Achieving Byzantine-Resilient Federated Learning via Layer-Adaptive Sparsified Model Aggregation
Supplementary Material

Jiahao Xu Zikai Zhang Rui Hu
University of Nevada, Reno
{jiahaox,zikaiz,ruihu}@unr.edu

A. Attack and defense models

Attack model. We follow the attack model in previous works [1, 4, 11, 13]. Specifically, the attacker controls a subset of f malicious clients within the FL system. These clients can either be fake clients injected into the system by the attacker or genuine clients that have been compromised. The goal of the attacker is to degrade the overall performance of the global model in FL. The attacker has full knowledge of all benign updates in each training round. For additional background knowledge of the attacker, we follow the same settings of the proposed attack works. The malicious clients need not follow the prescribed local training protocol of FL and may send arbitrary local model updates to the server. Let $\mathcal{B}$ denote the set of benign clients in the system so that $\mathcal{B} \subset \mathcal{N}$. Under the Byzantine attack, the local model update of a client $i \in \mathcal{N}$ can be represented as

$$\Delta_i = \begin{cases} \Delta_i, & \text{if } i \in \mathcal{B} \\ \beta_i, & \text{if } i \notin \mathcal{B} \end{cases} \tag{1}$$

where $\beta_i \in \mathbb{R}^d$ represents an arbitrary model depending on the specific attack method.

Defense goal. Like previous works [3, 4, 13], we assume the server to be the defender who can deploy a robust aggregation rule, denoted by F, to mitigate the negative impact of malicious local models on the global model. The server has full access to the global model and local model updates in each training round, but it does not have access to the local training data of clients. We assume the server does not know the number of malicious clients unless explicitly specified. In addition, we assume that clients' submissions are made anonymously so that the server cannot track clients' actions.

B. Experimental settings

We utilize six benchmark datasets of FL, including MNIST [8], Fashion-MNIST [12], FEMNIST [2], CIFAR-10 [7], CIFAR-100 [7], and Shakespeare [9] datasets, to conduct the performance evaluation. The MNIST dataset is composed of gray-scale images of size 28×28 pixels for image classification tasks. It has 60,000 images for training and 10,000 images for testing. Similar to MNIST, Fashion-MNIST (FMNIST) dataset contains 70,000 28×28 grayscale images for 10 categories of fashion products. The dataset is divided into 60,000 training images and 10,000 test images. For MNIST and FMNIST datasets, we evenly split the training data over 6,000 clients so that the distribution of private datasets on each client is IID. The Federated Extended MNIST (FEMNIST) dataset is a non-IID FL dataset extended from MNIST. It consists of 805,263 images hand-written by 3,550 users for a total of 62 classes, including 52 for upper and lower case characters and 10 for digits. We subsample 5% of the original data following [2], resulting in 1,827 clients with a total of 450,632 images. The number of samples for each client ranges from 3 to 525. The Shakespeare dataset is naturally a non-IID FL dataset for the next character prediction tasks. Following [10], we process the original data and result in a dataset consisting of 37,784 samples from 715 clients.

The CIFAR-10 and CIFAR-100 datasest [7] is a collection of 60,000 32×32 color images with 50,000 training samples and 10,000 testing samples. All images are evenly distributed among 10/100 different classes, respectively. We split the training dataset over 100 clients for IID cases. For non-IID cases, we use Dirichlet distribution to simulate the non-IID settings on CIFAR-10 and CIFAR-100 datasets, which is controlled by a non-IID degree hyperparameter α. The default value of α is set to 0.5 in our work.

For MNIST, FMNIST, and FEMNIST datasets, given their identical image format and size, we use the same neural network architecture in [6]. Specifically, we use a CNN model composed of two convolutional layers, each followed by max-pooling and ReLU activation functions. Two linear layers are utilized to map features to classes. For CIFAR-10/100 datasets, we use ResNet-18 [5]. For the Shakespeare dataset, we implement a Recurrent Neural Network (RNN)

model following [10]. The RNN model takes a sequence of characters as input and then uses an embedding layer to convert each character into an 8-dimensional feature representation. Subsequently, two Long Short-Term Memory (LSTM) layers process these embedded characters, and a final linear layer with the softmax activation is applied.

For all datasets except CIFAR-10/100, the server randomly selects $h = 100$ clients per round to perform local computations. While for CIFAR-10/100, we set $h = 25$. We use SGD with momentum as the local solver, with the decay ratio and momentum parameters set to 0.99 and 0.9, respectively, for all datasets except for Shakespeare, where it is set to 0.999 and 0.5, respectively. The learning rate is set as $\eta = 0.1$ for all datasets except for Shakespeare, where it is set to $\eta = 1.0$. By default, the filtering radius is set as $\lambda_m = \lambda_d = 1.0$ for CIFAR-10/100. While for other datasets, we set λ_m to 2.0. We define the sparsification level (SL) to be $1 - k/d$. A higher SL implies more parameters are zeroed out. In our experiments, SL is set as 0.3 for all datasets by default. We run each experiment with three random seeds and report the average of the best testing accuracies achieved in each individual training. The experiments are conducted using PyTorch and executed on NVIDIA RTX A6000 GPUs.

C. Evaluated attack methods

We consider eight attack methods including three naive attack methods, and five SOTA attack methods to comprehensively evaluate our method.

- *Random attack.* The malicious clients send randomized updates that follow a Gaussian distribution $N(\mu, \sigma^2 \mathbf{I}_d)$. We set $\mu = (0, \ldots, 0) \in \mathbb{R}^d$ and $\sigma = 0.5$.
- *Noise attack.* The malicious clients perturb benign updates by adding Gaussian noise used in random attacks.
- *Sign-flip attack.* The malicious clients manipulate their model updates by flipping the sign coordinately.
- *Min-Max/Min-Sum attack [11].* The malicious model updates are crafted in two steps. In the first step, the attacker generates a malicious update by perturbing the average of all benign updates. Then, for Min-Max attack, the attacker optimizes the malicious update so that its maximum Euclidean distance with any benign update is upper-bounded by the maximum distance between any two benign updates, i.e., $\max_{i,j \in \mathcal{H}} \|\Delta_i - \Delta_j\|_2$. For Min-Sum attack, the malicious update is optimized to ensure that the sum of its distances with each benign update is upper-bounded by the maximum total distance of a benign update among other benign updates, i.e.,$\max_{i \in \mathcal{H}} \sum_{j \in \mathcal{H}} \|\Delta_i - \Delta_j\|_2$. We additionally test a stealthy version of Min-Sum attack, where the distance of the malicious update from any benign update is bounded by the minimum (rather than maximum) total distance of benign updates. This stealthy version is tested on all the datasets except for MNIST. We follow [11] to keep the updates of all malicious clients the same.
- *AGR-tailored Trimmed-mean attack [11].* AGR-tailored Trimmed-mean (TailoredTrmean) attack is designed to attack the defense method Trmean proposed in [14] by maximizing the Euclidean distance between the aggregated result of simple average and Trmean, respectively.
- *Lie attack [1].* The malicious clients apply slight changes to their local benign updates, making it hard to be detected. Specifically, the malicious clients calculate the element-wise mean μ_j and standard error σ_j of all updates and generate the element of malicious updates by $(\beta_i)_j = \mu_j - z \times \sigma_j$, where $j \in [d]$. The scaling factor z is set to 0.5 for all experiments.
- *ByzMean attack [13].* The ByzMean attack makes the mean of updates arbitrary malicious updates. Specifically, it divides malicious clients into two groups, each with m_1 and m_2 clients, respectively. Clients in the first group select any existing attack methods to generate their malicious updates, denoted as $\beta_{i,\forall i \in [m_1]}$. The clients in the second group generate their malicious updates to make the average of all updates exactly equal to the average of malicious updates in $[m_1]$, which can be expressed as $\beta_{i,\forall i \in [m_2]} = \frac{(n-m_1) \times \beta_{i,\forall i \in [m_1]} - \sum_{i=f+1}^{n} \Delta_i}{m_2}$ assuming the first f updates are malicious. We follow the same setting in [13], where the Lie attack is selected as the base attack method for the first group, and the size of two groups is set as $m_1 = \lfloor f/2 \rfloor$ and $m_2 = f - m_1$.

D. Additional experimental results

D.1. More results

In this section, we set the attack ratio to 25%, and for FEMNIST and Shakeperare datasets, we set λ_d to 1.5. As shown in Table 1, LASA demonstrates its robustness against the naive and SOTA attack methods in IID settings, whereas almost all other defense methods are vulnerable to at least one attack method. Under no attack, LASA achieves a test accuracy comparable to FedAvg on MNIST dataset. This demonstrates the effectiveness of LASA in maintaining accuracy, not just in adversarial environments, but also in benign environments.

For MNIST dataset, LASA achieves the best performance against naive attacks with the highest accuracy of

Table 1. The main results for MNIST, FEMNIST, and Shakespeare are presented.

Datasets (Model)	Defense Methods	No Attack	Naive Attacks			State-of-the-art Attacks					Average w/ Attacks
			Random	Noise	Sign-flip	TailoredTrmean	Min-Max	Min-Sum	Lie	ByzMean	
MNIST (CNN)	FedAvg	97.85	19.28	32.25	96.89	11.01	94.16	94.22	96.86	10.24	56.36
	TrMean	96.14	94.11	94.50	95.19	11.35	88.35	88.41	93.67	10.74	72.67
	GeoMed	94.59	94.66	94.66	94.21	94.76	63.99	52.29	80.82	94.17	83.69
	Multi-Krum	97.00	96.50	96.73	96.97	11.35	67.33	69.51	93.82	10.24	67.43
	Bulyan	94.95	96.42	96.41	94.20	11.70	63.89	68.00	90.98	54.88	71.06
	DnC	97.69	96.57	96.58	97.14	46.31	64.57	89.89	96.17	28.29	76.69
	SignGuard	96.64	97.70	97.70	96.85	97.78	97.58	97.46	**97.58**	97.63	97.54
	SparseFed	**97.86**	19.18	31.69	96.85	11.01	94.11	94.22	96.86	10.24	56.27
	LASA (Ours)	97.35	**97.96**	**98.27**	**97.26**	**97.94**	**97.93**	**97.94**	97.54	**97.94**	**97.85**
FEMNIST (CNN)	FedAvg	**84.27**	42.60	48.15	81.30	5.58	58.76	81.68	81.11	1.28	50.43
	TrMean	82.23	78.26	78.81	79.13	5.70	29.80	76.72	75.79	5.73	53.12
	GeoMed	75.57	75.48	75.47	71.67	76.19	68.27	28.13	22.56	74.32	61.01
	Multi-Krum	82.85	76.13	76.48	80.00	5.58	25.83	77.25	74.91	6.48	52.58
	Bulyan	77.10	81.68	81.65	73.50	5.97	19.17	60.55	58.98	18.02	49.94
	DnC	83.89	75.41	76.08	80.96	63.93	66.60	80.37	78.97	22.84	68.52
	SignGuard	83.06	83.75	83.75	79.43	83.80	83.80	82.59	82.58	83.78	82.68
	SparseFed	**84.27**	42.24	48.07	81.29	5.58	60.06	81.71	81.05	1.28	50.41
	LASA (Ours)	83.69	**84.07**	**84.05**	**81.72**	**84.26**	**84.19**	**83.60**	**83.52**	**84.14**	**83.94**
Shakespeare (LSTM)	FedAvg	63.74	45.00	47.28	60.43	39.01	59.17	63.35	62.79	24.24	50.41
	TrMean	63.15	59.09	59.43	59.83	42.23	57.54	62.60	61.86	37.38	54.75
	GeoMed	57.63	57.67	57.67	52.55	57.89	57.72	57.89	56.24	56.28	57.24
	Multi-Krum	62.26	61.55	61.73	59.11	35.11	54.30	62.09	58.34	23.16	52.92
	Bulyan	60.89	62.73	62.76	58.05	49.39	54.61	60.71	59.11	52.90	57.41
	DnC	64.67	61.38	61.47	60.80	59.32	61.10	**64.70**	62.30	56.18	60.65
	SignGuard	63.65	65.26	65.26	59.84	64.76	64.76	60.83	62.35	64.76	61.97
	SparseFed	63.72	44.49	47.24	60.40	39.24	59.84	63.31	62.77	24.27	50.69
	LASA (Ours)	**65.08**	**66.25**	**66.24**	**62.56**	**66.32**	**65.63**	64.02	**64.25**	**65.99**	**65.16**

97.96% for Random attack, 98.27% for Noise attack, and 97.26% for Sign-Flip attack, outperforming all other defense methods. In contrast, SignGuard, DnC and LASA can effectively defend against TailoredTrmean and ByzMean attacks. Under TailoredTrmean attack, LASA achieves the highest accuracy of 97.94%, which is +0.17% and +51.63% higher than SignGuard and DnC, respectively; under ByzMean attacks, LASA achieves the highest accuracy of 97.94%, which is 0.31% and +69.66% higher than SignGuard and DnC, respectively.

Compared to FedAvg under no attack, we can see that LASA can maintain the accuracy of FL in the benign environment with only a -0.57% accuracy drop on FEMNIST dataset and even a +1.34% accuracy increase on Shakespeare dataset. We also observe that the performance of classic robust aggregation rules, including Trmean, GeoMed, Multi-Krum, and Bulyan, is poor on non-IID datasets. For example, Trmean and Multi-Krum completely failed against the ByzMean attack on FEMNIST dataset, yielding an accuracy of 5.73% and 6.48%, respectively. As we discussed in the related works, in non-IID settings, the divergence between benign model updates will increase, making these classic methods hard to filter out malicious model updates. For FEMNIST dataset, LASA outperforms all other defense methods. It achieves an accuracy of 84.26% at best under TailoredTrmean attack, which is identical to that of Mean under no attack. In addition, LASA outperforms SignGuard more significantly in non-IID settings, compared to their performance in IID settings. Specifically on Shakespeare dataset, the performance of SignGuard is not stable. For example, under Sign-Flip attack, the accuracy of SignGuard drops to 59.84%, while LASA achieves the highest accuracy of 62.56% (+2.72%). Under Min-Sum attack, SignGuard's accuracy drops to 60.83%, while LASA achieves an accuracy of 64.017% (+3.19%), which is comparable to the best accuracy achieved by DnC.

In a nutshell, the performance of LASA is not only manifested in attack scenarios but also in the absence of any attacks, which aligns with the design principles of LASA. Moreover, LASA shows robustness to both IID and more challenging non-IID cases. By adeptly integrating pre-aggregation sparsification and layer-wise adaptive aggregation, LASA effectively mitigates the impact of updates that diverge from others. The robustness of LASA, illustrated by the above-mentioned results, emphasizes its potential as a robust defense method in securing federated learning environments against a wide collection of attacks, ultimately enhancing the reliability of federated learning systems.

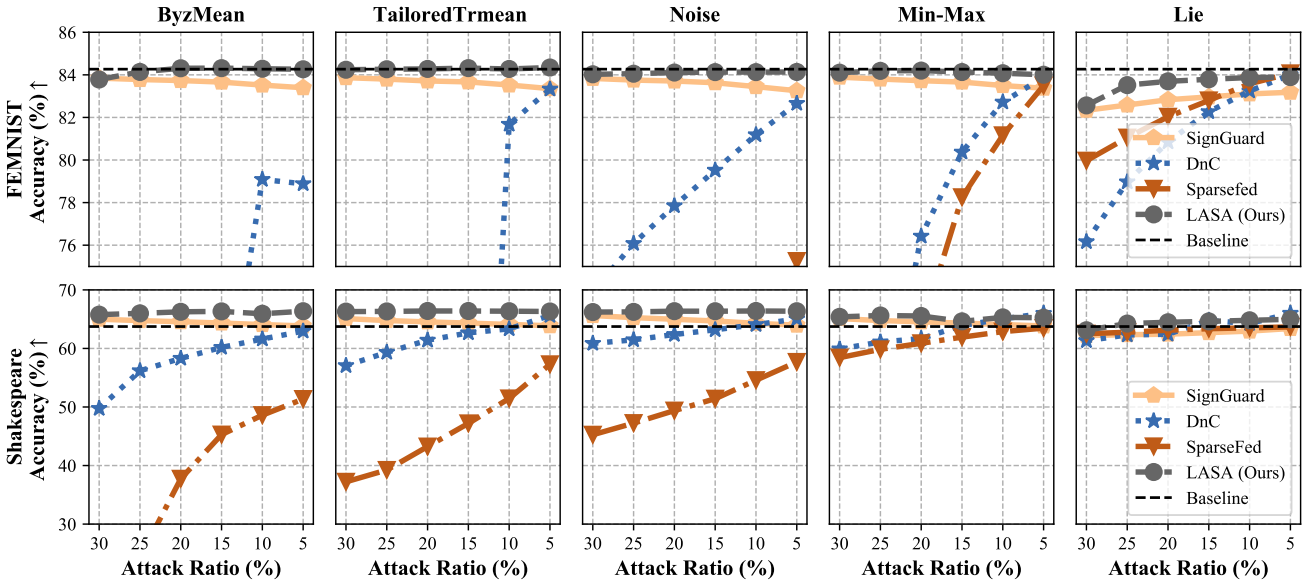

Figure 1. Testing Accuracy of LASA, SignGuard, DnC and SparseFed under Various Attack Ratios in non-IID Settings.

D.2. More results under various attack ratios

We evaluate the performance of three SOTA defense methods including DnC, SignGuard and SparseFed, and our method LASA under different attack ratios on non-IID datasets and report the results in Figure 1. Specifically, we conduct experiments under one naive attack and four SOTA attacks with the attack ratio varying from 5% to 30%. In Figure 1, the *Baseline* represents the non-robust method Mean under no attack. In general, DnC and SparseFed's accuracies increase as the attack ratio decreases, but they suffer from significant accuracy degradation when the attack ratio is high, especially under Byzmean and TailoredTrmean attacks. For instance, on FEMNIST dataset, even when the attack ratio is as low as 5%, SparseFed does not improve the robustness, achieving an accuracy of 7.44% under the ByzMean attack. Similarly, DnC struggles to defend against ByzMean attack effectively until the attack ratio is reduced to 10%, achieving a relatively low accuracy of 79.09%. SignGuard outperforms DnC and SparseFed significantly. However, under Byzmean, TailoredTrmean, Noise, and Min-Max attacks, the accuracy of SignGuard decreases as the attack ratio decreases. Compared to SignGuard, our method LASA achieves a better and more stable performance. As the attack ratio increases, LASA only has a minor decrease in accuracy.

D.3. Impact of sparsification level

As we stated in Section 4.1, the optimal sparsification parameter k should balance the tradeoff between sparsification error and robustness improvement. Here, we empirically study the impact of different k on learning performance. Recall that the SL is defined as $1 - k/d$, hence, a smaller k implies a higher SL and a heavier sparsification. We report the performance of LASA under Noise, Min-Max, and ByzMean attacks with SLs varying from 0.1 to 0.99 in Table 2, where M, FM, FEM, and Sha represent MNIST, FMNIST, FEMNIST, and Shakespeare datasets, respectively. The results demonstrate that there exists an optimal SL that maximizes robustness and a very high SL may lead to a significant accuracy drop. For example, as SL increases, the accuracy of LASA on FMNIST dataset increases to 87.94% and then decreases to 86.44% under ByzMean attack. This occurs because the sparsification error overwhelms the robustness improvement when SL is too large. We also observe that the sensitivity of LASA on SL depends on both

Table 2. Performance of LASA with Different Sparsification Levels.

Att.	Data.	Sparsification Level						
		0.1	0.3	0.5	0.7	0.9	0.95	0.99
ByzMean	M	97.833	97.943	97.821	97.543	**98.053**	97.880	97.490
	FM	87.820	87.647	**87.943**	87.867	87.803	87.740	86.437
	FEM	**84.143**	84.138	84.137	84.118	83.834	83.489	81.120
	Sha	**66.024**	65.990	65.842	65.409	64.355	63.055	60.463
Min-Max	M	97.310	**97.930**	97.493	97.307	97.557	96.950	97.593
	FM	87.917	87.907	87.920	**87.967**	87.330	87.707	86.353
	FEM	84.184	**84.264**	84.203	84.162	83.772	83.361	81.038
	Sha	64.723	**66.324**	65.472	65.032	63.654	62.527	60.090
Noise	M	98.223	98.270	**98.320**	97.643	98.050	97.923	97.800
	FM	87.877	87.870	87.893	**87.933**	87.623	87.897	86.423
	FEM	**84.061**	84.053	84.023	84.018	83.645	83.269	80.537
	Sha	**66.255**	66.244	66.102	65.754	64.506	63.546	60.686

Table 3. Performance of LASA with Different Filtering Radius

Con.		MNIST		FMNIST		FEMNIST	
λ_d	λ_m	Noise	ByzMean	Noise	ByzMean	Noise	ByzMean
1.0	1.0	97.963	97.803	87.950	**87.887**	83.922	84.158
1.0	1.5	97.883	97.843	**88.023**	87.720	83.946	**84.209**
1.0	2.0	**98.270**	97.943	87.870	87.647	84.007	84.119
1.0	4.0	91.743	97.840	77.400	77.930	69.408	84.048
1.5	2.0	97.927	**98.023**	87.937	87.640	84.053	84.138
2.0	2.0	97.593	97.487	87.950	84.000	84.136	77.399
3.0	2.0	97.883	66.897	87.917	67.250	**84.225**	28.300

the dataset and the attack method.

D.4. Impact of filtering radius

In this subsection, we study the performance of LASA with different filtering radius λ_m and λ_d. A smaller λ_m or λ_d indicates more stringent filtering and results in a smaller benign set for aggregation. As shown in Table 3, there exist optimal λ_m and λ_d that balance the filtering intensity and maximize the model accuracy. We also observe that the effectiveness of Noise attack is marginally affected by λ_d, as random noise perturbation does not change the sign purity in expectation. For all datasets, the optimal λ_d under Noise attack is 1.0 (note that for FEMNIST, the best accuracy when $\lambda_d = 3.0$ is comparable to the accuracy when $\lambda_d = 1.0$). However, as Noise attack adds Gaussian noise to the model updates to increase their magnitude (in L_2 norm), the effectiveness of Noise attack is sensitive to the values of λ_m. For different datasets, the optimal λ_m are different. For the advanced ByzMean attack, its effectiveness is marginally affected by λ_m, as the accuracy of LASA does not change much when λ_m increases from 1.0 to 2.0. This demonstrates that the magnitudes of malicious updates generated by ByzMean attack are close to that of benign models. In order to make the attack effective, ByzMean attack mainly focuses on manipulating the model direction, making it sensitive to the direction filtering radius λ_d: the accuracy of LASA vibrates a lot as λ_d increases. Additionally, both λ_m and λ_d should not be too large to compromise the effectiveness of the filtering.

E. Computational cost of LASA

We evaluate the computational cost of LASA in comparison to other methods. LASA incorporates pre-aggregation sparsification, leading to a complexity of $O(d \log d)$ due to the use of sorting algorithms like *merge sort* in the parameter space of local updates. Consequently, the worst-case computational expense for LASA is $O(nd \log d)$. Despite this, LASA's computational burden is on par with other methods such as Krum and Multi-Krum, which have a complexity of $O(dn^2)$, and Trmean with $O(dn \log n)$.

References

[1] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[2] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018. 1

[3] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *Proceedings of NDSS*, 2021. 1

[4] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to byzantine-robust federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Rui Hu, Yuanxiong Guo, and Yanmin Gong. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *IEEE Transactions on Mobile Computing*, 2023. 1

[7] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009. 1

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[9] 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. 1

[10] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020. 1, 2

[11] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021. 1, 2

[12] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

[13] Jian Xu, Shao-Lun Huang, Linqi Song, and Tian Lan. Byzantine-robust federated learning through collaborative malicious gradient filtering. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pages 1223–1235. IEEE, 2022. 1, 2

[14] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018. 2

F. Proof

F.1. Proof preliminaries

Lemma 1. *Given any two vectors $a, b \in \mathbb{R}^d$,*

$$2\langle a, b\rangle \leq \alpha \|a\|^2 + \frac{1}{\alpha}\|b\|^2, \forall\, \alpha > 0.$$

Lemma 2. *Given any two vectors $a, b \in \mathbb{R}^d$,*

$$\|a+b\|^2 \leq (1+\delta)\|a\|^2 + (1+\delta^{-1})\|b\|^2, \forall\, \delta > 0.$$

Lemma 3. *Given arbitrary set of n vectors $\{a_i\}_{i=1}^n$, $a_i \in \mathbb{R}^d$,*

$$\left\|\sum_{i=1}^n a_i\right\|^2 \leq n\sum_{i=1}^n \|a_i\|^2.$$

Lemma 4. *If the learning rate $\eta \leq 1/2\tau$, under Assumption 2 and 3, the local divergence of benign model updates are bounded as follows:*

$$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 \leq 2\bar{\nu} + \bar{\zeta} \tag{2}$$

Proof. Given that $\Delta_i = \eta\sum_{s=0}^{\tau-1} g_i^s$ where η is the learning rate and g_i^s is the local stochastic gradient over the mini-batch s. We have

$$\begin{aligned}
\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\eta\sum_{s=0}^{\tau-1} g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\eta\sum_{s=0}^{\tau-1} g_i^s\right\|^2 \qquad (3)\\
&= \frac{\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\sum_{s=0}^{\tau-1} g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} g_i^s\right\|^2\\
&\leq \frac{\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2\\
&= \frac{\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|(g_i^s - \nabla\mathcal{L}_i(\theta_i^s)) + \left(\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right) + (\nabla\mathcal{L}_i(\theta_i^s) - \nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s))\right\|^2\\
&\leq \frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1}\underbrace{\mathbb{E}\left\|g_i^s - \nabla\mathcal{L}_i(\theta_i^s))\right\|^2}_{T_1} + \frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1}\underbrace{\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2}_{T_2}\\
&\quad + \underbrace{\frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^s) - \nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s)\right\|^2}_{T_3}, \qquad (4)
\end{aligned}$$

where the first inequality follows Lemma 3, and the last second follows Lemma 2. For T_1, with Assumption 2, we have

$$T_1 \leq \bar{\nu}. \tag{5}$$

For T_2, we have

$$T_2 = \mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2 = \mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(\nabla\mathcal{L}_i(\theta_i^s) - g_i^s)\right\|^2 \leq \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^s) - g_i^s\right\|^2 \leq \bar{\nu}, \tag{6}$$

where the first inequality follows Lemma 3, and the last inequality follow Assumption 2. For T_3, we have

$$T_3 = \frac{3\tau\eta^2}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \sum_{s=0}^{\tau-1} \mathbb{E} \left\| \nabla \mathcal{L}_i(\theta_i^s) - \nabla \mathcal{L}_{\mathcal{B}}(\theta_i^s) \right\|^2 \leq 3\tau\eta^2 \sum_{s=0}^{\tau-1} \bar{\zeta} = 3\tau^2\eta^2\bar{\zeta} \tag{7}$$

by Assumption 3.
Plugging 5, 6, and 7 back to 4, with $\eta \leq 1/2\tau$, we have

$$\frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \mathbb{E} \left\| \Delta_i - \bar{\Delta}_{\mathcal{B}} \right\|^2 \leq 2\bar{\nu} + \bar{\zeta}.$$

This concludes the proof. □

F.2. Proof of Lemma 1

Proof. Recall that LASA denoted by $F(\cdot)\colon \mathbb{R}^{d\times n} \to \mathbb{R}^d$ is a layer-wise aggregation rule, i.e., there exist L real-valued functions $F_1, \ldots, F_L : \mathbb{R}^{d\times n} \to \mathbb{R}^d$ such that for all $\Delta_1, \ldots, \Delta_n \in \mathbb{R}^d$, $[F(\Delta_1, \ldots, \Delta_n)]_l = F_l(\Delta_1^l, \ldots, \Delta_n^l)$. As LASA utilizes layer-wise aggregation, we have

$$F_l(\Delta_1, \ldots, \Delta_n) = \frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} \hat{\Delta}_i^l,$$

where $\hat{\Delta}_i^l$ be the l-th layer of the Top-k sparsified model $\hat{\Delta}_i$ and $\mathcal{S}^l$ is the indices set of benign updates in l-th layer shown in Algorithm 1. We denote the indices set of Top-k parameters of a model/layer by $\mathcal{K}$ and the set of remaining parameters by $\mathcal{K}^-$. Let $[\Delta]_{\mathcal{K}}$ represent a sparsified model with only parameters in $\mathcal{K}$ (the rest are zero), then we have

$$\begin{aligned}
\mathbb{E}\|F(\Delta_1, \ldots, \Delta_n) - \bar{\Delta}_{\mathcal{B}}\|^2 &= \mathbb{E}\sum_{l=1}^{L} \left\|F_l(\Delta_1, \ldots, \Delta_n) - \bar{\Delta}_{\mathcal{B}}^l\right\|^2 \\
&= \mathbb{E}\sum_{l=1}^{L} \left\|\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{S}^l} \hat{\Delta}_i^l - \bar{\Delta}_{\mathcal{B}}^l\right\|^2 \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|^2} \left\|\sum_{i\in\mathcal{S}^l} \hat{\Delta}_i^l - \bar{\Delta}_{\mathcal{B}}^l\right\|^2 \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|^2} \left\|\sum_{i\in\mathcal{S}^l} \left[\hat{\Delta}_i^l - \bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^l} + \sum_{i\in\mathcal{S}^l} \left[-\bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^{l-}}\right\|^2 \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|^2} \left\|\sum_{i\in\mathcal{S}^l} \left[\Delta_i^l - \bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^l} + \sum_{i\in\mathcal{S}^l} \left[-\bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^{l-}}\right\|^2 \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|^2} \left(\left\|\sum_{i\in\mathcal{S}^l} \left[\Delta_i^l - \bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^l}\right\|^2 + \left\|\sum_{i\in\mathcal{S}^l} \left[\bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^{l-}}\right\|^2\right).
\end{aligned}$$

Let $c_i^l := \left\|\left[\Delta_i^l - \bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^l}\right\|^2 / \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2$, $b_{\mathcal{B}}^l := \left\|\left[\bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^{l-}}\right\|^2 / \left\|\bar{\Delta}_{\mathcal{B}}\right\|^2$, $C_{\mathcal{B}}^2 := \left\|\bar{\Delta}_{\mathcal{B}}\right\|^2$, $b_{\mathcal{B}} := \sum_{l=1}^{L} b_{\mathcal{B}}^l$, and $c_i := \sum_{l=1}^{L} c_i^l$, we have

$$\begin{aligned}
\mathbb{E}\|F(\Delta_1, \ldots, \Delta_n) - \bar{\Delta}_{\mathcal{B}}\|^2 &\le \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} \left(\left\|\left[\Delta_i^l - \bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^l}\right\|^2 + \left\|\left[\bar{\Delta}_{\mathcal{B}}^l\right]_{\mathcal{K}_i^{l-}}\right\|^2\right) \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} \left(c_i^l \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + b_{\mathcal{B}}^l \left\|\bar{\Delta}_{\mathcal{B}}\right\|^2\right) \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} \left(c_i^l \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + b_{\mathcal{B}}^l C_{\mathcal{B}}^2\right), \\
&= \mathbb{E}\sum_{l=1}^{L} \frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} c_i^l \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + C_{\mathcal{B}}^2 \sum_{l=1}^{L} b_{\mathcal{B}}^l \\
&= \underbrace{\mathbb{E}\frac{1}{|\mathcal{S}^l|} \sum_{i\in\mathcal{S}^l} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2}_{T_1} + C_{\mathcal{B}}^2 b_{\mathcal{B}}, \qquad (8)
\end{aligned}$$

where the first inequality follows Lemma 3. Note that $c_i = \left\|\left[\Delta_i - \bar{\Delta}_{\mathcal{B}}\right]_{\mathcal{K}_i}\right\|^2 / \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2$ and $b_{\mathcal{B}} = \left\|\left[\bar{\Delta}_{\mathcal{B}}\right]_{\mathcal{K}_i^-}\right\|^2 / \left\|\bar{\Delta}_{\mathcal{B}}\right\|^2$.
Now we treat T_1. If $\mathcal{S}^l \subseteq \mathcal{B}$, we have

$$T_1 = \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{S}^l} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] \leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right]. \tag{9}$$

If $\mathcal{S}^l \nsubseteq \mathcal{B}$, let $\mathcal{P} = \mathcal{S}^l \cap \mathcal{B}$, and $\mathcal{R} = \mathcal{S}^l \backslash \mathcal{B}$, let $C^2_{\mathcal{M},i} := \|\Delta_i\|^2, \forall i \in [N]\backslash\mathcal{B}$, then we have

$$\begin{aligned}
T_1 = \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{S}^l} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] &= \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\left(\sum_{i\in\mathcal{P}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \sum_{i\in\mathcal{R}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right)\right] \\
&\leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] + \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{R}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] + \mathbb{E}\left[\frac{2}{|\mathcal{S}^l|}\sum_{i\in\mathcal{R}} c_i \left(\|\Delta_i\|^2 + \left\|\bar{\Delta}_{\mathcal{B}}\right\|^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{|\mathcal{S}^l|}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2\right] + \mathbb{E}\left[\frac{2}{|\mathcal{S}^l|}\sum_{i\in\mathcal{R}} c_i \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right)\right],
\end{aligned} \tag{10}$$

where the second inequality follows Lemma 2.
Due to the use of MZ-score, models in $\mathcal{S}^l$ are centered around the median within a λ_m (and λ_d) radius. If the radius parameter λ_m or λ_d equals to zero, only the median model (based on l_2-norm or PDP) will be selected for averaging. To maximize benign model inclusion in averaging, the radius parameters λ_m and λ_d are set sufficiently large to ensure $|\mathcal{S}^l| \geq n/2 - f$. More precisely, assume there exist two positive constants λ_m^+ and λ_d^+, and if the radius parameters λ_m and λ_d in Algorithm 1 satisfy $\lambda_m \geq \lambda_m^+, \lambda_d \geq \lambda_d^+$, we have $|\mathcal{S}^l| \geq n/2 - f, \forall l \in [L]$. Integrated with 9 and 10, we have

$$\begin{aligned}
T_1 &\leq \begin{cases} \frac{2}{n-2f}\mathbb{E}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2, & \text{if } \mathcal{S}^l \subseteq \mathcal{B} \\ \frac{2}{n-2f}\mathbb{E}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \frac{4}{n-2f}\mathbb{E}\sum_{i\in\mathcal{R}} c_i \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right), & \text{if } \mathcal{S}^l \nsubseteq \mathcal{B} \end{cases} \\
&\leq \frac{2}{n-2f}\mathbb{E}\sum_{i\in\mathcal{B}} c_i \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \frac{4}{n-2f}\mathbb{E}\sum_{i\in\mathcal{R}} c_i \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right) \\
&\leq \frac{2c_{max}}{n-2f}\mathbb{E}\sum_{i\in\mathcal{B}} \left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \frac{4c_{max}}{n-2f}\sum_{i\in\mathcal{R}} \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right) \\
&= \frac{2c_{max}|\mathcal{B}|}{n-2f}\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \frac{4c_{max}}{n-2f}\sum_{i\in\mathcal{R}} \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right) \\
&= \frac{2c_{max}(n-f)}{n-2f}\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 + \frac{4c_{max}}{n-2f}\sum_{i\in\mathcal{R}} \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right) \\
&\leq \frac{2(n-f)}{n-2f}(2\bar{\nu} + \bar{\zeta})c_{max} + \frac{4c_{max}}{n-2f}\underbrace{\sum_{i\in\mathcal{R}} \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right)}_{T_2},
\end{aligned} \tag{11}$$

where the second inequality holds as $c_{max} := \max\{c_i, i \in [N]\}$ and the last inequality follows Lemma 4.
Assume the benign model update is bounded as $\|\Delta_i\|^2 \leq C^2, \forall i \in \mathcal{B}$, which can be achieved by using gradient clipping in practice. Assume the malicious model update is bounded as $\|\Delta_i\|^2 \leq C^2_{\lambda_m}, \forall i \in [N]\backslash\mathcal{B}$, which depends on the specific attack method and our magnitude-based filtering that is controlled by λ_m in Algorithm 1. We have

$$T_2 = \sum_{i\in\mathcal{R}} \left(C^2_{\mathcal{M},i} + C^2_{\mathcal{B}}\right) \leq |\mathcal{R}| \left(C^2_{\lambda_m} + C^2\right) \leq f\left(C^2_{\lambda_m} + C^2\right), \tag{12}$$

as $|\mathcal{R}| \leq |[N] \backslash \mathcal{B}| \leq f$. Therefore,

$$\begin{aligned} T_1 &\leq c_{max} \left(\frac{2(n-f)}{n-2f}(2\bar{\nu} + \bar{\zeta}) + \frac{4f}{n-2f} \left(C_{\lambda_m}^2 + C^2 \right) \right) \\ &\leq c_k \left(\frac{2(n-f)}{n-2f}(2\bar{\nu} + \bar{\zeta}) + \frac{4f}{n-2f} \left(C_{\lambda_m}^2 + C^2 \right) \right) \\ &\leq c_k \left(1 + \frac{f}{n-2f} \right) (4\bar{\nu} + 2\bar{\zeta} + 4C_{\lambda_m}^2 + 4C^2), \end{aligned} \tag{13}$$

if the sparsification applied to the local model update satisfies Assumption 4 so that $c_{max} \leq c_k$.
Summarizing to (8), we have

$$\begin{aligned} \mathbb{E}\|F(\Delta_1, \ldots, \Delta_n) - \bar{\Delta}_{\mathcal{B}}\|^2 &\leq T_1 + C_{\mathcal{B}}^2 b_{\mathcal{B}} \\ &\leq T_1 + b_k C^2 \\ &\leq c_k \left(1 + \frac{f}{n-2f} \right) (4\bar{\nu} + 2\bar{\zeta} + 4C_{\lambda_m}^2 + 4C^2) + b_k C^2 \end{aligned} \tag{14}$$

Discussion on the selection of k: When no sparsification is applied, i.e., when $k = d$, we have $c_k = 1$ and $b_k = 0$. In this case, the robustness upper bound is

$$\kappa_1 = \left(1 + \frac{f}{n-2f} \right) \left(4\bar{\nu} + 2\bar{\zeta} + 4C_{\lambda_m}^2 + 4C^2 \right) = O\left(1 + \frac{f}{n-2f} \right).$$

When $k = 0$, we have $c_k = 0$ and $b_k = 1$, then

$$\kappa = C^2,$$

which indicates the greatest sparsification error affecting robustness. When $0 < k < d$, the robustness upper bound is

$$\kappa_2 = (1+\epsilon) c_k \left(1 + \frac{f}{n-2f} \right) \left(4\bar{\nu} + 2\bar{\zeta} + 4C_{\lambda_m}^2 + 4C^2 \right) = O\left(c_k \left(1 + \frac{f}{n-2f} \right) \right)$$

if the sparsification parameter k is selected to satisfy that

$$\text{Condition 1} : c_k \leq \frac{1}{1+\epsilon}$$

and

$$\text{Condition 2} : \frac{b_k}{c_k} \leq \epsilon \left(\frac{4\bar{\nu} + 2\bar{\zeta} + 4C_{\lambda_m}^2}{C^2} + 4 \right)$$

with a positive constant ϵ. As $(1+\epsilon)c_k \leq 1$, we have

$$\kappa_2 \leq \kappa_1,$$

which demonstrates the effectiveness of sparsification for improving robustness. This finally concludes the proof. □

F.3. Proof of Theorem 1

Proof. Given the update rule $\theta^{t+1} = \theta^t - \bar{\Delta}^t = \theta^t - \eta\tilde{\Delta}^t$ where $\tilde{\Delta}_i^t := \sum_{r=0}^{\tau-1} g_i^{t,r} = \tau d_i^t$, for ease of expression, we let $\tilde{\Delta}_{\mathcal{B}^t} := \frac{1}{|\mathcal{B}^t|}\sum_{i\in\mathcal{B}^t}\tilde{\Delta}_i^t$ and $h_i^t := \mathbb{E}[d_i^t] = \mathbb{E}\left[(1/\tau)\sum_{r=0}^{r=\tau-1} g_i^{t,r}\right] = (1/\tau)\sum_{r=0}^{r=\tau-1}\nabla\mathcal{L}_i(\theta_i^{t,r})$. With Assumption 1, we have the following for all $t \in [0, T-1]$:

$$\begin{aligned}
\mathcal{L}_{\mathcal{B}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{B}}(\theta^t) &\leq \mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \theta^{t+1}-\theta^t\right\rangle + \frac{\mu}{2}\mathbb{E}\left\|\theta^{t+1}-\theta^t\right\|^2 \\
&= -\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}^t\right\rangle + \frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t\right\|^2 \\
&= -\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}^t + \tilde{\Delta}_{\mathcal{B}^t} - \tilde{\Delta}_{\mathcal{B}^t}\right\rangle + \frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t\right\|^2 \\
&= -\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}_{\mathcal{B}^t}\right\rangle - \eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\rangle + \frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t\right\|^2 \\
&= \underbrace{-\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \frac{1}{|\mathcal{B}^t|}\sum_{i\in\mathcal{B}^t}\tilde{\Delta}_i^t\right\rangle}_{T_1} + \underbrace{\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}_{\mathcal{B}^t} - \tilde{\Delta}^t\right\rangle}_{T_2} + \underbrace{\frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t\right\|^2}_{T_3}. \qquad (15)
\end{aligned}$$

Now we treat T_1, T_2, and T_3 respectively. We decompose T_1 by

$$\begin{aligned}
T_1 &= -\eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \frac{1}{|\mathcal{B}^t|}\sum_{i\in\mathcal{B}^t}\tilde{\Delta}_i^t\right\rangle = -\eta\tau\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \frac{1}{|\mathcal{B}^t|}\sum_{i\in\mathcal{B}^t} d_i^t\right\rangle = -\eta\tau\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\rangle \\
&= \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2, \qquad (16)
\end{aligned}$$

where we use the fact that $-2\langle a, b\rangle = \|a-b\|^2 - \|a\|^2 - \|b\|^2$.

We decompose T_2 as

$$T_2 = \eta\mathbb{E}\left\langle\nabla\mathcal{L}_{\mathcal{B}}(\theta^t), \tilde{\Delta}_{\mathcal{B}^t} - \tilde{\Delta}^t\right\rangle \leq \frac{\eta\alpha}{2}\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta}{2\alpha}\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2, \qquad (17)$$

where the first inequality follows Lemma 1 with a $\alpha > 0$.

We decompose T_3 as

$$\begin{aligned}
T_3 = \frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t\right\|^2 &= \frac{\mu\eta^2}{2}\mathbb{E}\left\|\tilde{\Delta}^t + \tilde{\Delta}_{\mathcal{B}^t} - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 \\
&\leq \mu\eta^2\mathbb{E}\left\|\tilde{\Delta}_{\mathcal{B}^t}\right\|^2 + \mu\eta^2\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 \\
&= \mu\eta^2\mathbb{E}\left\|\frac{1}{|\mathcal{B}^t|}\sum_{i\in\mathcal{B}^t}\tilde{\Delta}_i^t\right\|^2 + \mu\eta^2\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 \\
&\leq \frac{\mu\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2 + \mu\eta^2\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2, \qquad (18)
\end{aligned}$$

where the first inequality follows Lemma 2 with $\delta = 1$ and the second inequality follows Lemma 3.

Combining 16, 17, 18 and, 15, we get

$$\begin{aligned}
\mathcal{L}_{\mathcal{B}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{B}}(\theta^t) \leq & \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2 \\
& + \frac{\eta\alpha}{2}\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta}{2\alpha}\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 + \frac{\mu\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2 + \mu\eta^2\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 \\
= & -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \\
& + \left(\mu\eta^2 + \frac{\eta}{2\alpha}\right)\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2 + \frac{\mu\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2 \\
= & -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \\
& + \underbrace{\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right)\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2}_{T_4} - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2 + \frac{\mu\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2. \qquad (19)
\end{aligned}$$

T_4 can be decomposed as

$$T_4 = \left(\mu\eta^2 + \frac{\eta}{2\alpha}\right)\mathbb{E}\left\|\tilde{\Delta}^t - \tilde{\Delta}_{\mathcal{B}^t}\right\|^2 \leq \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \qquad (20)$$

where the first inequality holds as LASA is κ-robust aggregation rule with κ.

Plugging 20 back to 19, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{B}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{B}}(\theta^t) \leq & -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \\
& + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2 + \frac{\mu\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2 \\
= & -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \\
& + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) + \mu\eta^2\underbrace{\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2}_{T_5} - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i^t\right\|^2 \qquad (21)
\end{aligned}$$

T_5 can be charcterized as

$$
\begin{aligned}
T_5 = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\tilde{\Delta}_i^t\right\|^2 &= \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|d_i^t\right\|^2 = \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\mathbb{E}\left\|d_i^t - h_i^t\right\|^2 + \mathbb{E}\left\|h_i^t\right\|^2\right) \\
&= \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\mathbb{E}\left\|\frac{1}{\tau}\sum_{s=0}^{\tau-1}\left(g_i^{t,s} - \nabla\mathcal{L}_i(\theta_i^{t,s})\right)\right\|^2 + \mathbb{E}\left\|h_i^t\right\|^2\right) \\
&\leq \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\frac{1}{\tau}\sum_{s=0}^{\tau-1}\mathbb{E}\left\|g_i^{t,s} - \nabla\mathcal{L}_i(\theta_i^{t,s})\right\|^2 + \mathbb{E}\left\|h_i^t\right\|^2\right) \\
&\leq \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\frac{1}{\tau}\sum_{s=0}^{\tau-1}\nu_i^2 + \mathbb{E}\left\|h_i^t\right\|^2\right) \\
&= \frac{\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\nu_i^2 + \mathbb{E}\left\|h_i^t\right\|^2\right), \qquad (22)
\end{aligned}
$$

where the first inequality follows Lemma 3 and the second inequality follows Assumption 2.

Plugging 22 back to 21, we have

$$
\begin{aligned}
\mathcal{L}_{\mathcal{B}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{B}}(\theta^t) &\leq -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 - \frac{\eta\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}h_i^t\right\|^2 \\
&\quad + \frac{\mu\eta^2\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\nu_i^2 + \mathbb{E}\left\|h_i^t\right\|^2\right) + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \\
&= -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta^2\tau}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}h_i^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\mu\eta^2\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|h_i^t\right\|^2 \\
&\quad + \mu\eta\tau^2\bar{\nu}^2 + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \qquad (23) \\
&\leq -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{\tau}\sum_{r=0}^{\tau-1}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^{t,r}) - \nabla\mathcal{L}_i(\theta^t)\right\|^2 \\
&\quad + \frac{\mu\eta^2\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(2\mathbb{E}\left\|h_i^t - \nabla\mathcal{L}_i(\theta^t)\right\|^2 + \frac{2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta^t)\right\|^2\right) + \mu\eta^2\tau^2\bar{\nu}^2 + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \\
&\leq -\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \frac{\eta\tau}{2}\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{\mu^2}{\tau}\sum_{r=0}^{\tau-1}\mathbb{E}\left\|\theta_i^{t,r} - \theta^t\right\|^2 + \frac{\mu\eta^2\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{\mu^2}{\tau}\sum_{r=0}^{\tau-1}2\mathbb{E}\left\|\theta_i^{t,r} - \theta^t\right\|^2 \\
&\quad + \frac{4\mu\eta^2\tau^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}(\bar{\zeta} + \left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2) + \mu\eta^2\tau^2\bar{\nu}^2 + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \\
&= \left[-\left(\frac{\eta\tau}{2} - \frac{\eta\alpha}{2}\right) + 4\mu\eta^2\tau^2\right]\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 + \left(\frac{\eta\mu^2}{2} + 2\eta^2\tau\mu^3\right)\sum_{r=0}^{\tau-1}\underbrace{\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r} - \theta^t\right\|^2}_{T_6} \\
&\quad + 4\mu\eta^2\tau^2\bar{\zeta} + \mu\eta^2\tau^2\bar{\nu}^2 + \kappa\left(\mu\eta^2 + \frac{\eta}{2\alpha}\right) \qquad (24)
\end{aligned}
$$

where the second inequality follows Lemma 2 and the third inequality follow Assumption 1.

Now we treat T_6 as

$$
\begin{aligned}
T_6 &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r}-\theta^t\right\|^2 = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t-\eta g_i^{t,s-1}\right\|^2 \\
&= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t-\eta g_i^{t,s-1}+\eta\nabla\mathcal{L}_i(\theta^{t,s-1})-\eta\nabla\mathcal{L}_i(\theta^{t,s-1})+\eta\nabla\mathcal{L}_i(\theta^t)-\eta\nabla\mathcal{L}_i(\theta^t)\right\|^2 \\
&= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t-\eta\nabla\mathcal{L}_i(\theta^{t,s-1})+\eta\nabla\mathcal{L}_i(\theta^t)-\eta\nabla\mathcal{L}_i(\theta^t)\right\|^2 + \frac{\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|g_i^{t,s-1}-\nabla\mathcal{L}_i(\theta^{t,s-1})\right\|^2 \\
&\le \left(1+\frac{1}{2\tau-1}\right)\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{2\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_i(\theta^{t,s-1})+\nabla\mathcal{L}_i(\theta^t)-\nabla\mathcal{L}_i(\theta^t)\right\|^2+\eta^2\bar{\nu} \\
&\le \left(1+\frac{1}{2\tau-1}\right)\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{4\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_i(\theta^{t,s-1})-\nabla\mathcal{L}_i(\theta^t)\right\|^2 + \frac{4\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_i(\theta^t)\right\|^2+\eta^2\bar{\nu} \\
&\le \left(1+\frac{1}{2\tau-1}\right)\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{4\tau\mu^2\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{4\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_i(\theta^t)-\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)+\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+\eta^2\bar{\nu} \\
&\le \left(1+\frac{1}{2\tau-1}+4\tau\mu^2\eta^2\right)\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{8\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+8\tau\bar{\zeta}\eta^2+\eta^2\bar{\nu} \\
&\le \left(1+\frac{1}{\tau-1}\right)\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r-1}-\theta^t\right\|^2 + \frac{8\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+8\tau\bar{\zeta}\eta^2+\eta^2\bar{\nu}, \qquad (25)
\end{aligned}
$$

where the first and second inequality follows Lemma 2 with $\delta = 2\tau$ and $\delta = 1$, respectively. The third inequality follows Assumption 1, and the last inequality holds if $\eta \le 1/3\tau\mu$. Consequently, we have

$$
\begin{aligned}
T_6 = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\theta_i^{t,r}-\theta^t\right\|^2 &\le \sum_{h=0}^{s-1}\left(1+\frac{1}{\tau-1}\right)^h\left[8\tau\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+8\tau\bar{\zeta}\eta^2+\eta^2\bar{\nu}\right] \\
&\le (\tau-1)\left[\left(1+\frac{1}{\tau-1}\right)^\tau-1\right]\times\left[8\tau\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+8\tau\bar{\zeta}\eta^2+\eta^2\bar{\nu}\right] \\
&\le 32\tau^2\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+32\tau^2\bar{\zeta}\eta^2+4\tau\eta^2\bar{\nu}, \qquad (26)
\end{aligned}
$$

where the last inequality results from the fact that $\left(1+\frac{1}{\tau-1}\right)^t \le 5$ when $\tau > 1$.

Plugging 26 back to 24, we have

$$
\begin{aligned}
\mathcal{L}_{\mathcal{B}}(\theta^{t+1})-\mathcal{L}_{\mathcal{B}}(\theta^t) &\le \left[-\left(\frac{\eta\tau}{2}-\frac{\eta\alpha}{2}\right)+4\mu\eta^2\tau^2\right]\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+4\mu\eta^2\tau^2\bar{\zeta}+\mu\eta^2\tau^2\bar{\nu}^2+\kappa\left(\mu\eta^2+\frac{\eta}{2\alpha}\right) \\
&\quad+\left(\frac{\eta\mu^2}{2}+2\eta^2\tau\mu^3\right)\sum_{r=0}^{\tau-1}\left[32\tau^2\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+32\tau^2\bar{\zeta}\eta^2+4\tau\eta^2\bar{\nu}\right] \\
&= \left[-\left(\frac{\eta\tau}{2}-\frac{\eta\alpha}{2}\right)+4\mu\eta^2\tau^2\right]\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+4\mu\eta^2\tau^2\bar{\zeta}+\mu\eta^2\tau^2\bar{\nu}^2+\kappa\left(\mu\eta^2+\frac{\eta}{2\alpha}\right) \\
&\quad+\left(\frac{\eta\mu^2}{2}+2\eta^2\tau\mu^3\right)\left[32\tau^3\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+32\tau^3\bar{\zeta}\eta^2+4\tau^2\eta^2\bar{\nu}\right] \\
&= \left[\left[-\left(\frac{\eta\tau}{2}-\frac{\eta\alpha}{2}\right)+4\mu\eta^2\tau^2\right]+\left(\frac{\eta\mu^2}{2}+2\eta^2\tau\mu^3\right)\left(32\eta^2\tau^3\right)\right]\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \\
&\quad+\left(\frac{\eta\mu^2}{2}+2\eta^2\tau\mu^3\right)\left(32\tau^3\bar{\zeta}\eta^2+4\tau^2\eta^2\bar{\nu}\right)+4\mu\eta^2\tau^2\bar{\zeta}+\mu\eta^2\tau^2\bar{\nu}+\kappa\left(\mu\eta^2+\frac{\eta}{2\alpha}\right) \\
&\le -\eta\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+\left(\frac{\eta\mu^2}{2}+2\eta^2\tau\mu^3\right)\left(32\tau^3\bar{\zeta}\eta^2+4\tau^2\eta^2\bar{\nu}\right)+4\mu\eta^2\tau^2\bar{\zeta}+\mu\eta^2\tau^2\bar{\nu}+\kappa\left(\mu\eta^2+\frac{\eta}{2\alpha}\right) \\
&\le -\eta\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2+\kappa\left(\mu\eta^2+\frac{\eta}{4}\right)+7\eta\tau\bar{\zeta}+(1+\tau)\eta\bar{\nu} \qquad (27)
\end{aligned}
$$

where the second inequality holds with $\alpha \geq 2$, and $\eta \leq 1/3\mu\tau$.
Times $1/\eta$ to the both sides of 27, rearranging and summing it form $t = 0$ to $t = T - 1$ and dividing by T, one yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 \leq \frac{\left(\mathcal{L}_{\mathcal{B}}(\theta^0) - \mathcal{L}_{\mathcal{B}}(\theta^*)\right)}{T\eta} + \kappa\left(\mu\eta + 1\right) + 7\tau\bar{\zeta} + (1+\tau)\bar{\nu}.$$

Assume $\widetilde{\theta}$ is uniformly sampled from the sequence of outputs $\{\theta^0, \theta^1, \ldots, \theta^T\}$ generated by FL with LASA as the F, then we have

$$\mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\widetilde{\theta})\right\|^2 = \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2,$$

which concludes the proof. □