ResIST: Layer-Wise Decomposition of ResNets for Distributed Training (Supplementary Material)

Chen Dun[1] **Cameron R. Wolfe**[1] **Christopher M. Jermaine**[1] **Anastasios Kyrillidis**[1]

[1]Computer Science Dept., Rice University, Houston, Texas, USA

A ABLATIONS

These experiments provide an understanding of the algorithm's behavior, as well as empirical support for its design.

A.1 DESIGNING RESIST

Extensive ablation experiments are conducted on the CIFAR10 dataset, outlined in Fig. 1, to empirically motivate the design choices made within ResIST (i.e., see Sec. **??**). For the two sub-ResNet case, the naive implementation of ResIST, which evenly splits all convolutional blocks between subnetworks, is shown to perform poorly (i.e., $<$70% on CIFAR10). The accuracy of ResIST is improved over 25% by only allowing select layers to be partitioned and ensuring activations are scaled correctly when performing inference with the full network. The pre-activation ResNet is shown to yield an improvement in accuracy, leading ResIST to perform near optimally with two sub-ResNets.

When ResIST is expanded to eight sub-ResNets, we initially observe a significant decrease in model accuracy. However, as can be seen in Fig. 1, this gap can be closed by enforcing a minimum depth on sub-ResNets and tuning the number of local iterations. By making these extra modifications, ResIST begins to perform similarly with two to eight sub-ResNets, yielding compelling performance.

A.2 SHALLOW ENSEMBLES

The ResIST algorithm requires that independently-trained sub-ResNets must have their parameters synchronized intermittently. Such synchronization, however, can be completely avoided by training each sub-ResNet separately and forming an ensemble (i.e., ResIST without any aggregation). Although maintaining an ensemble has several drawbacks (e.g., slower inference, more parameters, etc.), the training time of the ensemble would nonetheless be reduced in comparison to ResIST by avoiding communication altogether. Therefore, the performance of such an ensemble should be compared to the models trained with ResIST.

Table 1: Performance of indpendently-trained ensembles of shallow ResNets in comparison to ResIST on CIFAR10 and CIFAR100 (denoted as C10 and C100, respectively).

Dataset	Method	2 Model	4 Model	8 Model
C10	Ensemble	92.27 % $\pm$ 0.00	92.56% $\pm$ 0.03	90.67 % $\pm$ 0.04
	ResIST	91.95% $\pm$ 0.32	92.35% $\pm$ 0.22	91.45% $\pm$ 0.30
C100	Ensemble	72.08% $\pm$ 0.05	72.12% $\pm$ 0.04	67.98 % $\pm$ 0.12
	ResIST	70.06% $\pm$ 0.51	71.30% $\pm$ 0.20	70.26% $\pm$ 0.21

Modification	Naive Model						ResIST
Share strided layers		✓	✓	✓	✓	✓	✓
Only Partition Section 3			✓	✓	✓	✓	✓
Scale Activations				✓	✓	✓	✓
Pre-Act ResNet					✓	✓	✓
Minimum Depth						✓	✓
Tune Local Iterations							✓
2 Sub-ResNet Acc.	66.3%	68.3%	84.3%	91.5%	92.0%	-	92.0%
8 Sub-ResNet Acc.	-	-	-	-	86.5%	89.9%	91.3%

Figure 1: Test accuracies on the CIFAR10 dataset for a single run for the major ablation experiments performed with `ResIST`.

Table 2: Test accuracy on CIFAR10 (C10) and CIFAR100 (C100) for deeper architectures trained with `ResIST` and local SGD (LSGD). All tests were performed with 100 local iterations between synchronization rounds. All models were trained for 80 epochs.

			ResNet152			ResNet200		
Dataset	# Machines	Method	Time	Test Acc.	Speedup	Time	Test Acc.	Speedup
C10	2	LSGD	3512s	92.27% ± 0.003		4575s	92.31% ± 0.001	
		`ResIST`	2215s	92.01% ± 0.002	**1.58**×	2380s	92.10% ± 0.001	**1.92**×
	4	LSGD	3598s	91.39% ± 0.001		4357s	91.35% ± 0.000	
		`ResIST`	1054s	90.67% ± 0.001	**3.41**×	1161s	90.27% ± 0.001	**3.75**×
C100	2	LSGD	3528s	70.50% ± 0.003		4639s	71.05% ± 0.005	
		`ResIST`	2291s	70.32% ± 0.005	**1.53**×	2202s	70.71% ± 0.002	**2.10**×
	4	LSGD	3518s	68.39% ± 0.004		4391s	69.05% ± 0.003	
		`ResIST`	1164s	67.27% ± 0.003	**3.02**×	1195s	67.62% ± 0.001	**3.67**×

The performance of sub-ResNet ensembles in comparison to models trained with ResIST is displayed in Table 1. For 8 Sub-ResNets, the shallow ensembles achieve inferior performance in comparison to ResIST. When two and four Sub-ResNets are used, the performance of shallow ensembles and ResIST is comparable (i.e., $< 1\%$ performance difference in most cases). However, it should be noted that such shallow ensembles of two or four sub-ResNets, in comparison to ResIST, cause a $2\times$ to $4\times$ slowdown in inference time (i.e., inference time for a single Sub-ResNet is not significantly faster than that of the global ResNet). Furthermore, the ensembles consume more parameters in comparison to global ResNet trained with ResIST.

A.3 ROBUSTNESS TO LOCAL ITERATIONS

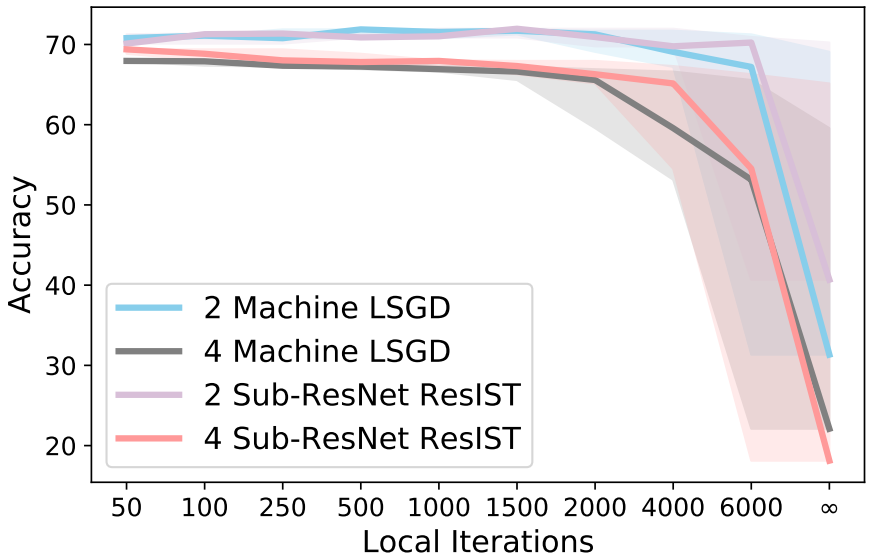

Figure 2: Test accuracy on CIFAR100 for ResNet-101 trained with both ResIST and local SGD (LSGD) with different numbers of local iterations. ∞ local iterations refers to aggregating parameters only once at the end of training (i.e., single-shot averaging). Shaded regions reflect deviations in accuracy.

ResIST is robust to various numbers of local iterations [Lin et al., 2018, Zhang et al., 2016, McMahan et al., 2017]. An extensive sweep over possible values of ℓ is performed on CIFAR100. The results of this experiment are depicted in Fig. 2. As can be seen, ResIST achieves high accuracy even with thousands of local SGD iterations (i.e., previous work typically uses much fewer [Lin et al., 2018]). However, if more sub-ResNets are used, performance tends to deteriorate more quickly as local iterations increase. Due to the robustness of ResIST to large numbers of local iterations, training can be accelerated without deteriorating model performance by simply increasing the value of ℓ. Local SGD was found to demonstrate similar robustness to the number of local iterations, as shown in Fig. 2.

A.4 DEEPER ARCHITECTURES

The ResIST methodology is easily applicable to deeper architectures. To demonstrate this, results are replicated for CIFAR10 and CIFAR100 datasets with ResNet152 and ResNet200. These deeper architectures are identical to the original ResNet101 architecture (i.e., see Fig. **??**). However, more residual blocks are added to the third section of the ResNet (i.e., the highlighted portion of Fig. **??**) to increase the model's depth. It should be noted that convolutional blocks within the third section of the ResNet are partitioned in ResIST by default (see Sec. **??**). As a result, all extra residual blocks within these deeper architectures are partitioned to sub-ResNets by ResIST (i.e., no extra blocks are shared between sub-ResNets), allowing ResIST to achieve greater acceleration in comparison to local SGD.

The results of experiments with deeper ResNets are presented in Table 2. ResIST performs competitively with localSGD in all cases. Furthermore, ResIST achieves a significant speedup in comparison to local SGD that becomes more pronounced as the model becomes deeper. E.g., for 4-GPUs, ResIST completes training $> 3\times$ faster than local SGD for ResNet200 on both datasets. This speedup is caused by a greater ratio of total network blocks being partitioned to sub-ResNets in ResIST. While local SGD must communicate all parameters between machines, ResIST achieves a relative decrease in

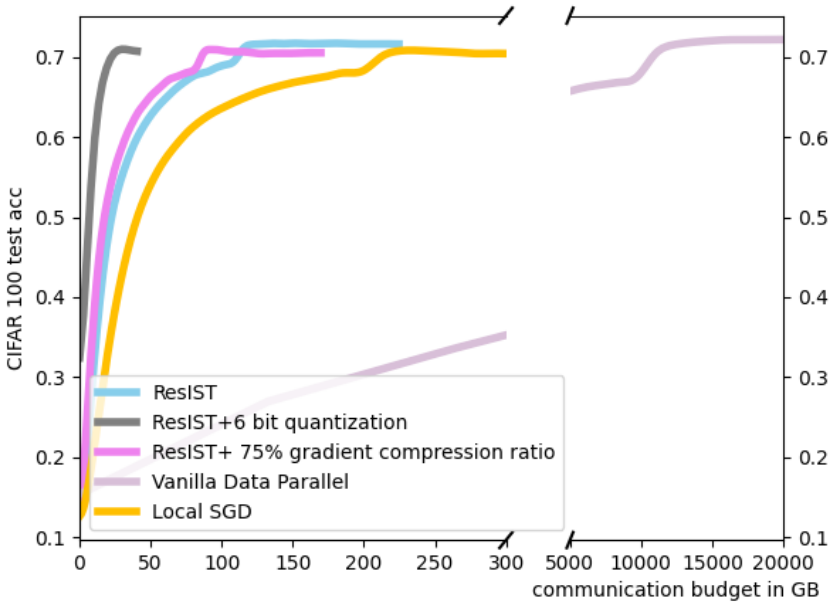

Figure 3: Test accuracy vs. communication budget for ResIST, ResIST+quantization, ResIST+gradient compression, local SGD and vanilla data parallel on CIFAR100. All models are trained over a 4-GPU cluster.

communication by partitioning all extra residual blocks evenly between sub-ResNets.

A.5 RESIST AND QUANTIZATION/SPARSE GRADIENTS

Many quantization [Alistarh et al., 2017, Yu et al., 2019] and sparsification [Aji and Heafield, 2017, Jiang and Agrawal, 2018] techniques have been proposed for reducing communication costs in distributed training. Such techniques focus on compressing communicated data, and they do not interfere with our methodology, which provides a novel approach to model synchronization and training. The proposed approach can be easily combined with existing compression techniques to further reduce communication costs and accelerate training *with no extra tuning or modifications*. To demonstrate that ResIST works well with quantization, we compress all communicated parameters using both four-bit and eight-bit compression. Table 3 shows that ResIST retains its performance until the compression level reaches five-bit and lower. We also perform experiments with sparsification of communicated weights by only keeping 25% of total weights within each synchronization round. Such a strategy reaches a validation peformance of 71.25% on CIFAR100. We summarize the results of all quantization experiments in Fig. 3, where we compare communication budgets across different compression techniques with ResIST. From this figure, it is clear that ResIST is most efficient with six-bit quantization and is compatible with most main-stream compression techniques.

Table 3: Test Accuracy for ResIST combined with quantization on CIFAR10 and CIFAR100 (denoted as C10 and C100).

Dataset	8 bit	7 bit	6 bit	5 bit	4 bit
C10	92.14%	92.26%	91.91%	91.35%	76.33%
C100	71.38%	72.15%	71.37%	68.29%	40.48%

A.6 FURTHER ANALYSIS ON COMMUNICATION COST REDUCTION OF RESIST

The significant communication cost reduction of ResIST, in comparison with Local SGD, comes from that fact that (1) it reduces the communication volume at each global synchronization by only communicating subnetworks (2) it has the

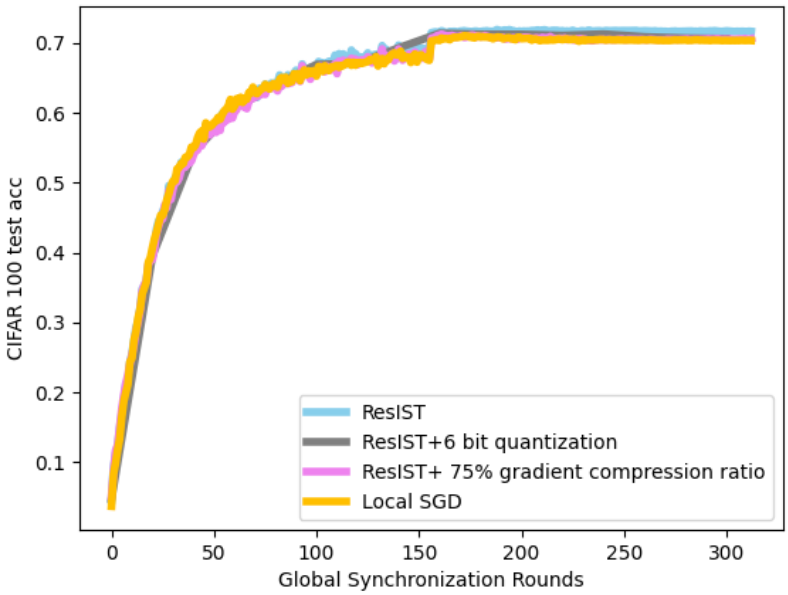

Figure 4: Test accuracy vs. global synchronization rounds for `ResIST`, `ResIST`+quantization, `ResIST`+gradient compression, local SGD and vanilla data parallel on CIFAR100. All models are trained over a 4-GPU cluster.

similar convergence speed in terms of number of global synchronization rounds, as shown in Fig 4. This is also true when `ResIST` is combined with other compression techniques.

A.7 COMPARISON BETWEEN RESIST AND FEDERATED DROPOUT

Federated Dropout, a concurrent work with our paper, also explores splitting a global model into smaller ones, in order to achieve acceleration and preserve final accuracy. Yet, we have not found any work in the literature that handles ResNets and residual blocks specifically and utilized similar additional technique we have applied (such as layer scaling in subnetworks as shown in Figure 1). More importantly, Federated Dropout and its variants do not utilize multiple rounds of local subnetwork training before global synchronization. In other words, Federated Dropout variants need to communicate and synchronize at every training iteration. On the other hand, ResIST and its baseline local SGD locally train for a number of iterations (e.g., 50) on its local model before each synchronization. Thus, compared with Federated Dropout, ResIST significantly reduces the communication frequency and accordingly the total communication volume/cost, as we show in the experiments. Further, Federated Dropout and its variants can only support around 25% dropout rate for each subnetwork, where ResIST, with our additional technique, can support above 50% dropout rate for each subnetwork. This further reduces the total communication cost, local computation cost and local memory cost. In CIFAR100 experiments with 4 workers, to reach test accuracy of 71%, the total communication cost/volume for ResIST is 112.32 GB, while for Federated Dropout the cost is 8138.81 GB. In other words, ResIST achieves 72.46x reduction in the communication cost.

B PROOF FOR RESIST

Suppose we have S workers, for subnetwork v at local training step $l_t \leq \ell$ and global synchronization step $t \leq T$:

$$\mathbf{x}^{(1)}_{v,l_t,t} = \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(1)}_{v,l_t,t}\mathbf{x}_{\mathbf{v},\mathbf{l_t},\mathbf{t}}\right),$$
$$\mathbf{x}^{(h)}_{v,l_t,t} = \mathbf{x}^{(h-1)}_{v,l_t,t} + \frac{c_{res}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}_{v,l_t,t}\mathbf{x}^{(h-1)}_{v,l_t,t}\right)M^{(h)}_{v,t}$$

$$\text{for } 2 \leq h \leq H,$$
$$f_{res}(\mathbf{x}, \theta) = \mathbf{a}_{v,l_t,t}^{\top} \mathbf{x}_{v,l_t,t}^{(H)}$$

where $0 < c_{res} < 1$ is a small constant and $M_{v,t}^{(h)}$ is random binary variable in layer dropout or the indicator in `ResIST` that indicates whether this layer is partitioned to this subnetwork. such mask variable is constant during local training steps and re-sampled/re-assigned at global synchronization step. In `ResIST` and other research on layer dropout for ResNet, last layer is never dropped/paritioned but shared with all workers. Thus, in the following proof, we will follow this setting.Note here we use a $\frac{c_{res}}{H\sqrt{m}}$ scaling. We follow the general assumption made in Du et al. [2019] on some technical conditions on the activation functions σ: There exists a constant $c > 0$ such that $|\sigma(0)| \leq c$ and for any $z, z' \in \mathbb{R}$,

$$\begin{aligned} |\sigma(z) - \sigma(z')| \leq & c|z - z'|, \\ \text{and } |\sigma'(z) - \sigma'(z)| \leq & c|z - z'|. \end{aligned}$$

Also, we assume $\sigma(\cdot)$ is analytic and is not a polynomial function.

In practice, several actication function satisfy the above two assumptions. The guiding example is softplus: $\sigma(z) = \log(1 + \exp(z))$. For softplus both Lipschitz constant and smoothness constant are 1. In this paper, we view all activation function related parameters as constants.

The gradient for subnetwork is

$$\frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}} = \frac{c_{res}}{H\sqrt{m}} \sum_{i=1}^{n} (y_i - u_i) \mathbf{x}_{i,v,l_t,t}^{(h-1)} \cdot \left[\mathbf{a}_{v,l_t,t}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,v,l_t,t}^{(l)} \mathbf{W}_{v,l_t,t}^{(l)} M_{v,t}^{(l)} \right) \mathbf{J}_{i,v,l_t,t}^{(h)} M_{v,t}^{(h)} \right]$$

For subnetwork, $\mathbf{G}^{(H)}$ has the same form as in layer drop ResNet.

The accumulated gradients of all the subnetworks:

$$\begin{aligned} \mathcal{W}_{t+1}^{(h)} - \mathcal{W}_t^{(h)} &= \eta \frac{\sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}}}{\sum_{v=1}^{S} M_{v,t}^{(h)}} \\ &= \frac{\eta}{\sum_{v=1}^{S} M_{v,t}^{(h)}} \sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{c_{res}}{H\sqrt{m}} \sum_{i=1}^{n} (y_i - u_i) \mathbf{x}_{i,v,l_t,t}^{(h-1)} \cdot \left[\mathbf{a}_{v,l_t,t}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,v,l_t,t}^{(l)} \mathbf{W}_{v,l_t,t}^{(l)} M_{v,t}^{(l)} \right) \mathbf{J}_{i,v,l_t,t}^{(h)} M_{v,t}^{(h)} \right] \end{aligned}$$

The whole network at global synchronization step t+1

$$\begin{aligned} \mathbf{x}_t^{(1)} &= \sqrt{\frac{c_\sigma}{m}} \sigma \left(\frac{\sum_{v=1}^{S} \mathbf{W}_{v,\ell,t}^{(1)}}{S} \mathbf{x_t} \right), \\ \mathbf{x}_t^{(h)} &= \mathbf{x}_t^{(h-1)} + \frac{c_{res}}{H\sqrt{m}} \sigma \left(\frac{\sum_{v=1}^{S} \mathbf{W}_{v,\ell,t}^{(h)} M_{v,t}^{(h)}}{\sum_{v=1}^{S} M_{v,t}^{(h)}} \mathbf{x}_t^{(h-1)} \right) \\ &\text{for } 2 \leq h \leq H, \\ f_{res}(\mathbf{x}, \theta) &= \frac{\sum_{v=1}^{S} \mathbf{a}_{v,\ell,t}}{S}^{\top} \mathbf{x}_t^{(H)} \end{aligned}$$

Let $\mathcal{W}_t^{(h)} = \frac{\sum_{v=1}^{S} \mathbf{W}_{v,\ell,t}^{(h)} M_{v,t}^{(h)}}{\sum_{v=1}^{S} M_{v,t}^{(h)}}$, $\mathbf{a}_t = \frac{\sum_{v=1}^{S} \mathbf{a}_{v,\ell,t}}{S}$

The whole network at global synchronization step 0

$$\mathbf{x}_0^{(1)} = \sqrt{\frac{c_\sigma}{m}} \sigma \left(\mathbf{W}_0^{(1)} \mathbf{x} \right),$$

$$\mathbf{x}_0^{(h)} = \mathbf{x}_0^{(h-1)} + \frac{c_{res}}{H\sqrt{m}}\sigma\left(\mathbf{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right) \quad \text{for } 2 \le h \le H,$$

$$f_{res}(\mathbf{x}, \theta) = \mathbf{a}_0^\top \mathbf{x}_0^{(H)}$$

B.1 PROOF SKETCH

We can write the loss of the whole network at global syncrhonization step t+1 as

$$L(\theta(t), \mathbf{M}_{1,t}, \mathbf{M}_{2,t}...\mathbf{M}_{S,t}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{u}(t, \mathbf{M}_{1,t}, \mathbf{M}_{2,t}...\mathbf{M}_{S,t})\right\|_2^2.$$

where $\mathbf{M}_{v,t} = \{M_{v,t}^{(1)}, M_{v,t}^{(2)}...M_{v,t}^{(H)}\}$ Let $\mathcal{M}_t = \{\mathbf{M}_{1,t}, \mathbf{M}_{2,t}...\mathbf{M}_{S,t}\}$

For convience, we drop all mask notation in the following proof. Let $\hat{\mathbf{u}}(t)$ be the output of the whole network at global synchronization step $t+1$. Now recall the progress of loss function:

$$\left\|\mathbf{y} - \hat{\mathbf{u}}(t+1)\right\|_2^2 = \left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2 - 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\left(\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right) + \left\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right\|_2^2$$

Following Du et al. [2019], we apply Taylor expansion on $(\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t))$ and look at the ith coordinate.

$$\begin{aligned}\hat{u}_i(t+1) - \hat{u}_i(t) &= -\langle\theta(t+1) - \theta(t), \hat{u}_i'\left(\theta(t)\right)\rangle + \int_{s=0}^{1}\langle\theta(t+1) - \theta(t), \hat{u}_i'\left(\theta(t)\right) - \hat{u}_i'\left(\theta(t) - s(\theta(t) - \theta(t+1))\right)\rangle ds \\ &\triangleq I_1^i(t) + I_2^i(t)\end{aligned}$$

However, it is not obvious that $I_1(t)$ and $I_2(t)$ can be directed bounded to show the decrease of the loss of the whole network as both of them involve the accumulated gradient change from distributed local subnetwork training. Thus, we introduce a new term $I_1'^i(t)$ as below, which relates to the hypothetical global gradient direction as if the whole network trained centrally.

$$\begin{aligned}I_1'^i(t) &= -\eta\ell\langle L'(\theta(t)), \hat{u}_i'\left(\theta(t)\right)\rangle \\ &= -\eta\ell\sum_{j=1}^{n}(\hat{u}_j - y_j)\langle\hat{u}_j'(\theta(t)), \hat{u}_i'\left(\theta(t)\right)\rangle \\ &\triangleq -\eta\ell\sum_{j=1}^{n}(\hat{u}_j - y_j)\sum_{h=1}^{H+1}\hat{\mathbf{G}}_{ij}^{(h)}(t)\end{aligned}$$

Accordingly,

$$\begin{aligned}&\left\|\mathbf{y} - \hat{\mathbf{u}}(t+1)\right\|_2^2 \\ &= \left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2 - 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\left(\mathbf{I}_1(t) + \mathbf{I}_2(t) + \mathbf{I'}_1(t) - \mathbf{I'}_1(t)\right) + \left\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right\|_2^2 \\ &= \left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2 - 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\mathbf{I'}_1(t) + 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\left(\mathbf{I'}_1(t) - \mathbf{I}_1(t)\right) - 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\mathbf{I}_2(t) + \left\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right\|_2^2 \\ &\le \left(1 - \eta\ell\lambda_{\min}\left(\hat{\mathbf{G}}^{(H)}(t)\right)\right)\left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2 + 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\left(\mathbf{I'}_1(t) - \mathbf{I}_1(t)\right) \\ &\quad - 2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\mathbf{I}_2(t) + \left\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right\|_2^2.\end{aligned}$$

Our hypothesis is:

Condition b.1. *At the $t+1$-th global synchronization, for the whole network, we have*

$$\left\|\mathbf{y} - \hat{\mathbf{u}}(t, \mathcal{M}_t)\right\|_2^2 \le (1 - \frac{\eta\ell\lambda_0}{2})^t\left\|\mathbf{y} - \hat{\mathbf{u}}(0)\right\|_2^2.$$

In order to prove this, we need to show $2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\left(\mathbf{I'}_1(t) - \mathbf{I}_1(t)\right)$,$-2\left(\mathbf{y} - \hat{\mathbf{u}}(t)\right)^\top\mathbf{I}_2(t)$ and $\left\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\right\|_2^2$ are proportional to $\eta^2\left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2$ so if we set η sufficiently small, this term is smaller than $\eta\lambda_{\min}\left(\hat{\mathbf{G}}^{(H)}(t)\right)\left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2^2$ and thus the loss function decreases with a linear rate.

Further, similar to Du et al. [2019], to prove the induction hypothesis, it suffices to prove $\lambda_{\min}\left(\hat{\mathbf{G}}^{(H)}(t)\right) \geq \frac{\lambda_0}{2}$ for $t' = 0, \ldots, t$, where λ_0 is independent of m. Similar to Du et al. [2019], we can show at the beginning

$$\lambda_{\min}\left(\hat{\mathbf{G}}^{(H)}(0)\right) \geq \frac{3}{4}\lambda_0.$$

Now for the t-th global iteration, by matrix perturbation analysis, we know it is sufficient to show $\left\|\hat{\mathbf{G}}^{(H)}(t) - \hat{\mathbf{G}}^{(H)}(0)\right\|_2 \leq \frac{1}{4}\lambda_0$. To do this, we show as long as m is large enough, every weight matrix is close its initialization in a relative error sense.

Lemma b.1 (Lemma on Initialization Norms for the whole network)**.** *If $\sigma(\cdot)$ is $L-$Lipschitz and $m = \Omega\left(\frac{n}{\delta}\right)$, assuming $\left\|\mathcal{W}_0^{(h)}\right\|_2 \leq c_{w,0}\sqrt{m}$ for $h \in [2, H]$ and $c_{w,0} \approx 2$ for Gaussian initialization. We have with probability at least $1-\delta$ over random initialization, for every $h \in [H]$ and $i \in [n]$,*

$$\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}_{i,0}^{(h)}\right\|_2 \leq c_{x,0}$$

for some universal constant $c_{x,0} > 1$

Proof of Lemma b.1. As the global model at initialization is the same with original ResNet in Du et al. [2019], we can use the same proof in Lemma C.1 in Du et al. [2019]. □

The following lemma lower bounds $\hat{\mathbf{G}}^{(H)}(0)$'s least eigenvalue.

Lemma b.2 (Least Eigenvalue at the Initialization)**.** *If $m = \Omega\left(\frac{n^2\log(Hn/\delta)}{\lambda_0^2}\right)$, we have*

$$\lambda_{\min}(\hat{\mathbf{G}}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.$$

Proof of Lemma b.2. As the global model at initialization is the same with original ResNet in Du et al. [2019], we can use the same proof in Lemma C.2 in Du et al. [2019]. □

Lemma b.3. *Suppose $\sigma(\cdot)$ is L-Lipschitz and for $h \in [H]$, $\left\|\mathcal{W}_0^{(h)}\right\|_2 \leq c_{w,0}\sqrt{m}$, $\left\|\mathbf{x}_0^{(h)}\right\|_2 \leq c_{x,0}$ and $\left\|\mathbf{W}_{v,l_t,t}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \leq \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0} > 0$ and $R \leq c_{w,0}$. Then we have*

$$\left\|\mathbf{x}_{v,l_t,t}^{(h)} - \mathbf{x}_0^{(h)}\right\|_2 \leq \left(\sqrt{c_\sigma}L + \frac{c_{x,0}}{c_{w,0}} + \frac{c_{x,0}}{R}\right)e^{2c_{res}c_{w,0}L}R \triangleq c_x'R.$$

Proof of Lemma b.3. We prove this lemma by induction. Our induction hypothesis is

$$\left\|\mathbf{x}_{v,l_t,t}^{(h)} - \mathbf{x}_0^{(h)}\right\|_2 \leq g(h),$$

where

$$g(h) = \left[1 + \frac{2c_{res}c_{w,0}L}{H}\right]g(h-1) + \frac{c_{res}Lc_{x,0}}{H}(c_{w,0} + R).$$

For $h = 1$, we have

$$\begin{aligned}\left\|\mathbf{x}_{v,l_t,t}^{(1)} - \mathbf{x}_0^{(1)}\right\|_2 &\leq \sqrt{\frac{c_\sigma}{m}}\left\|\sigma\left(\mathbf{W}_{v,l_t,t}^{(1)}\mathbf{x}\right) - \sigma\left(\mathcal{W}_0^{(1)}\mathbf{x}\right)\right\|_2 \\ &\leq \sqrt{\frac{c_\sigma}{m}}L\left\|\mathbf{W}_{v,l_t,t}^{(1)} - \mathcal{W}_0^{(1)}\right\|_F \leq \sqrt{c_\sigma}LR,\end{aligned}$$

which implies $g(1) = \sqrt{c_\sigma}LR$, for $2 \leq h \leq H$, we have

$$\left\|\mathbf{x}_{v,l_t,t}^{(h)} - \mathbf{x}_0^{(h)}\right\|_2 \leq \frac{c_{res}}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right)M_{v,t}^{(h)} - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)1\right\|_2$$

$$
\begin{aligned}
&+ \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}} \left\|[\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)]M_{v,t}^{(h)} + \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)(M_{v,t}^{(h)} - 1)\right\|_2 \\
&+ \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}} \left\|[\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)]M_{v,t}^{(h)}\right\|_2 + \frac{c_{res}}{H\sqrt{m}} \left\|\sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)(M_{v,t}^{(h)} - 1)\right\|_2 \\
&+ \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}} \left\|[\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)]\right\|_2 \left\|M_{v,t}^{(h)}\right\|_2 + \frac{c_{res}}{H\sqrt{m}} \left\|\sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)\right\|_2 \left\|(M_{v,t}^{(h)} - 1)\right\|_2 \\
&+ \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}} \left\|[\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_0^{(h-1)}\right) + \sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)]\right\|_2 \\
&+ \frac{c_{res}L}{H\sqrt{m}} \left\|\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right\| + \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}} \left\|\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_{v,l_t,t}^{(h-1)}\right) - \sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_0^{(h-1)}\right)\right\|_2 \\
&+ \frac{c_{res}}{H\sqrt{m}} \left\|\sigma\left(\mathbf{W}_{v,l_t,t}^{(h)}\mathbf{x}_0^{(h-1)}\right) - \sigma\left(\mathcal{W}_0^{(h)}\mathbf{x}_0^{(h-1)}\right)\right\|_2 \\
&+ \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 + \frac{c_{res}Lc_{w,0}c_{x,0}}{H} \\
&\le \frac{c_{res}L}{H\sqrt{m}} \left(\left\|\mathcal{W}_0^{(h)}\right\|_2 + \left\|\mathbf{W}_{v,l_t,t}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F\right) \cdot \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 \\
&+ \frac{c_{res}L}{H\sqrt{m}} \left\|\mathbf{W}_{v,l_t,t}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \left\|\mathbf{x}_0^{(h-1)}\right\|_2 + \left\|\mathbf{x}_{v,l_t,t}^{(h-1)} - \mathbf{x}_0^{(h-1)}\right\|_2 + \frac{c_{res}Lc_{w,0}c_{x,0}}{H} \\
&\le \left[1 + \frac{c_{res}L}{H\sqrt{m}}\left(c_{w,0}\sqrt{m} + R\sqrt{m}\right)\right] g(h-1) + \frac{c_{res}L}{H\sqrt{m}}\sqrt{m}Rc_{x,0} + \frac{c_{res}Lc_{w,0}c_{x,0}}{H} \\
&\le \left(1 + \frac{2c_{res}c_{w,0}L}{H}\right) g(h-1) + \frac{c_{res}}{H}Lc_{x,0}R + \frac{c_{res}Lc_{w,0}c_{x,0}}{H}.
\end{aligned}
$$

Lastly, simple calculations show $g(h) \le \left(\sqrt{c_\sigma}L + \frac{c_{x,0}}{c_{w,0}} + \frac{c_{x,0}}{R}\right) e^{2c_{res}c_{w,0}L}R$.

□

Lemma b.4. *Suppose $\sigma(\cdot)$ is L-Lipschitz and for $h \in [H]$, $\left\|\mathcal{W}_0^{(h)}\right\|_2 \le c_{w,0}\sqrt{m}$, $\left\|\mathbf{x}_0^{(h)}\right\|_2 \le c_{x,0}$ and $\left\|\mathcal{W}_t^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \le \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0} > 0$ and $R \le c_{w,0}$. Then we have*

$$
\left\|\mathbf{x}_t^{(h)} - \mathbf{x}_0^{(h)}\right\|_2 \le \left(\sqrt{c_\sigma}L + \frac{c_{x,0}}{c_{w,0}}\right) e^{2c_{res}c_{w,0}L}R \triangleq c_xR.
$$

Proof of Lemma b.4. The proof is exactly the same with proof of C.3 in Du et al. [2019] □

Next, we characterize how the perturbation on the weight matrices affect $\hat{\mathbf{G}}^{(H)}$.

Lemma b.5. *Suppose $\sigma(\cdot)$ is differentiable, $L-$Lipschitz and $\beta-$smooth. Suppose for $h \in [H]$, $\left\|\mathcal{W}_0^{(h)}\right\|_2 \le c_{w,0}\sqrt{m}$, $\|\mathbf{a}_0\|_2 \le a_{2,0}\sqrt{m}$, $\|\mathbf{a}_0\|_4 \le a_{4,0}m^{1/4}$, $\frac{1}{c_{x,0}} \le \left\|\mathbf{x}_0^{(h)}\right\|_2 \le c_{x,0}$, if $\left\|\mathcal{W}_r^{(h)} - \mathcal{W}_0^{(h)}\right\|_F, \|\mathbf{a}_r - \mathbf{a}_0\|_2 \le \sqrt{m}R$ where $R \le c\lambda_0H^2n^{-1}$ and $R \le c$ for some small constant c , we have*

$$
\left\|\hat{\mathbf{G}}^{(H)}(t) - \hat{\mathbf{G}}^{(H)}(0)\right\|_2 \le \frac{\lambda_0}{2}.
$$

Proof of Lemma b.5. Similar to C.4 in Du et al. [2019] Because Frobenius-norm of a matrix is bigger than the operator norm, it is sufficient to bound $\left\|\hat{\mathbf{G}}^{(H)}(t)-\hat{\mathbf{G}}^{(H)}(0)\right\|_F$. For simplicity define $z_{i,q}(t)=\mathcal{W}_{t,q}^{(H)\top}\mathbf{x}_{i,t}^{(H-1)}$, we have

$$
\begin{aligned}
&\left|\hat{\mathbf{G}}_{i,j}^{(H)}(t)-\hat{\mathbf{G}}_{i,j}^{(H)}(0)\right| \\
=&\frac{c_{res}^2}{H^2m}\big|\mathbf{x}_{i,t}^{(H-1)\top}\mathbf{x}_{j,t}^{(H-1)}\sum_{q=1}^{m}a_q(t)^2\sigma'\left(z_{i,q}(t)\right)\sigma'\left(z_{j,q}(t)\right) \\
&-\mathbf{x}_{i,0}^{(H-1)\top}\mathbf{x}_{j,0}^{(H-1)}\sum_{q=1}^{m}a_q(0)^2\sigma'\left(z_{i,q}(0)\right)\sigma'\left(z_{j,q}(0)\right)\big| \\
\leq&\frac{c_{res}^2}{H^2}L^2a_{2,0}^2\left|\mathbf{x}_{i,t}^{(H-1)\top}\mathbf{x}_{j,t}^{(H-1)}-\mathbf{x}_{i,0}^{(H-1)\top}\mathbf{x}_{j,0}^{(H-1)}\right| \\
&+\frac{c_{res}^2}{H^2}\frac{c_{x,0}^2}{m}\left|\sum_{q=1}^{m}a_q(0)^2\left(\sigma'\left(z_{i,q}(t)\right)\sigma'\left(z_{j,q}(t)\right)-\sigma'\left(z_{i,q}(0)\right)\sigma'\left(z_{j,q}(0)\right)\right)\right| \\
&+\frac{c_{res}^2}{H^2m}\left|\mathbf{x}_{i,t}^{(H-1)\top}\mathbf{x}_{j,t}^{(H-1)}\right|\left|\sum_{q=1}^{m}\left(a_q(t)^2-a_q(0)^2\right)\sigma'\left(z_{i,q}(t)\right)\sigma'\left(z_{j,q}(t)\right)\right| \\
\triangleq&\frac{c_{res}^2}{H^2}(I_1^{i,j}+I_2^{i,j}+I_3^{i,j}).
\end{aligned}
$$

For $I_1^{i,j}$, using Lemma b.4, we have

$$
\begin{aligned}
I_1^{i,j}=&L^2a_{2,0}^2\left|\mathbf{x}_{i,t}^{(H-1)\top}\mathbf{x}_{j,t}^{(H-1)}-\mathbf{x}_{i,0}^{(H-1)\top}\mathbf{x}_{j,0}^{(H-1)}\right| \\
\leq&L^2a_{2,0}^2\left|(\mathbf{x}_{i,t}^{(H-1)}-\mathbf{x}_{i,0}^{(H-1)}\top)\mathbf{x}_{j,t}^{(H-1)}\right|+L^2a_{2,0}^2\left|\mathbf{x}_{i,0}^{(H-1)\top}(\mathbf{x}_{i,t}^{(H-1)}-\mathbf{x}_{i,0}^{(H-1)})\right| \\
\leq&c_xL^2a_{2,0}^2R\cdot(c_{x,0}+c_xR)+c_{x,0}c_xL^2a_{2,0}^2R \\
\leq&3c_{x,0}c_xL^2a_{2,0}^2R,
\end{aligned}
$$

Same with C.4 Du et al. [2019], to bound $I_2^{i,j}$, we have

$$
\begin{aligned}
I_2^{i,j}=&c_{x,0}^2\frac{1}{m}\left|\sum_{q=1}^{m}a_q(0)^2\sigma'\left(z_{i,q}(t)\right)\sigma'\left(z_{j,q}(t)\right)-a_q(0)^2\sigma'\left(z_{i,q}(0)\right)\sigma'\left(z_{j,q}(0)\right)\right| \\
\leq&\frac{\beta La_{4,0}^2c_{x,0}^2}{\sqrt{m}}\left(\sqrt{\sum_{q=1}^{m}\left|z_{i,q}(t)-z_{i,q}(0)\right|^2}+\sqrt{\sum_{q=1}^{m}\left|z_{j,q}(t)-z_{j,q}(0)\right|^2}\right).
\end{aligned}
$$

Using the same proof for Lemma b.4, it is easy to see

$$
\sum_{q=1}^{m}\left|z_{i,q}(t)-z_{i,q}(0)\right|^2\leq\left(2c_xc_{w,0}+c_{x,0}\right)^2L^2mR^2.
$$

Thus

$$
I_2^{i,j}\leq2\beta c_{x,0}^2\left(2c_xc_{w,0}+c_{x,0}\right)L^2R.
$$

The bound of $I_3^{i,j}$ is the same to that $I_3^{i,j}$ in Du et al. [2019] C.4,

$$
I_3^{i,j}\leq12L^2c_{x,0}^2a_{2,0}R.
$$

Therefore we can bound the perturbation

$$
\left\|\hat{\mathbf{G}}^{(H)}(t)-\hat{\mathbf{G}}^{(H)}(0)\right\|_F=\sqrt{\sum_{(i,j)}^{n,n}\left|\hat{\mathbf{G}}_{i,j}^{(H)}(t)-\hat{\mathbf{G}}_{i,j}^{(H)}(0)\right|^2}
$$

$$\leq \frac{c_{res}^2}{H^2}\sqrt{n^2(3c_{x,0}c_xL^2a_{2,0}^2R + 2\beta c_{x,0}^2\left(2c_xc_{w,0}+c_{x,0}\right)L^2R + 12L^2c_{x,0}^2a_{2,0}R)}$$
$$= \frac{c_{res}^2}{H^2}n(3c_{x,0}c_xL^2a_{2,0}^2R + 2\beta c_{x,0}^2\left(2c_xc_{w,0}+c_{x,0}\right)L^2R + 12L^2c_{x,0}^2a_{2,0}R)$$

.

Plugging in the bound on R, we have the desired result. □

Now we prove theorem **??** by induction, assume the condition b.1, we want to bound the change of weight to satisfy lemma b.5 and then we want to show $2\left(\mathbf{y}-\hat{\mathbf{u}}(t)\right)^\top\left(\mathbf{I}'_1(t)-\mathbf{I}_1(t)\right)$,$-2\left(\mathbf{y}-\hat{\mathbf{u}}(t)\right)^\top\mathbf{I}_2(t)$ and $\|\hat{\mathbf{u}}(t+1)-\hat{\mathbf{u}}(t)\|_2^2$ are proportional to $\eta^2\|\mathbf{y}-\hat{\mathbf{u}}(t)\|_2^2$ so if we set η sufficiently small, this term is smaller than $\eta\lambda_{\min}\left(\hat{\mathbf{G}}^{(H)}(t)\right)\|\mathbf{y}-\hat{\mathbf{u}}(t)\|_2^2$ and thus the loss function decreases with a linear rate.

Lemma b.6. *If Condition b.1 holds for $t'=0,\ldots,t-1$, we have for any $1\leq v\leq S$, $0\leq l_t\leq\ell$*

$$\left\|\mathbf{W}_{v,l_t,t}^{(h)}-\mathcal{W}_0^{(h)}\right\|_F, \|\mathbf{a}_{v,l_t,t}-\mathbf{a}_0\|_2 \leq R'\sqrt{m},$$
$$\left\|\mathbf{W}_{v,l_t,t}^{(h)}-\mathbf{W}_{v,l_t-1,t}^{(h)}\right\|_F, \|\mathbf{a}_{v,l_t,t}-\mathbf{a}_{v,l_t-1,t}\|_2 \leq \eta Q'(l_t-1,t),$$

where $R'=\frac{16c_{res}c_{x,0}a_{2,0}Le^{2c_{res}c_{w,0}L}\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{H\lambda_0\sqrt{m}} < c$ for some small constant c,
$Q'(l_t,t)=4c_{res}c_{x,0}a_{2,0}Le^{2c_{res}c_{w,0}L}\sqrt{n}\|\mathbf{y}-\mathbf{u}_{t,l_t}\|_2/H$ and
$Q'(t)=4c_{res}c_{x,0}a_{2,0}Le^{2c_{res}c_{w,0}L}\sqrt{n}\|\mathbf{y}-\hat{\mathbf{u}}_t\|_2/H$.

Proof of Lemma b.6. We will prove this corollary by induction. The induction hypothesis is

$$\left\|\mathbf{W}_{v,l_t,t}^{(h)}-\mathcal{W}_0^{(h)}\right\|_F \leq R'\sqrt{m}$$
$$\|\mathbf{a}_{v,l_t,t}-\mathbf{a}_0\|_2 \leq R'\sqrt{m}.$$

First we want to prove it holds for $t'=0$ and $0\leq l_t\leq\ell$.

We prove it by induction w.r.t l_t: It is easy to see that it holds for $t'=0$ and $l'_t=0$. Suppose it holds for $0\leq l'_t\leq l_t$, we want to prove it holds for $l'_t=l_t+1$ Following C.5 in Du et al. [2019], note $\left\|\mathbf{J}_{i,v,l_t,t}^{(k)}\right\|_2\leq L$. We have

$$\begin{aligned}
&\left\|\mathbf{W}_{v,l_t+1,t}^{(h)}-\mathbf{W}_{v,l_t,t}^{(h)}\right\|_F \\
\leq&\eta\frac{c_{res}}{H\sqrt{m}}\|\mathbf{a}_{\mathbf{v},\mathbf{l_t},\mathbf{t}}\|_2\sum_{i=1}^n|y_i-u_{i,v,l_t,t}|\left\|\mathbf{x}_{i,v,l_t,t}^{(h-1)}\right\|_2\prod_{k=h+1}^H\left\|\mathbf{I}+\frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(k)}\mathbf{W}_{v,l_t,t}^{(k)}M_{v,t}^{(k)}\right\|_2\left\|\mathbf{J}_{i,v,l_t,t}^{(k)}\right\|_2\left\|M_{v,t}^{(h)}\right\|_2 \\
\leq&\eta\frac{Lc_{res}}{H\sqrt{m}}\|\mathbf{a}_{\mathbf{v},\mathbf{l_t},\mathbf{t}}\|_2\sum_{i=1}^n|y_i-u_{i,v,l_t,t}|\left\|\mathbf{x}_{i,v,l_t,t}^{(h-1)}\right\|_2\prod_{k=h+1}^H\left\|\mathbf{I}+\frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(k)}\mathbf{W}_{v,l_t,t}^{(k)}M_{v,t}^{(k)}\right\|_2
\end{aligned}$$

Further

$$\begin{aligned}
&\prod_{k=h+1}^H\left\|\mathbf{I}+\frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(k)}\mathbf{W}_{v,l_t,t}^{(k)}M_{v,t}^{(k)}\right\|_2 \\
\leq&\prod_{k=h+1}^H\|\mathbf{I}\|_2+\left\|\frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(k)}\mathbf{W}_{v,l_t,t}^{(k)}M_{v,t}^{(k)}\right\|_2 \\
\leq&\prod_{k=h+1}^H\|\mathbf{I}\|_2+\frac{c_{res}}{H\sqrt{m}}\left\|\mathbf{J}_{i,v,l_t,t}^{(k)}\right\|_2\left\|\mathbf{W}_{v,l_t,t}^{(k)}\right\|_2\left\|M_{v,t}^{(k)}\right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq \prod_{k=h+1}^{H} \|\mathbf{I}\|_2 + \frac{c_{res}L}{H\sqrt{m}}(\left\|\mathcal{W}_0^{(k)}\right\|_F + \left\|\mathbf{W}_{v,l_t,t}^{(k)} - \mathcal{W}_0^{(k)}\right\|_F) \\
&\leq \prod_{k=h+1}^{H} 1 + \frac{c_{res}L}{H}(c_{w,0} + R') \\
&\leq \prod_{k=h+1}^{H} 1 + \frac{c_{res}L}{H} 2c_{w,0} \\
&\leq e^{2c_{res}x_{w,0}L}
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\left\|\mathbf{W}_{v,l_t+1,t}^{(h)} - \mathbf{W}_{v,l_t,t}^{(h)}\right\|_F \\
\leq &\eta\frac{Lc_{res}}{H\sqrt{m}} \|\mathbf{a}_{v,l_t,t}\|_2 \sum_{i=1}^{n} |y_i - u_i(s)| \left\|\mathbf{x}_{i,v,l_t,t}^{(h-1)}\right\|_2 \prod_{k=h+1}^{H} \left\|\mathbf{I} + \frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(k)}\mathbf{W}_{v,l_t,t}^{(k)}M_{v,t}^{(k)}\right\|_2 \\
\leq &\eta\frac{Lc_{res}}{H\sqrt{m}} \|\mathbf{a}_{v,l_t,t}\|_2 \sum_{i=1}^{n} |y_i - u_i(s)| \left\|\mathbf{x}_{i,v,l_t,t}^{(h-1)}\right\|_2 e^{2c_{res}x_{w,0}L} \\
\leq &\eta c_{res}(c_{x,0} + c_x R')La_{2,0}e^{2c_{res}c_{w,0}L}\sqrt{n} \|\mathbf{y} - \mathbf{u}_{l_t,t}\|_2 / H \\
\leq &3\eta c_{res}c_{x,0}La_{2,0}e^{2c_{res}c_{w,0}L}\sqrt{n} \|\mathbf{y} - \mathbf{u}_{l_t,t}\|_2 / H \\
\leq &\eta Q'(l_t, t) \\
\leq &(1 - \frac{\eta\lambda_0}{2})^{s/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\|\mathbf{a}_{v,l_t+1,t} - \mathbf{a}_{v,l_t,t}\|_2 \leq &3\eta c_{x,0} \sum_{i=1}^{n} |y_i - u_{l_t,t}| \\
\leq &\eta Q'(l_t, t) \\
\leq &(1 - \frac{\eta\lambda_0}{2})^{l_t/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\left\|\mathbf{W}_{v,l_t+1,t}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \\
\leq &\left\|\mathbf{W}_{v,l_t+1,t}^{(h)} - \mathcal{W}_{v,l_t,t}^{(h)}\right\|_F + \left\|\mathbf{W}_{v,l_t,t}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \\
\leq &\sum_{l_t'=0}^{l_t} \eta(1 - \frac{\eta\lambda_0}{2})^{l_t'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\|\mathbf{a}_{v,l_t+1,t} - \mathbf{a}_0\|_2 \\
\leq &\sum_{l_t'=0}^{l_t} \eta(1 - \frac{\eta\lambda_0}{2})^{l_t'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}
$$

Now suppose the hypothesis hold for t'=0,1..,t and for $0 \leq l_t \leq \ell$. We want to prove for $t' = t + 1$, the hypothesis holds. By Lemma b.7, we know $\left\|\mathcal{W}_t^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \leq \sqrt{m}R'$ Thus, $\left\|\mathbf{W}_{v,l_t=0,t+1}^{(h)} - \mathcal{W}_0^{(h)}\right\|_F \leq \sqrt{m}R'$ Thus, by using the same induction on l_t above, we can prove the hypothesis for $t + 1$.

□

Lemma b.7. *Assume*

$$\left\|\mathbf{W}_{v,l_t,t}^{(h)}-\mathcal{W}_0^{(h)}\right\|_F, \|\mathbf{a}_{v,l_t,t}-\mathbf{a}_0\|_2 \leq \sqrt{m}R'$$

We have

$$\left\|\mathcal{W}_t^{(h)}-\mathcal{W}_0^{(h)}\right\|_F, \|\mathbf{a}_t-\mathbf{a}_0\|_2 \leq \sqrt{m}R'$$

Proof of Lemma b.7.

$$\begin{aligned}\left\|\mathcal{W}_t^{(h)}-\mathcal{W}_0^{(h)}\right\|_F &= \left\|\frac{\sum_{v=1}^S \mathbf{W}_{v,\ell,t}^{(h)} M_{v,t}^{(h)}}{\sum_{v=1}^S M_{v,t}^{(h)}}-\mathcal{W}_0^{(h)}\right\|_F \\ &\leq \frac{\sum_{v:M_{v,t}^{(h)}=1}\left\|\mathbf{W}_{v,\ell,t}^{(h)}-\mathcal{W}_0^{(h)}\right\|_F}{\sum_{v=1}^S M_{v,t}^{(h)}} \\ &\leq \sqrt{m}R'\end{aligned}$$

Similarly,

$$\begin{aligned}\|\mathbf{a}_t-\mathbf{a}_0\|_2 &\leq \frac{\sum_{v=1}^S \|\mathbf{a}_{v,l_t,t}-\mathbf{a}_0\|_2}{S} \\ &\leq \sqrt{m}R'\end{aligned}$$

□

Lemma b.8. *If Condition b.1 holds for $t'=0,\ldots,t-1$ and $\eta \leq c\lambda_0 H^2 n^{-2}\ell^{-2}S^{-1}$ for some small constant c, we have $\left\|\mathbf{I'}_1^i(t)-\mathbf{I}_1^i(t)\right\|_2 \leq C_{I_1}^*\eta^2\|y_i-\hat{u}_{i,t-1}\|_2$ where $C_{I_1}^*$ is a constant and thus $\|\mathbf{I'}_1(t)-\mathbf{I}_1(t)\|_2 \leq \frac{1}{16}\eta\lambda_0\|\mathbf{y}-\hat{\mathbf{u}}(k)\|_2$.*

Proof of Lemma b.8.

$$\begin{aligned}\left\|\mathbf{I'}_1^i(t)-\mathbf{I}_1^i(t)\right\|_2 &= \|\langle \eta\ell L'(\theta(t))-(\theta(t+1)-\theta(t)), \hat{u}_i'(\theta(t))\rangle\|_2 \\ &\leq \sum_{h=1}^H \left\|\eta\frac{\sum_{v=1}^S\sum_{l_t=1}^\ell \frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}}}{\sum_{v=1}^S M_{v,t}^{(h)}}-\eta\ell\frac{\partial L}{\partial \mathcal{W}_{t-1}^{(h)}}\right\|_F \|\hat{u}_i'(\theta(t)\|_2 \\ &+ \left\|\eta\frac{\sum_{v=1}^S\sum_{l_t=1}^\ell \frac{\partial L}{\partial \mathbf{a}_{v,l_t,t}}}{S}-\eta\ell\frac{\partial L}{\partial \mathbf{a}_t}\right\|_2 \|\hat{u}_i'(\theta(t)\|_2\end{aligned}$$

Let $M_{t,h}=\sum_{v=1}^S M_{v,t}^{(h)}$

$$\begin{aligned}&\left\|\eta\frac{\sum_{v=1}^S\sum_{l_t=1}^\ell \frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}}}{\sum_{v=1}^S M_{v,t}^{(h)}}-\eta\ell\frac{\partial L}{\partial \mathcal{W}_{t-1}^{(h)}}\right\|_F \\ \leq&\eta 1/M_{t,h}\sum_{v=1}^S\sum_{l_t=1}^\ell\left\|\frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}}-\frac{\partial L}{\partial \mathcal{W}_{t-1}^{(h)}}\right\|_F \\ \leq&\eta 1/M_{t,h}\sum_{v=1}^S\sum_{l_t=1}^\ell||\frac{c_{res}}{H\sqrt{m}}\sum_{i=1}^n(y_i-u_{i,v,l_t,t})\mathbf{x}_{i,v,l_t,t}^{(h-1)}\cdot\left[\mathbf{a}_{v,l_t,t}^\top\prod_{l=h+1}^H\left(\mathbf{I}+\frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,v,l_t,t}^{(l)}\mathbf{W}_{v,l_t,t}^{(l)}M_{v,t}^{(l)}\right)\mathbf{J}_{i,v,l_t,t}^{(h)}M_{v,t}^{(h)}\right]\end{aligned}$$

$$
\begin{aligned}
&- \frac{c_{res}}{H\sqrt{m}} \sum_{i=1}^{n} (y_i - \hat{u}_{i,t-1}) \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,t-1}^{(l)} \mathcal{W}_{t-1}^{(l)} \right) \mathbf{J}_{i,t-1}^{(h)} \right] ||_F \\
\leq & \eta 1/M_{t,h} \sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{c_{res}}{H\sqrt{m}} \sum_{i=1}^{n} ||(y_i - u_{i,v,l_t,t}) \mathbf{x}_{i,v,l_t,t}^{(h-1)} \cdot \left[\mathbf{a}_{v,l_t,t}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,v,l_t,t}^{(l)} \mathbf{W}_{v,l_t,t}^{(l)} M_{v,t}^{(l)} \right) \mathbf{J}_{i,v,l_t,t}^{(h)} M_{v,t}^{(h)} \right] \\
&- (y_i - \hat{u}_{i,t-1}) \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,t-1}^{(l)} \mathcal{W}_{t-1}^{(l)} \right) \mathbf{J}_{i,t-1}^{(h)} \right] ||_F
\end{aligned}
$$

Through standard calculations, we have

$$
\begin{aligned}
\left\| \mathcal{W}_{t-1}^{(l)} - \mathbf{W}_{v,l_t,t}^{(l)} \right\|_F \leq & \eta \ell Q'(0,t), \\
\left\| \mathbf{a}_{t-1} - \mathbf{a}_{v,l_t,t} \right\|_F \leq & \eta \ell Q'(0,t), \\
\left\| \mathbf{x}_{i,t-1}^{(h-1)} - \mathbf{x}_{i,v,l_t,t}^{(h-1)} \right\|_F \leq & \eta \ell c_x' \frac{Q'(0,t)}{\sqrt{m}}, \\
\left\| \mathbf{J}_{i,t-1}^{(l)} - \mathbf{J}_{i,v,l_t,t}^{(l)} \right\|_F \leq & 2\ell \left(c_{x,0} + c_{w,0} c_x' \right) \eta \beta Q'(0,t),
\end{aligned}
$$

where $c_x' \triangleq \left(\sqrt{c_\sigma} L + \frac{c_{x,0}}{c_{w,0}} + \frac{c_{x,0}}{R} \right) e^{2 c_{res} c_{w,0} L}$. As we know $\|y_i - u_{i,v,l_t,t}\| \leq \|y_i - \hat{u}_{i,t-1}\|$, suppose $\|u_{i,v,l_t,t} - \hat{u}_{i,t-1}\| \leq C_u$

According to Lemma G.1 in Du et al. [2019], we have

$$
\begin{aligned}
&\eta 1/M_{t,h} \sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{c_{res}}{H\sqrt{m}} \sum_{i=1}^{n} ||(y_i - u_{i,v,l_t,t}) \mathbf{x}_{i,v,l_t,t}^{(h-1)} \cdot \left[\mathbf{a}_{v,l_t,t}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,v,l_t,t}^{(l)} \mathbf{W}_{v,l_t,t}^{(l)} M_{v,t}^{(l)} \right) \mathbf{J}_{i,v,l_t,t}^{(h)} M_{v,t}^{(h)} \right] \\
&- (y_i - \hat{u}_{i,t-1}) \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,t-1}^{(l)} \mathcal{W}_{t-1}^{(l)} \right) \mathbf{J}_{i,t-1}^{(h)} \right] ||_F \\
\leq & \eta 1/M_{t,h} S \ell n \frac{4}{H} c_{res} c_{x,0} L a_{2,0} e^{2 L c_{w,0}} (C_u \\
&+ \eta \ell \frac{Q'(0,t)}{\sqrt{m}} \left(\frac{c_x}{c_{x,0}} + \frac{2}{L} \left(c_{x,0} + c_{w,0} c_x \right) \beta \sqrt{m} + 4 c_{w,0} \left(c_{x,0} + c_{w,0} c_x \right) \beta + L + 1 \right)) \left\| y_i - \hat{u}_{i,t-1} \right\|_2
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\left\| \eta \frac{\sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{\partial L}{\partial \mathbf{a}_{v,l_t,t}}}{S} - \eta \ell \frac{\partial L}{\partial \mathbf{a}_t} \right\|_2 \leq & \eta 1/S \sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \left\| \frac{\partial L}{\partial \mathbf{a}_{v,l_t,t}} - \frac{\partial L}{\partial \mathbf{a}_t} \right\|_2 \\
\leq & \eta 1/S \sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \sum_{i=1}^{n} \left\| (y_i - u_{i,v,l_t,t}) \mathbf{x}_{i,v,l_t,t}^{(H)} - (y_i - \hat{u}_{i,t}) \mathbf{x}_{i,t-1}^{(H)} \right\|_2 \\
\leq & \eta \ell n (C_u + \eta \ell c_x' \frac{Q'(0,t)}{\sqrt{m}}) \left\| y_i - \hat{u}_{i,t-1} \right\|_2
\end{aligned}
$$

Also,

$$
\begin{aligned}
\left\| \hat{u}_i' \left(\theta(t) \right) \right\|_2 \leq & \frac{c_{res}}{H\sqrt{m}} \sum_{h=1}^{H} \left\| \frac{\partial \hat{u}_i \left(\theta(t) \right)}{\partial \mathcal{W}_{t-1}^{(h)}} \right\|_2 \\
= & \frac{c_{res}}{H\sqrt{m}} \sum_{h=1}^{H} \left\| \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^{\top} \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,t-1}^{(l)} \mathcal{W}_{t-1}^{(l)} \right) \mathbf{J}_{i,t-1}^{(h)} \right] \right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{c_{res}}{H\sqrt{m}} \sum_{h=1}^{H} \left\| \mathbf{x}_{i,t-1}^{(h-1)} \right\|_2 \left\| \mathbf{a}_{t-1} \right\|_2 \left\| \prod_{l=h+1}^{H} \left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}} \mathbf{J}_{i,t-1}^{(l)} \mathcal{W}_{t-1}^{(l)} \right) \right\|_2 \left\| \mathbf{J}_{i,t-1}^{(h)} \right\|_2 \\
&\leq \frac{c_{res}}{H} H 2 c_{x,0} a_{2,0} L e^{2 c_{res} x_{w,0} L} \\
&= 2 c_{res} c_{x,0} a_{2,0} L e^{2 c_{res} x_{w,0} L}
\end{aligned}
$$

Thus, combine all above and also according to Lemma b.9

$$
\begin{aligned}
\left\| \mathbf{I'}_1^i(t) - \mathbf{I}_1^i(t) \right\|_2 \leq & \sum_{h=1}^{H} \left\| \eta \frac{\sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{\partial L}{\partial \mathbf{W}_{v,l_t,t}^{(h)}}}{\sum_{v=1}^{S} M_{v,t}^{(h)}} - \eta \ell \frac{\partial L}{\partial \mathcal{W}_{t-1}^{(h)}} \right\|_2 \left\| \hat{u}_i' \left(\theta(t) \right) \right\|_2 \\
& + \left\| \eta \frac{\sum_{v=1}^{S} \sum_{l_t=1}^{\ell} \frac{\partial L}{\partial \mathbf{a}_{v,l_t,t}}}{S} - \eta \ell \frac{\partial L}{\partial \mathbf{a}_t} \right\|_2 \left\| \hat{u}_i' \left(\theta(t) \right) \right\|_2 \\
& \leq C_{I_1}^* \eta^2 \left\| y_i - \hat{u}_{i,t-1} \right\| \text{ where } C_{I_1}^* \text{ is a constant}
\end{aligned}
$$

Using the bound on η and following Du et al. [2019] $\left\| \mathbf{y} - \hat{\mathbf{u}} \right\|_2 = O(\sqrt{n})$,

$$
\left\| \mathbf{I'}_1(t) - \mathbf{I}_1(t) \right\| \leq \frac{1}{16} \eta \lambda_0 \left\| \mathbf{y} - \hat{\mathbf{u}}(k) \right\|_2
$$

□

Lemma b.9.

$$
\left\| u_{i,v,l_t,t} - \hat{u}_{i,t} \right\|_2 \leq \eta \ell Q'(0,t) B \textit{ where B is a constant}
$$

Proof of Lemma b.9.

$$
\begin{aligned}
\left\| u_{i,v,l_t,t} - \hat{u}_{i,t} \right\|_2 &= \left\| \mathbf{a}_{v,l_t,t}^{\top} \mathbf{x}_{v,l_t,t}^{(H)} - \mathbf{a}_t^{\top} \mathbf{x}_t^{(H)} \right\|_2 \\
&\leq \eta (2 a_{2,0} 3 c_{x,0} \ell Q'(0,t) (1 + \frac{c_x}{\sqrt{m}}))
\end{aligned}
$$

□

Lemma b.10. *If Condition b.1 holds for $t' = 0, \ldots, t-1$ and $\eta \leq c \lambda_0 H^2 n^{-2} \ell^{-2} S^{-1}$ for some small constant c, we have $\left\| \mathbf{I}_2(t) \right\|_2 \leq C_{I_2}^* \eta^2 \left\| y_i - \hat{u}_{i,t-1} \right\|_2$ where $C_{I_2}^*$ is a constant and thus $\left\| \mathbf{I}_2(t) \right\|_2 \leq \frac{1}{8} \eta \lambda_0 \left\| \mathbf{y} - \hat{\mathbf{u}}(k) \right\|_2$.*

Proof of Leamma b.10.

$$
I_2^i(t) = \int_{s=0}^{1} \langle \theta(t+1) - \theta(t), \hat{u}_i' \left(\theta(t) \right) - \hat{u}_i' \left(\theta(t) - s(\theta(t) - \theta(t+1)) \right) \rangle ds
$$

Define for $1 \leq h \leq H$

$$
\hat{u}_i'^{(h)} \left(\theta(t) \right) = \frac{\partial \hat{u}(\theta(t))}{\mathcal{W}_t^{(h)}}
$$

And

$$
\hat{u}_i'^{(H+1)} \left(\theta(t) \right) = \frac{\partial \hat{u}(\theta(t))}{\mathbf{a}_t}
$$

$$
\left| I_2^i(t) \right| \leq \max_{0 \leq s \leq 1} \sum_{h=1}^{H} \left\| \mathcal{W}_t^{(h)} - \mathcal{W}_{t-1}^{(h)} \right\|_F \left\| \hat{u}_i'^{(h)} \left(\theta(t) \right) - \hat{u}_i'^{(h)} \left(\theta(t) - s(\theta(t+1) - \theta(t)) \right) \right\|_F
$$

$$+ \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2 \left\|\hat{u}_i'^{(H+1)}(\theta(t)) - \hat{u}_i'^{(H+1)}(\theta(t) - s(\theta(t+1) - \theta(t)))\right\|_2.$$

From Lemma b.8 and Lemma b.6,

$$\left\|\mathcal{W}_t^{(h)} - \mathcal{W}_{t-1}^{(h)}\right\|_F \le \eta\ell\hat{Q}'(t-1)$$
$$\|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2 \le \eta\ell\hat{Q}'(t-1)$$

Let $\mathbf{x}_{i,t-1,s}^{(l)}$ be the activation of global network with $\mathcal{W}_{t-1,s} = \mathcal{W}_{t-1} - s(\mathcal{W}_{t-1} - \mathcal{W}_t)$. We similarly define $\mathbf{J}_{i,t-1,s}^{(l)}$ and $\mathbf{a}_{t-1,s}$

$$\begin{aligned}
&\left\|\hat{u}_i'^{(h)}(\theta(t)) - \hat{u}_i'^{(h)}(\theta(t) - s(\mathcal{W}_{t-1} - \mathcal{W}_t))\right\|_F \\
\le&\frac{c_{res}}{H\sqrt{m}}||\mathbf{x}_{i,t-1,s}^{(h-1)} \cdot \left[\mathbf{a}_{t-1,s}^\top \prod_{l=h+1}^{H}\left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,t-1,s}^{(l)}\mathcal{W}_{t-1,s}^{(l)}\right)\mathbf{J}_{i,t-1,s}^{(h)}\right] \\
&- \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^\top \prod_{l=h+1}^{H}\left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,t-1}^{(l)}\mathcal{W}_{t-1}^{(l)}\right)\mathbf{J}_{i,t-1}^{(h)}\right]||_F
\end{aligned}$$

Through similar calculation in Lemma b.8,

$$\begin{aligned}
\left\|\mathcal{W}_{t-1,s}^{(l)} - \mathcal{W}_{t-1}^{(l)}\right\|_F =& s\left\|(\mathcal{W}_{t-1}^{(l)} - \mathcal{W}_t^{(l)})\right\|_F \\
\le& \left\|(\mathcal{W}_{t-1}^{(l)} - \mathcal{W}_t^{(l)})\right\|_F \\
\le& \eta\ell\hat{Q}'(t-1)
\end{aligned}$$

$$\begin{aligned}
\left\|\mathbf{x}_{i,t-1,s}^{(l)} - \mathbf{x}_{i,t-1}^{(l)}\right\|_2 &\le \frac{c_{res}}{H\sqrt{m}}\left\|\sigma\left(\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1,s}^{(l-1)}\right) - \sigma\left(\mathcal{W}_{t-1}^{(l)}\mathbf{x}_{t-1}^{(l-1)}\right)\right\|_2 + \left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}}\left\|\sigma\left(\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1,s}^{(l-1)}\right) - \sigma\left(\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1}^{(l-1)}\right)\right\|_2 \\
&+ \frac{c_{res}}{H\sqrt{m}}\left\|\sigma\left(\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1}^{(l-1)}\right) - \sigma\left(\mathcal{W}_{t-1}^{(l)}\mathbf{x}_{t-1}^{(l-1)}\right)\right\|_2 + \left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&\le \frac{c_{res}L}{H\sqrt{m}}\left\|\mathcal{W}_{t-1,s}^{(l)}\right\|_F\left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&+ \frac{c_{res}L}{H\sqrt{m}}\left\|\mathcal{W}_{t-1,s}^{(l)} - \mathcal{W}_{t-1}^{(l)}\right\|_F\left\|\mathbf{x}_{t-1}^{(l-1)}\right\|_2 + \left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&\le (1 + \frac{c_{res}L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R'\sqrt{m} + \eta\ell\hat{Q}'(t-1)))\left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&+ \frac{c_{res}L}{H\sqrt{m}}\eta\ell\hat{Q}'(t-1)(c_xR' + c_{x,0})
\end{aligned}$$

Also

$$\begin{aligned}
\left\|\mathbf{x}_{i,t-1,s}^{(0)} - \mathbf{x}_{i,t-1}^{(0)}\right\|_2 &= \frac{c_{res}}{H\sqrt{m}}\left\|\sigma\left(\mathcal{W}_{t-1,s}^{(1)}\mathbf{x}_i\right) - \sigma\left(\mathcal{W}_{t-1}^{(0)}\mathbf{x}_i\right)\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}}L\left\|\mathcal{W}_{t-1,s}^{(1)} - \mathcal{W}_{t-1}^{(0)}\right\|_2 \\
&\le \frac{c_{res}}{H\sqrt{m}}L\eta\ell\hat{Q}'(t-1)
\end{aligned}$$

Thus

$$\left\|\mathbf{x}_{i,t-1,s}^{(l)} - \mathbf{x}_{i,t-1}^{(l)}\right\|_2 \le (\frac{c_{res}}{H\sqrt{m}}L\eta\ell\hat{Q}'(t-1) + \frac{\frac{c_{res}L}{H\sqrt{m}}\eta\ell\hat{Q}'(t-1)(c_xR' + c_{x,0})}{\frac{c_{res}L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R'\sqrt{m})})e^{\frac{c_{res}L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R'\sqrt{m} + \eta\ell\hat{Q}'(t-1))}$$

$$
\begin{aligned}
&\leq \eta\ell\hat{Q}'(t-1)(\frac{c_{res}}{H\sqrt{m}}L + \frac{(c_xR' + c_{x,0})}{(c_{w,0}\sqrt{m} + R'\sqrt{m})})e^{\frac{c_{res}L}{\sqrt{m}}(c_{w,0}\sqrt{m}+R'\sqrt{m}+\eta\ell\hat{Q}'(t-1))} \\
&\triangleq \eta\ell\hat{Q}'(t-1)C_x^*
\end{aligned}
$$

Similarly, through standard calculation we can get

$$
\left\|\mathbf{a}_{t-1,s} - \mathbf{a}_{t-1}\right\|_2 \leq \eta\ell\hat{Q}'(t-1)
$$

Lastly,

$$
\begin{aligned}
\left\|\mathbf{J}_{i,t-1,s}^{(l)} - \mathbf{J}_{i,t-1}^{(l)}\right\|_2 &= \left\|\sigma'(\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1,s}^{(l-1)}) - \sigma'(\mathcal{W}_{t-1}^{(l)}\mathbf{x}_{t-1}^{(l-1)})\right\|_2 \\
&\leq \beta\left\|\mathcal{W}_{t-1,s}^{(l)}\mathbf{x}_{t-1,s}^{(l-1)} - \mathcal{W}_{t-1}^{(l)}\mathbf{x}_{t-1}^{(l-1)}\right\|_2 \\
&\leq \beta(\left\|\mathcal{W}_{t-1,s}^{(l)}\right\|_F\left\|\mathbf{x}_{t-1,s}^{(l-1)} - \mathbf{x}_{t-1}^{(l-1)}\right\|_2 + \left\|\mathcal{W}_{t-1,s}^{(l)} - \mathcal{W}_{t-1}^{(l)}\right\|_F\left\|\mathbf{x}_{t-1}^{(l-1)}\right\|_2) \\
&\leq \beta(\frac{c_{res}L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R'\sqrt{m} + \eta\ell\hat{Q}'(t-1))\eta\ell\hat{Q}'(t-1)C_x^* + \eta\ell\hat{Q}'(t-1)(c_xR' + c_{x,0})) \\
&= \eta\ell\hat{Q}'(t-1)\beta(\frac{c_{res}L}{H\sqrt{m}}(c_{w,0}\sqrt{m} + R'\sqrt{m} + \eta\ell\hat{Q}'(t-1))C_x^* + (c_xR' + c_{x,0})) \\
&\triangleq \eta\ell\hat{Q}'(t-1)\beta C_J^*
\end{aligned}
$$

Thus, according to Lemma G.1 in Du et al. [2019], we have

$$
\begin{aligned}
&\left\|\hat{u}_i'^{(h)}(\theta(t)) - \hat{u}_i'^{(h)}(\theta(t) - s(\mathcal{W}_{t-1} - \mathcal{W}_t))\right\|_F \\
\leq&\frac{c_{res}}{H\sqrt{m}}||\mathbf{x}_{i,t-1,s}^{(h-1)} \cdot \left[\mathbf{a}_{t-1,s}^\top \prod_{l=h+1}^{H}\left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,t-1,s}^{(l)}\mathcal{W}_{t-1,s}^{(l)}\right)\mathbf{J}_{i,t-1,s}^{(h)}\right] \\
&- \mathbf{x}_{i,t-1}^{(h-1)} \cdot \left[\mathbf{a}_{t-1}^\top \prod_{l=h+1}^{H}\left(\mathbf{I} + \frac{c_{res}}{H\sqrt{m}}\mathbf{J}_{i,t-1}^{(l)}\mathcal{W}_{t-1}^{(l)}\right)\mathbf{J}_{i,t-1}^{(h)}\right]|| \\
&\leq \eta\ell\hat{Q}'(t-1)\frac{c_{res}}{H\sqrt{m}}2c_{x,0}2a_{2,0}Le^{2Lc_{w,0}}(\frac{C_x^*}{2c_{x,0}} + \frac{1}{2a_{2,0}} + \frac{c_{res}}{\sqrt{m}}\beta C_J^*)
\end{aligned}
$$

On the other hand

$$
\begin{aligned}
\left\|\hat{u}_i'^{(H+1)}(\theta(t)) - \hat{u}_i'^{(H+1)}(\theta(t) - s(\theta(t+1) - \theta(t)))\right\|_2 &\leq \left\|\mathbf{x}_{t-1,s}^{(H)} - \mathbf{x}_{t-1}^{(H)}\right\|_2 \\
&\leq \eta\ell\hat{Q}'(t-1)C_x^*
\end{aligned}
$$

In the end,

$$
\begin{aligned}
\left\|\mathbf{I}_2(t)\right\|_2 &\leq \eta\ell\hat{Q}'(t-1)\eta\ell\hat{Q}'(t-1)(\frac{c_{res}}{\sqrt{m}}2c_{x,0}2a_{2,0}Le^{2Lc_{w,0}}(\frac{C_x^*}{2c_{x,0}} + \frac{1}{2a_{2,0}} + \frac{c_{res}}{\sqrt{m}}\beta C_J^*) + C_x^*) \\
&\leq \eta^2\ell^2\hat{Q}'(t-1)^2(\frac{c_{res}}{\sqrt{m}}2c_{x,0}2a_{2,0}Le^{2Lc_{w,0}}(\frac{C_x^*}{2c_{x,0}} + \frac{1}{2a_{2,0}} + \frac{c_{res}}{\sqrt{m}}\beta C_J^*) + C_x^*) \\
&\leq \eta^2 C_{I_2}^* \left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2 \\
&\leq \frac{1}{16}\eta\lambda_0 \left\|\mathbf{y} - \hat{\mathbf{u}}(t)\right\|_2
\end{aligned}
$$

□

Lemma b.11. *If Condition b.1 holds for $t' = 0, \ldots, t-1$ and $\eta \leq c\lambda_0 H^2 n^{-2}\ell^{-2}S^{-1}$ for some small constant c, we have $\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\|_2^2 \leq \frac{1}{16}\eta\lambda_0\|\mathbf{y} - \hat{\mathbf{u}}(k)\|_2^2$.*

Proof of Lemma b.11.

$$
\|\hat{\mathbf{u}}(t+1) - \hat{\mathbf{u}}(t)\|_2^2 = \sum_{i=1}^{n}\left(\mathbf{a}_{t+1}^\top\mathbf{x}_{i,t+1}^{(H)} - \mathbf{a}_t^\top\mathbf{x}_{i,t}^{(H)}\right)^2
$$

$$
\begin{aligned}
=&\sum_{i=1}^{n}\left(\left[\mathbf{a}_{t+1}-\mathbf{a}_t\right]^\top \mathbf{x}_{i,t+1}^{(H)}+\mathbf{a}_t^\top\left[\mathbf{x}_{i,t+1}^{(H)}-\mathbf{x}_{i,t}^{(H)}\right]\right)^2\\
\leq& 2\left\|\mathbf{a}_{t+1}-\mathbf{a}_t\right\|_2^2\sum_{i=1}^{n}\left\|\mathbf{x}_{i,t+1}^{(H)}\right\|_2^2+2\left\|\mathbf{a}_t\right\|_2^2\sum_{i=1}^{n}\left\|\mathbf{x}_{i,t+1}^{(H)}-\mathbf{x}_{i,t}^{(H)}\right\|_2^2\\
\leq& 18n\eta^2\ell^2c_{x,0}^2Q'(t)^2+4n\left(\eta\ell a_{2,0}c_xQ'(t)\right)^2\\
\leq&\frac{1}{8}\eta\lambda_0\left\|\mathbf{y}-\hat{\mathbf{u}}(t)\right\|_2^2.
\end{aligned}
$$

□

References

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 440–445, 2017.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 2017.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. *Advances in Neural Information Processing Systems*, 31, 2018.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't Use Large Mini-Batches, Use Local SGD. art. arXiv:1808.07217, August 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Yue Yu, Jiaxiang Wu, and Longbo Huang. Double quantization for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: When does averaging help? art. arXiv:1606.07365, June 2016.