligned}
\mathbb{E}[\|\Delta_t^i\|^2] &= \mathbb{E}\left[\left\| -\eta \sum_{k=0}^{K-1} \mathbf{g}_{t,k}^i \right\|^2\right] \\
&= \mathbb{E}\left[\left\| -\eta \sum_{k=0}^{K-1} [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) + \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)] \right\|^2\right] \\
&= \mathbb{E}\left[\left\| \eta \sum_{k=0}^{K-1} [\mathbf{g}_{t,k}^i - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)] \right\|^2\right] + \mathbb{E}\left[\left\| \eta \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2\right] \\
&\leq K\eta_l^2 \sigma^2 + \eta_l^2 \mathbb{E}\left[\left\| \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2\right], \tag{33}
\end{aligned}$$

where the third equation holds by the unbiased-ness of the stochastic gradient, and the inequality holds by Assumption 5.2.

$$\begin{aligned}
&\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) \right\|^2\right] \\
&= \mathbb{E}\left[\left\| \sum_{k=0}^{K-1} [\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) + \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_t) + \nabla_{b_t} f_i(\boldsymbol{\theta}_t) - \nabla_{b_t} f(\boldsymbol{\theta}_t) + \nabla_{b_t} f(\boldsymbol{\theta}_t)] \right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} [\nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i) - \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i)] \right\|^2\right] + 2\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) \right\|^2\right] \\
&\leq 2K \sum_{k=0}^{K-1} L^2 \mathbb{E}\left[\left\| \boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i \right\|^2\right] + 2K^2 \mathbb{E}\left[\left\| \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,0}^i) \right\|^2\right], \tag{34}
\end{aligned}$$

where

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] \leq 5K\eta_l^2 \sigma^2 + 30K^2\eta_l^2 \sigma_g^2 + \frac{30K^2\eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2], \tag{35}$$

and

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] \leq \frac{2L^2}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] + 2\mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2], \quad (36)$$

where there is

$$\mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \leq 2\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] + 2\eta^2 \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2]. \quad (37)$$

Then

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_t^i\|^2] &\leq K\eta_l^2 \sigma^2 + \frac{\eta_l^2}{N} \sum_{i=1}^N \mathbb{E}\left[\left\|\sum_{k=0}^{K-1} \nabla_{b_t} f_i(\boldsymbol{\theta}_{t,k}^i)\right\|^2\right] \\ &\leq K\eta_l^2 \sigma^2 + \frac{2K\eta_l^2 L^2}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E}[\|\boldsymbol{\theta}_{t,k}^i - \boldsymbol{\theta}_{t,0}^i\|^2] + \frac{2K^2\eta_l^2}{N} \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] \\ &\leq K\eta_l^2 \sigma^2 + 10K^2\eta_l^4 L^2 \sigma^2 + 60K^3\eta_l^4 L^2 \sigma_g^2 + \left(\frac{60K^3\eta_l^4 L^2}{N} + \frac{2K^2\eta_l^2}{N}\right) \sum_{i=1}^N \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_{t,0}^i)\|^2] \\ &\leq K\eta_l^2 \sigma^2 + 10K^2\eta_l^4 L^2 \sigma^2 + 60K^3\eta_l^4 L^2 \sigma_g^2 + \left(\frac{120K^3\eta_l^4 L^4}{N} + \frac{4K^2\eta_l^2 L^2}{N}\right) \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\theta}_{t,0}^i - \boldsymbol{\theta}_t\|^2] \\ &\quad + (120K^3\eta_l^4 L^2 + 4K^2\eta_l^2) \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2] \\ &\leq K\eta_l^2 \sigma^2 + 10K^2\eta_l^4 L^2 \sigma^2 + 60K^3\eta_l^4 L^2 \sigma_g^2 + (240K^3\eta_l^2 \eta^4 L^4 + 8K^2\eta^2 \eta_l^2 L^2) \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}^i\|^2] \\ &\quad + (240K^3\eta_l^2 \eta^4 L^4 + 8K^2\eta^2 \eta_l^2 L^2) \mathbb{E}[\|\boldsymbol{\Delta}_{t-1}\|^2] + (120K^3\eta_l^4 L^2 + 4K^2\eta_l^2) \mathbb{E}[\|\nabla_{b_t} f(\boldsymbol{\theta}_t)\|^2]. \end{aligned} \quad (38)$$

First we previously assume that  $\eta_l \leq \frac{1}{8KL}$ , and also (1) for simplicity, if we have a sequence  $x_t \leq \alpha x_{t-1} + \alpha y_{t-1} + \beta z_t + C$ , then we have

$$\begin{aligned} x_t &\leq \alpha x_{t-1} + \alpha y_{t-1} + \beta z_t + C \\ &\leq \alpha(\alpha x_{t-2} + \alpha y_{t-2} + \beta z_{t-1} + C) + \alpha y_{t-1} + \beta z_t + C \\ &\dots \\ &\leq \alpha^t x_0 + \sum_{i=1}^t \alpha^i y_{i-1} + \sum_{i=1}^t \alpha^{i-1} \beta z_i + C \sum_{i=1}^t \alpha^{i-1}. \end{aligned}$$

(2) for simplicity, if we have a sequence  $x_t \leq \alpha x_{t-1} + \alpha y_{t-1} + \beta z_t + C$ , then we have

$$\begin{aligned} x_t &\leq \alpha x_{t-1} + \alpha y_{t-1} + \beta z_t + C \\ \sum_{t=0}^{T-1} x_t &\leq \alpha \sum_{t=0}^{T-1} x_{t-1} + \alpha \sum_{t=0}^{T-1} y_{t-1} + \beta \sum_{t=0}^{T-1} z_t + C * T \\ &\Rightarrow \\ \sum_{t=0}^{T-1} x_t &\leq \alpha \sum_{t=0}^{T-1} x_t + \alpha \sum_{t=0}^{T-1} y_{t-1} + \beta \sum_{t=0}^{T-1} z_t + C * T \\ (1 - \alpha) \sum_{t=0}^{T-1} x_t &\leq \alpha \sum_{t=0}^{T-1} y_{t-1} + \beta \sum_{t=0}^{T-1} z_t + C * T \\ \sum_{t=0}^{T-1} x_t &\leq \alpha(1 - \alpha)^{-1} \sum_{t=0}^{T-1} y_{t-1} + \beta(1 - \alpha)^{-1} \sum_{t=0}^{T-1} z_t + C(1 - \alpha)^{-1} * T, \end{aligned}$$

we want that  $\frac{1}{2} \leq 1 - \alpha < 1$ , which means  $0 < \alpha \leq \frac{1}{2}$  therefore, we have  $1 < (1 - \alpha)^{-1} \leq 2$ . Moreover, since  $\alpha = 240K^3\eta^2\eta_l^4L^4 + 8K^2\eta^2\eta_l^2L^2 \leq \frac{1}{2}$ , we have  $240K^3\eta_l^4L^2 + 8K^2\eta_l^2 \leq \frac{1}{2\eta^2L^2}$

$$\begin{aligned} \sum_{t=0}^{T-1} x_t &\leq \sum_{t=0}^{T-1} y_t + 2\beta \sum_{t=0}^{T-1} z_t + 2C * T \\ &\Rightarrow \end{aligned} \tag{39}$$

$$\begin{aligned} \sum_{t=0}^{T-1} \left[ \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\Delta_t^i\|^2] \right] &\leq \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] + (240K^3\eta_l^4L^2 + 8K^2\eta_l^2) \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\theta_t)\|^2] \\ &\quad + 2TK\eta_l^2\sigma^2 + 20TK^2\eta_l^4L^2\sigma^2 + 120TK^3\eta_l^4L^2\sigma_g^2 \\ &\leq \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] + \frac{1}{2\eta^2L^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_{b_t} f(\theta_t)\|^2] \\ &\quad + 2TK\eta_l^2\sigma^2 + 20TK^2\eta_l^4L^2\sigma^2 + 120TK^3\eta_l^4L^2\sigma_g^2. \end{aligned} \tag{40}$$

□

**Lemma C.3.** For  $\theta = [\theta^1, \theta^2, \dots, \theta^B]$ , i.e., there is a block partition  $b = 1, 2, \dots, B$  partitioned  $\theta$  into  $B$  blocks, then we have  $\|\theta\|^2 = \sum_{b=1}^B \|\theta^b\|^2$ .

*Proof.*

$$\begin{aligned} &\|\theta^1\|^2 + \|\theta^2\|^2 + \dots + \|\theta^B\|^2 \\ &= \left( \sum_{i=1}^{d_1} (x^{1,i})^2 \right) + \left( \sum_{i=1}^{d_2} (x^{2,i})^2 \right) + \dots + \left( \sum_{i=1}^{d_B} (x^{B,i})^2 \right) \\ &= \sum_{i=1}^d (x^i)^2 = \|\theta\|^2. \end{aligned} \tag{41}$$

□

**Lemma C.4.** For  $\theta = [\theta^1, \theta^2, \dots, \theta^B]$  and  $y = [y^1, y^2, \dots, y^B]$ , i.e., there is a block partition  $b = 1, 2, \dots, B$  partitioned  $\theta$  and  $y$  into  $B$  blocks, then we have  $\langle \theta, y \rangle = \sum_{b=1}^B \langle \theta^b, y^b \rangle$ .

*Proof.*

$$\begin{aligned} &\langle \theta^1, y^1 \rangle + \langle \theta^2, y^2 \rangle + \dots + \langle \theta^B, y^B \rangle \\ &= \sum_{i=1}^{d_1} x^{1,i} y^{1,i} + \sum_{i=1}^{d_2} x^{2,i} y^{2,i} + \dots + \sum_{i=1}^{d_B} x^{B,i} y^{B,i} \\ &= \sum_{i=1}^d x^i y^i = \langle \theta, y \rangle. \end{aligned} \tag{42}$$

□