REFERENCES

Mack W Alford, Jean-Pierre Ansart, Günter Hommel, Leslie Lamport, Barbara Liskov, Geoff P Mullery, and Fred B Schneider. *Distributed systems: methods and tools for specification. An advanced course*. Springer-Verlag, 1985.

Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.

Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1505–1521, 2021.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Liad Blumrosen and Noam Nisan. Algorithmic game theory. *Introduction to Mechanism Design, Cambridge University Press, New York, USA*, 2007.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *the NIPS 2016 workshop on Private Multi-Party Machine Learning*, 2016.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2021.

Xiaoyu Cao, Jinyuan Jia, Zaixi Zhang, and Neil Zhenqiang Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In *Proceedings of the IEEE Symposium on Security and Privacy 2023*, 2023.

Alessandra Casella and Luis Sanchez. Storable votes and quadratic voting. an experiment on four california propositions. Technical report, National Bureau of Economic Research, 2019.

Bharat Chandar and E Glen Weyl. Quadratic voting in finite populations. *Available at SSRN 2571026*, 2019.

Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, pp. 903–912. PMLR, 2018.

Tianyue Chu, Alvaro Garcia-Recuero, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. Securing federated sensitive topic classification against poisoning attacks. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2022.

Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. In *International Conference on Learning Representations(ICLR)*, 2020.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pp. 1–12. Springer, 2006.

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)*, pp. 1605–1622, 2020.

Hector Garcia-Molina. Elections in a distributed computing system. *IEEE transactions on Computers*, 31(01):48–59, 1982.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Hanxi Guo, Hao Wang, Tao Song, Yang Hua, Zhangcheng Lv, Xiulang Jin, Zhengui Xue, Ruhui Ma, and Haibing Guan. Siren: Byzantine-robust federated learning via proactive alarming. In *Proceedings of the ACM Symposium on Cloud Computing*, pp. 47–60, 2021.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

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*, pp. 770–778, 2016.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Konstantinos Konstantinidis and Aditya Ramamoorthy. Byzshield: An efficient and robust system for distributed training. *Proceedings of Machine Learning and Systems*, 3:812–828, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pp. 33–37, 2018.

Steven P Lalley, E Glen Weyl, et al. Quadratic voting. *Available at SSRN*, 2016.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations(ICLR)*, 2020.

Yinghao Liu, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. Fedvoting: A cross-silo boosting tree construction method for privacy-preserving long-term human mobility prediction. *Sensors*, 21(24): 8282, 2021.

Xu Ma, Yuqing Zhou, Laihua Wang, and Meixia Miao. Privacy-preserving byzantine-robust federated learning. *Computer Standards & Interfaces*, 80:103561, 2022.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations(ICLR)*, 2018.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

Eric A Posner and E Glen Weyl. Voting squared: Quadratic voting in democratic politics. *Vand. L. Rev.*, 68:441, 2015.

David Quarfoot, Douglas von Kohorn, Kevin Slavin, Rory Sutherland, David Goldstein, and Ellen Konar. Quadratic voting in the wild: real people, real votes. *Public Choice*, 172(1):283–303, 2017.

Shashank Rajput, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Detox: A redundancy-based framework for faster and more robust gradient aggregation. *Advances in Neural Information Processing Systems*, 32, 2019.

Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

Giovanni Sartori. *The theory of democracy revisited*, volume 2. NJ, 1987.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Sparse binary compression: Towards distributed deep learning with minimal communication. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *Network and Distributed Systems Security (NDSS) Symposium*, 2021.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1354–1371. IEEE, 2022.

Jinhyun So, Başak Güler, and A Salman Avestimehr. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7):2168–2181, 2020.

Jy-yong Sohn, Dong-Jun Han, Beongjun Choi, and Jaekyun Moon. Election coding for distributed learning: Protecting signsgd against byzantine attacks. *Advances in Neural Information Processing Systems*, 33:14615–14625, 2020.

Nicolaus Tideman and Florenz Plassmann. Efficient collective decision-making, marginal cost pricing, and quadratic voting. *Public Choice*, 172(1):45–73, 2017.

E Glen Weyl. The robustness of quadratic voting. *Public choice*, 172(1):75–107, 2017.

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

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pp. 6893–6901. PMLR, 2019.

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

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33,01, pp. 5693–5700, 2019.

Kai Yue, Richeng Jin, Chau-Wai Wong, and Huaiyu Dai. Federated learning via plurality vote. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022. doi: 10.1109/TNNLS.2022.3225715.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning (ICML)*, pp. 26429–26446. PMLR, 2022.

Bo Zhao, Peng Sun, Tao Wang, and Keyu Jiang. Fedinv: Byzantine-robust federated learning by inversing local model updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 36(8), 9171–9179, 2022.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

We present the related supplements in the following sections. It contains the proof of theoretical analysis section for Theorem A.9, Theorem 4.1 and Theorem 4.5, experimental details, and extra results.

A Proof of Theoretical Analysis

A.1 Assumptions

Assumption A.1. The loss functions are L-smooth, which means they are continuously differentiable and their gradients are Lipschitz-continuous with Lipschitz constant $L > 0$, whereas:

$$\forall i \in N,\ \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d,\ \|\nabla \mathcal{L}(\boldsymbol{w}_1)) - \nabla \mathcal{L}(\boldsymbol{w}_2))\|_2 \leq L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$$
$$\|\nabla \ell(\boldsymbol{w}_1; \mathcal{D}) - \nabla \ell(\boldsymbol{w}_2; \mathcal{D})\|_2 \leq L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$$

Assumption A.2. The loss function $\ell(\boldsymbol{w}_i, D)$ are μ-strongly convex:

$$\exists \mu > 0, \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d, \nabla \ell(\boldsymbol{w}^*; \mathcal{D}) = 0, \nabla \mathcal{L}(\boldsymbol{w}^*) = 0$$
$$2\left(\mathcal{L}(\boldsymbol{w}_1) - \mathcal{L}(\boldsymbol{w}_2)\right) \geq 2\left\langle \nabla \mathcal{L}(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\rangle + \mu \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2$$
$$2\left(\ell(\boldsymbol{w}_1; \mathcal{D}) - l(\boldsymbol{w}_2; \mathcal{D})\right) \geq 2\left\langle \nabla \ell(\boldsymbol{w}_2; \mathcal{D}), \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\rangle + \mu \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2$$

Assumption A.3. The expected square norm of gradients $\boldsymbol{w}$ is bounded:

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \exists \mathcal{G}_{\boldsymbol{w}} < \infty, \mathbb{E}\|\nabla \ell(\boldsymbol{w}; \mathcal{D})\|_2^2 \leq \mathcal{G}_{\boldsymbol{w}}^2$$

Assumption A.4. The variance of gradients $\boldsymbol{w}$ is bounded:

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \exists \mathcal{V}_{\boldsymbol{w}} < \infty, \mathbb{E}\|\nabla \ell(\boldsymbol{w}; \mathcal{D}) - \mathbb{E}(\nabla \ell(\boldsymbol{w}; \mathcal{D})\|_2^2 \leq \mathcal{V}_{\boldsymbol{w}}$$

A.2 Proof of Theorem A.9 and Theorem 4.1

A.2.1 Lemmas

The lemmas we utilize in the proof of Theorem A.9 and Theorem 4.1, are presented here due to the page limit.

Lemma A.5. *Assume Assumption A.4 holds, according to our Algorithm 1, it follows that*

$$\mathbb{E}\left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \nabla \mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2 \leq (1 - 2\theta)\, \mathcal{C}\mathcal{V}_{\boldsymbol{w}} \sqrt{B}$$

Where

$$\mathcal{F}(\boldsymbol{w}^{t-1}) = \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \nabla \ell(\boldsymbol{w}_i^{t-1}; \mathcal{D}_i^{t-1})$$

Lemma A.6. *From Assumption A.1 and A.2, $\mathcal{L}(\boldsymbol{w})$ is L-smooth and μ-strongly convex. Then $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, one has*

$$\left\langle \nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\rangle \geq \frac{L\mu}{L + \mu} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + \frac{1}{L + \mu} \|\nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2)\|_2^2$$

Lemma A.7. *Assume Assumption A.1, Assumption A.2 and Lemma A.6 hold, we have*

$$\begin{aligned}\left\|\boldsymbol{w}^{t-1} - r\nabla \mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \leq & \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 \\ & + \left(r^2\left(1 + L^2\right) - \frac{2rL\mu + 1}{L + \mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \end{aligned} \tag{4}$$

Lemma A.8. *Assume Assumption A.3 holds, it follows that*

$$\mathbb{E}\sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 \leq (E-1)^2 r^2 \mathcal{G}_{\boldsymbol{w}}^2$$

A.2.2 PROOF OF LEMMAS

Lemmas A.5, Lemmas A.8, Lemmas A.6 and Lemmas A.7 are all the lemmas we utilise during the proof of Theorem A.9, and we prove them in that order. Notice, Lemmas A.6 are used in the proof Lemmas A.7, and Theorem A.9 is proved using Lemmas A.5, Lemmas A.8 and Lemmas A.7.

Proof of Lemma A.5

Proof. Due to Assumption A.4 and Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2 = \mathrm{Var}\left(\mathcal{F}(\boldsymbol{w}^{t-1})\right) &= \mathbb{E}_{\mathcal{S}^{t-1}}\left\|\sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1}\left(\nabla\ell(\boldsymbol{w}_i^{t-1};\mathcal{D}_i^{t-1}) - \nabla\ell(\boldsymbol{w}_i^{t-1})\right)\right\|_2^2 \\
&= \sum_{i\in\mathcal{S}^{t-1}}\left(p_i^{t-1}\right)^2\mathbb{E}\left\|\nabla\ell(\boldsymbol{w}_i^{t-1};\mathcal{D}_i^{t-1}) - \nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\
&\leq \sum_{i\in\mathcal{S}^{t-1}}\left(p_i^{t-1}\right)^2\mathcal{V}_{\boldsymbol{w}} \leq \mathcal{V}_{\boldsymbol{w}}\sum_{i\in\mathcal{S}^{t-1}}\left(\frac{v_i^{t-1}}{\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}}\right)^2 \\
&\leq \mathcal{V}_{\boldsymbol{w}}\frac{\sum_{i\in\mathcal{S}^{t-1}}\left(v_i^{t-1}\right)^2}{\left(\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}\right)^2} \leq \mathcal{V}_{\boldsymbol{w}}\frac{\sum_{i\in\mathcal{S}^{t-1}}\left(v_i^{t-1}\right)^2}{\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}} \\
&\leq \mathcal{V}_{\boldsymbol{w}}\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1} \leq (1-2\theta)\,qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B} \qquad (5)
\end{aligned}$$

□

Proof of Lemma A.6

Proof. Let $g(\boldsymbol{w}) = \ell(\boldsymbol{w}) - \frac{\varsigma}{2}\|\boldsymbol{w}\|_2^2$. Base on the Assumption A.2, we have $g(\boldsymbol{w})$ is $(L-\varsigma)$-strongly convex. from Bubeck et al. (2015) Equation 3.6, we have

$$\langle\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \geq \frac{1}{L}\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\|_2^2 \qquad (6)$$

Hence,

$$\langle\nabla g(\boldsymbol{w}_1) - \nabla g(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \geq \frac{1}{L-\varsigma}\|\nabla g(\boldsymbol{w}_1) - \nabla g(\boldsymbol{w}_2)\|_2^2 \qquad (7)$$

Now We have

$$\begin{aligned}
&\langle\nabla\left(\ell(\boldsymbol{w}_1) - \frac{\varsigma}{2}\|\boldsymbol{w}_1\|_2^2\right) - \nabla\left(\ell(\boldsymbol{w}_2) - \frac{\varsigma}{2}\|\boldsymbol{w}_2\|_2^2\right), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \\
&\qquad \geq \frac{1}{L+\mu}\left\|\nabla\left(\ell(\mathbf{w_1}) - \frac{\varsigma}{2}\|\boldsymbol{w}_1\|_2^2\right) - \nabla\left(\ell(\boldsymbol{w}_2) - \frac{\varsigma}{2}\|\boldsymbol{w}_2\|_2^2\right)\right\|_2^2 \qquad (8)
\end{aligned}$$

And therefore

$$\begin{aligned}
&\langle\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle - \langle\varsigma\boldsymbol{w}_1 - \varsigma\boldsymbol{w}_2, \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \\
&\qquad \geq \frac{1}{L-\varsigma}\|(\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)) - (\varsigma\boldsymbol{w}_1 - \varsigma\boldsymbol{w}_2)\|_2^2 \qquad (9)
\end{aligned}$$

Refer to Assumption A.1, we obtain

$$\begin{aligned}\langle \nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle &\geq \frac{L\varsigma}{L-\varsigma} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 - \frac{2\varsigma}{L-\varsigma} \langle \nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle \\ &+ \frac{1}{L-\varsigma} \|\nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2)\|_2^2 \\ &\geq -\frac{L\varsigma}{L-\varsigma} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + \frac{1}{L-\varsigma} \|\nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2)\|_2^2 \end{aligned} \tag{10}$$

Let $\varsigma = -\mu$, then we conclude the proof of Lemma A.6.

□

Proof of Lemma A.7

Proof. We have

$$\left\|\boldsymbol{w}^{t-1} - r_{t-1} \nabla \mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 = \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \underbrace{-2r_{t-1} \left\langle \nabla \mathcal{L}(\boldsymbol{w}^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}^* \right\rangle}_{\mathbf{A1}} + \underbrace{r_{t-1}^2 \left\|\nabla \mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{A2}} \tag{11}$$

For part $\mathbf{A1}$ under the Assumption A.2, Lemma A.6 and Maclaurin inequality, we have

$$\begin{aligned}\mathbf{A1} &= -2r_{t-1} \sum_{i=1}^N p_i^{t-1} \left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}^* \right\rangle \\ &= -2r_{t-1} \sum_{i=1}^N p_i^{t-1} \left(\left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1} \right\rangle \right) \\ &- 2r_{t-1} \sum_{i=1}^N p_i^{t-1} \left(\left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}_i^{t-1} - \boldsymbol{w}^* \right\rangle \right) \\ &\leq \sum_{i=1}^N p_i^{t-1} \left(r_{t-1}^2 \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 \right) - \\ &2r_{t-1} \sum_{i=1}^N p_i^{t-1} \left(\frac{1}{L+\mu} \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \frac{L\mu}{L+\mu} \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \right) \\ &= \left(r_{t-1}^2 - \frac{1}{L+\mu} \right) \sum_{i=1}^N p_i^{t-1} \left(\left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \right) \\ &+ \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \end{aligned}$$

From Assumption A.1 and Jensen inequality, we can derive:

$$\left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \nabla \ell(\boldsymbol{w}^*)\right\|_2^2 \leq L^2 \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \tag{12}$$

Hence for $\mathbf{A1}$, by Jensen inequality and Equation 12, we have

$$\begin{aligned}
\mathbf{A1} &\leq \left(r_{t-1}^2 - \frac{1}{L+\mu}\right) \sum_{i=1}^{N} p_i^{t-1} \left(\left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2\right) \\
&+ \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&\leq \left(r_{t-1}^2 - \frac{1}{L+\mu}\right) \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&\leq \left(r_{t-1}^2 - \frac{2r_{t-1}L\mu + 1}{L+\mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2
\end{aligned}$$

Similar for $\mathbf{A2}$, we have

$$\begin{aligned}
\mathbf{A2} &= r_{t-1}^2 \left\|\sum_{i=1}^{N} p_i^{t-1} \nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \leq r_{t-1}^2 \sum_{i=1}^{N} p_i^{t-1} \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\
&\leq r_{t-1}^2 L^2 \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&= r_{t-1}^2 L^2 \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2
\end{aligned}$$

Then we combine results of $\mathbf{A1}$ and $\mathbf{A2}$ for Equation 11, it follows that

$$\begin{aligned}
\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla \mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \leq &\left(r_{t-1}^2 \left(1 + L^2\right) - \frac{2r_{t-1}L\mu + 1}{L+\mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 \qquad (13)
\end{aligned}$$

□

Proof of Lemma A.8

Proof. For each E step FL necessitates a communication. As a result, for any $t-1 \geq 0$, $\exists t^* \leq t-1$ that $t - t^* \leq E, t^* \in T$, accordingly $\forall i, j \in \mathcal{S}^{t^*}, \boldsymbol{w}_i^{t^*} = \boldsymbol{w}_j^{t^*} = \boldsymbol{w}^{t^*}$. Then, based on $\mathbb{E}\left\|\mathbf{X} - \mathbb{E}\,\mathbf{X}\right\|_2^2 \leq \mathbb{E}\left\|\mathbf{X}\right\|_2^2$, Jensen inequality and Assumption A.3, we have

$$
\begin{aligned}
\mathbb{E}\sum_{i=1}^{N} p_i^{t-1}\left\|\boldsymbol{w}^{t-1}-\boldsymbol{w}_i^{t-1}\right\|_2^2 &= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)-\left(\boldsymbol{w}^{t-1}-\boldsymbol{w}^{t^*}\right)\right\|_2^2 \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\left[\mathbb{E}_{\mathcal{S}^{t^*}}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)-\mathbb{E}_{\mathcal{S}^{t^*}}\left[\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right]\right\|_2^2\right) \\
&\le \mathbb{E}_{\mathcal{S}^{t^*}}\left[\mathbb{E}_{\mathcal{S}^{t^*}}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)\right\|_2^2\right] \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right\|_2^2 \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\sum_{t=t^*}^{t-2}\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\,\mathbb{E}_{\mathcal{S}^{t^*}}\left(t-1-t^*\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\left\|\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\left\|\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2 \\
&\le \left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2 \qquad (14)
\end{aligned}
$$

□

A.2.3 THEOREM A.9

Theorem A.9. *Under Assumptions A.1, A.2, A.3 and A.4, and $m = 0$. Choose $\alpha = \frac{L+\mu}{\mu L}$ and $\beta = 2\frac{(L+1)(L+\mu)}{\mu L}$, then* FEDQV *satisfies*

$$
\mathbb{E}\,\mathcal{L}(\boldsymbol{w}^T)-\mathcal{L}(\boldsymbol{w}^*) \le \frac{L}{2\varphi+T}\left(\varphi\,\mathbb{E}\left\|\boldsymbol{w}^0-\boldsymbol{w}^*\right\|_2^2+\frac{\alpha^2}{2}\Delta\right) \qquad (15)
$$

Where

$$
\Delta = \left(E-1\right)^2\mathcal{G}_{\boldsymbol{w}}^2+\left(1-2\theta\right)\mathcal{C}\mathcal{V}_{\boldsymbol{w}}\sqrt{B},\ \varphi=\alpha\left(L+1\right),\ \boldsymbol{w}^t=\sum_{i=1}^{N} p_i^t\boldsymbol{w}_i^t,\ p_i^t=\frac{1}{\mathcal{C}}\mathbb{1}_{i\in\mathcal{S}^t}
$$

A.2.4 PROOF OF THEOREM A.9

Proof. In t round, due to $m = 0$, we have:

$$
\begin{aligned}
\left\|\boldsymbol{w}^t-\boldsymbol{w}^*\right\|_2^2 &= \left\|\boldsymbol{w}^{t-1}-r_{t-1}\mathcal{M}(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\right\|_2^2 = \left\|\boldsymbol{w}^{t-1}-r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\right\|_2^2 \\
&= \underbrace{\left\|\boldsymbol{w}^{t-1}-r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\right\|_2^2}_{\mathbf{A}}+\underbrace{r_{t-1}^2\left\|\mathcal{F}(\boldsymbol{w}^{t-1})-\nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{B}} \\
&\quad+\underbrace{2r_{t-1}\left\langle\boldsymbol{w}^{t-1}-r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*,\mathcal{F}(\boldsymbol{w}^{t-1})-\nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\rangle}_{\mathbf{C}} \qquad (16)
\end{aligned}
$$

Where

$$
\mathcal{M}(\boldsymbol{w}^{t-1}) = \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1}\mathcal{M}_i(\boldsymbol{w}_i^{t-1})
$$

Note that $\mathbb{E}\,\mathbf{C} = 0$. For the expectation of A, from Lemma A.7 and Lemma A.8, it follows that

$$\begin{aligned}\mathbb{E}[\mathbf{A}] &= \mathbb{E}\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \\ &\leq \left(r_{t-1}^2\left(1+L^2\right) - \frac{2r_{t-1}L\mu+1}{L+\mu}\right)\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\ &+ \left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2\end{aligned} \tag{17}$$

We use Lemma A.5 to bound $\mathbf{B}$, we have

$$\mathbb{E}[\mathbf{B}] \leq r_{t-1}^2\left(1-2\theta\right)qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B} \tag{18}$$

Hence, we have

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq\ & r_{t-1}^2\left(1+L^2\right)\mathbb{E}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\ & - \frac{2r_{t-1}L\mu+1}{L+\mu}\mathbb{E}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 + r_{t-1}^2\Delta\end{aligned} \tag{19}$$

where

$$\Delta = \left(E-1\right)^2\mathcal{G}_{\boldsymbol{w}}^2 + \left(1-2\theta\right)qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B}$$

For the learning rate r_t, $\exists\alpha > \frac{L+\mu}{2\mu L}, \exists\beta > 0$, such that $r_t = \frac{\alpha}{\beta+t} \leq \frac{1}{L+1}$. We use mathematical induction to prove the following statement:

Proposition: $\forall t \in \mathbb{N}, \mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+t}$, where $\gamma = \max\left\{\frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu}, \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\right\}$.

Let $P(t)$ be the statement $\mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+t}$, we give a proof by induction on t.

Base case: The statement $P(0)$ holds for $t = 0$:

$$\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta}$$

Inductive step: Assume the induction hypothesis that for a particular j, the single case $t = j$ holds, meaning $P(j)$ is true:

$$\mathbb{E}\left\|\boldsymbol{w}^j - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+j}$$

It follows that:

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{w}^{j+1} - \boldsymbol{w}^*\right\|_2^2 &\leq \left(r_t^2\left(1+L^2\right) - \frac{2r_tL\mu+1}{L+\mu}\right)\mathbb{E}\left\|\boldsymbol{w}^j - \boldsymbol{w}^*\right\|_2^2 + r_t^2\Delta \\ &\leq \left(1 - \frac{2L\mu\alpha}{(L+\mu)(\beta+j)}\right)\frac{\gamma}{\beta+j} + \left(\frac{\alpha}{\beta+j}\right)^2\Delta \\ &= \left[\frac{\alpha^2\Delta}{(\beta+j)^2} - \frac{2\alpha\mu L - L - \mu}{(\beta+j)^2(L+\mu)}\gamma\right] + \frac{\beta+j-1}{(\beta+j)^2}\gamma \\ &\leq \frac{\gamma}{\beta+j+1}\end{aligned}$$

Therefore, the statement $P(j+1)$ also holds true, establishing the inductive step. Since both the base case and the inductive step have been proved as true, by mathematical induction the statement $P(t)$ holds for $\forall t \in \mathbb{N}$.

We choose $\alpha = \frac{L+\mu}{\mu L}$ and $\beta = 2\frac{(L+1)(L+\mu)}{\mu L}$, and we have

$$\begin{aligned}\gamma &= \max\left\{\frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu}, \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\right\} \\ &\leq \frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu} + \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 \\ &= \alpha^2\Delta + 2(L+1)\alpha\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\end{aligned}$$

Then based on Assumption A.1 and Taylor expansion, we have the quadratic upper-bound of $\mathcal{L}(\cdot)$:

$$\mathcal{L}(\boldsymbol{w}_1) - \mathcal{L}(\boldsymbol{w}_2) \leq (\boldsymbol{w}_1 - \boldsymbol{w}_2)^T \nabla \mathcal{L}(\boldsymbol{w}_2) + \frac{L}{2} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2$$

It follows that

$$\begin{aligned}
\mathbb{E}\, \mathcal{L}(\boldsymbol{w}^T) - \mathcal{L}(\boldsymbol{w}^*) &\leq \frac{L}{2} \mathbb{E} \left\|\boldsymbol{w}^T - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma L}{2(\beta + T)} \\
&\leq \frac{L}{2\alpha(L+1) + T} \left(\frac{\alpha^2}{2} \Delta + \alpha(L+1)\, \mathbb{E} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 \right) \\
&= \frac{L}{2\varphi + T} \left(\varphi\, \mathbb{E} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \frac{\alpha^2}{2} \Delta \right)
\end{aligned}$$

Where

$$\Delta = (E-1)^2\, \mathcal{G}_{\boldsymbol{w}}^2 + (1 - 2\theta)\, \mathcal{C} \mathcal{V}_{\boldsymbol{w}} \sqrt{B},\ \varphi = \alpha\,(L+1)\,,\ \boldsymbol{w}^t = \sum_{i=1}^{N} p_i^t \boldsymbol{w}_i^t,\ p_i^t = \frac{1}{\mathcal{C}} \mathbb{1}_{i \in \mathcal{S}^t}$$

□

A.2.5 Proof of Theorem 4.1

Proof. In the t round, we have:

$$\begin{aligned}
\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 &= \left\|\boldsymbol{w}^{t-1} - r_{t-1} \mathcal{M}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \\
&= \left\|\boldsymbol{w}^{t-1} - r_{t-1} \mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^* + r_{t-1} \mathcal{F}(\boldsymbol{w}^{t-1}) - r_{t-1} \mathcal{M}(\boldsymbol{w}^{t-1})\right\|_2^2 \\
&= \underbrace{\left\|\boldsymbol{w}^{t-1} - r_{t-1} \mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2}_{\mathbf{A}} + \underbrace{r_{t-1}^2 \left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \mathcal{M}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{B}} \\
&\quad + \underbrace{2 r_{t-1} \left\langle \boldsymbol{w}^{t-1} - r_{t-1} \mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*, \mathcal{F}(\boldsymbol{w}^{t-1}) - \mathcal{M}(\boldsymbol{w}^{t-1}) \right\rangle}_{\mathbf{C}}
\end{aligned} \tag{20}$$

Where

$$\mathcal{M}(\boldsymbol{w}^{t-1}) = \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \mathcal{M}_i(\boldsymbol{w}_i^{t-1})$$

For the expectation of A, from Theorem A.9, it follows that

$$\mathbb{E}[\mathbf{A}] \leq \frac{1}{2\varphi + t} \left(2\varphi\, \mathbb{E} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \alpha^2 \Delta \right) \tag{21}$$

For $\mathbf{B}$, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{B}] &= r_{t-1}^2 \left\| \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \nabla \ell(\boldsymbol{w}_i^{t-1}) - \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\
&= r_{t-1}^2 \left\| \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \left(\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right) \right\|_2^2 \\
&\leq r_{t-1}^2 \left\| \sum_{i \in mN} p_i^{t-1} \left(\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right) \right\|_2^2
\end{aligned} \tag{22}$$

Where m is the percentage of the malicious parties.

Due to Equation 1, we have

$$\theta \leq \frac{\left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\rangle}{\left\|\nabla \ell(\boldsymbol{w}_i^{t-1}\right\| \cdot \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|} \leq 1 - \theta \tag{23}$$

Thus,

$$\theta \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| \leq \left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\rangle \leq (1-\theta) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| \tag{24}$$

Due to this, we have

$$\begin{aligned} &\left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 - 2(1-\theta) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}\right\| \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| + \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\ &\leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\ &\leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 - 2\theta \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| + \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 \end{aligned} \tag{25}$$

Hence we have

$$\begin{aligned} &\theta(2-\theta) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \left\|(1-\theta) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| - \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|\right\|_2^2 \\ &\leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\ &\leq (1-\theta^2) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \left\|\theta \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| - \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|\right\|_2^2 \end{aligned} \tag{26}$$

Hence,

$$\theta(2-\theta) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 \tag{27}$$

Due to the Triangle Inequality, we have

$$\sqrt{\theta(2-\theta)} \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| \leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| \leq \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| + \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| \tag{28}$$

It follows that:

$$\left(\sqrt{\theta(2-\theta)} - 1\right) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| \leq \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\| \tag{29}$$

By incorporating Equation 26 and leveraging the AM-GM inequality, we can derive the following expression

$$\begin{aligned} \left\|\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|_2^2 &\leq (1-\theta^2) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \left\|\theta \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\| - \left\|\mathcal{M}_i(\boldsymbol{w}_i^{t-1})\right\|\right\|_2^2 \\ &\leq \left(1 - \theta^2 + \left(1 + \theta + \sqrt{\theta(2-\theta)}\right)^2\right) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\ &\leq \left(4 + 6\theta - \theta^2\right) \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \end{aligned} \tag{30}$$

Therefore,

$$\mathbb{E}[\mathbf{B}] \leq r_{t-1}^2 \left\| \sum_{i \in mN} p_i^{t-1} \left(\sqrt{4 + 6\theta - \theta^2} \left\|\nabla \ell(\boldsymbol{w}_i^{t-1})\right\|\right) \right\|_2^2 \leq \left(4 + 6\theta - \theta^2\right) m^2 N^2 r_{t-1}^2 \mathcal{G}_{\boldsymbol{w}}^2 \tag{31}$$

Hence for $\mathbf{C}$, we have

$$\mathbb{E}[\mathbf{C}] \leq \frac{2mN\mathcal{G}_{\boldsymbol{w}} r_{t-1}^2 \sqrt{4 + 6\theta - \theta^2}}{2\varphi + t} \left(2\varphi \, \mathbb{E} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \alpha^2 \Delta\right) \tag{32}$$

Then based on Assumption A.1 and Taylor expansion, we have the quadratic upper-bound of $\mathcal{L}(\cdot)$:

$$\mathcal{L}(\boldsymbol{w}_1) - \mathcal{L}(\boldsymbol{w}_2) \leq (\boldsymbol{w}_1 - \boldsymbol{w}_2)^T \nabla \mathcal{L}(\boldsymbol{w}_2) + \frac{L}{2} \left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2^2$$

It follows that

$$\begin{aligned} \mathbb{E}\, \mathcal{L}(\boldsymbol{w}^T) - \mathcal{L}(\boldsymbol{w}^*) &\leq \frac{L}{2} \mathbb{E} \left\|\boldsymbol{w}^T - \boldsymbol{w}^*\right\|_2^2 \\ &\leq \frac{L + 2Lr_{T-1}\varpi}{2\varphi + T} \left(\varphi \, \mathbb{E} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \frac{\alpha^2}{2}\Delta\right) + \frac{L\varpi^2}{2} \end{aligned}$$

Where $\varphi = \alpha\,(L+1)$, $\varpi = mN\mathcal{G}_{\boldsymbol{w}} r_{T-1} \sqrt{4 + 6\theta - \theta^2}$ □

A.3 PROOF OF THEOREM 4.5

A.3.1 LEMMAS

Lemma A.10. *f is monotone: : $\forall v_{-i}$ and $\forall v_i^{'} > v_i$, if $f(v_i, v_{-i}) \in W_i$, then $f(v_i^{'}, v_{-i}) \in W_i$.*

Lemma A.11. *In* FEDQV, *$\forall i, v_i, v_{-i}$ that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = \Phi_i(v_{-i})$, where Φ_i is the critical value of a monotone function f on a single parameter domain that $\Phi_i(v_{-i}) = \sup_{v_i : f(v_i, v_{-i}) \notin W_i} v_i$.*

A.3.2 PROOF OF LEMMAS

Proof of Lemmas A.10

Proof. $\forall v_{-i}$ and $\forall v_i^{'} > v_i$, based on the voting scheme, if the party i who submit s_i join the aggregation with v_i, which means $f(v_i, v_{-i}) \in W_i$, then this party can also submit $\forall s_i^{'} < s_i$ that lead to $v_i^{'} > v_i$, and still join the aggregation. In other words, $f(v_i^{'}, v_{-i}) \in W_i$. Thus, f is monotone. □

Proof of Lemmas A.11

Proof. The number of parties is $\mathcal{C}$ in each round. In voting scheme that follows Equation 1, the parties whose $s_i \leq \theta$ and $s_i \geq 1 - \theta$ pay 0 credits voice. After Equation 2, the parties with 0 credit voice or 0 budget gain 0 vote. Assuming there are the top $k (k < \mathcal{C})$ parties in ranking whose payments are $c_{j \in k}$ ($c_{j \in k} > 0$). Notice in FEDQV, the payment function $p_i(v_i, v_{-i}) = c_i = v_i^2$.

$\forall j \in k$, if party j pays $c_j^{'} > p_j(v_j, v_{-j}) = \Phi_i(v_{-i}) = \sup_{v_i : f(v_i, v_{-i}) \notin W_i} v_i$, it will still remain in top k and join the aggregation. On the other hand, if party j pays $c_j^{'} < p_j(v_j, v_{-j}) = \Phi_i(v_{-i})$, then it will be replaced by the party $k + 1$ in the ranking, and party j will not be able to join the aggregation regardless of whether party $k + 1$ joins or not. As a result, in order to participate in the aggregation, the parties need to pay critical value, that is, $\forall i, v_i, v_{-i}$ that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = \Phi_i(v_{-i})$ □

A.3.3 PROOF OF THEOREM 4.5

Proof. According to Theorem 9.36 Blumrosen & Nisan (2007): a normalised mechanism on a single parameter domain is incentive compatible(truthful) if and only if:
(i) The selection rule is monotone.
(ii) For every party i participants in the aggregation ($v_i > 0$) pays the critical value $\Phi_i(v_{-i}) = \sup_{v_i : f(v_i, v_{-i}) \notin W_i} v_i$.
The first condition (i) and the second one (ii) are proofed in Lemma A.10 and Lemma A.11 respectively. Thus, the proposed scheme FEDQV is incentive-compatible (truthful). □

B FEDQV WITH ADAPTIVE BUDGETS ALGORITHM

Here we present a concise elucidation of key components of the Algorithm 2 as followings:

- **IRLS** (Iteratively Reweighted Least Squares): IRLS serves as an optimisation technique employed to solve specific regression problems. Within Chu et al. (2022), IRLS is utilised to compute the Subjective Observations of participating clients based on their parameter's confidence score, which is calculated using the repeated-median regression technique.
- **Subjective Observations**: Positive observations denoted by P_i^t signify acceptance of an update, while negative observations denoted by N_i^t indicate rejection. Consequently, positive observations enhance a client's reputation, and negative ones have the opposite effect.
- **Reputation Score Calculation**: The reputation score of a client is determined using a subjective logic model, formulated as follows:

$$R_i^t = \frac{\kappa P_i^t + W a}{\kappa P_i^t + \eta N_i^t + W}$$

Table 3: Default experimental settings

Explanation	Notation	Default Setting
Budget	B	25
Similarity threshold	θ	0.1
The number of parties	N	100
The fraction of selected parties	$\mathcal{C}$	10
The number of total steps	T	500
The number of local epochs	E	5
Learning rate	r	0.01
Local batch size		10
Loss function	$\mathcal{L}(\cdot)$	Cross-entropy
Repeating times	3	

Regarding the integration of the reputation model, our objective is to demonstrate how combining FEDQV with the reputation model enables the allocation of unequal budgets, thereby enhancing the robustness of standard FEDQV. This integration's adaptability extends beyond a single reputation model, allowing customisation to suit various needs. The example presented in the paper serves to showcase the concept's viability.

C EXPERIMENTAL DETAILS AND EXTRA RESULTS

C.1 EXPERIMENTAL DETAILS

Our simulation experiments are implemented with Pytorch framework Paszke et al. (2017) on the cloud computing platform Google Colaboratory Pro (Colab Pro) with access to Nvidia K80s, T4s, P4s and P100s with 25 GB of Random Access Memory. Table 3 shows the default setting in our experiments.

C.2 OVERVIEW OF FEDQV

Figure 3 provides an overview of our QV-based aggregation algorithm, which comprises two integral components: "similarity computation" executed on the party side and "voting scheme" managed on the server side. This visual representation encapsulates the essential steps involved in our approach.

C.3 STATE-OF-THE-ART ATTACKS

Labelflip Attack Fang et al. (2020): In the Label-Flip scenario, all the labels of the training data for the malicious clients are set to zero. This scenario simulates a directed attack, with the goal to disproportionally bias the jointly trained model towards one specific class. This is a data poisoning

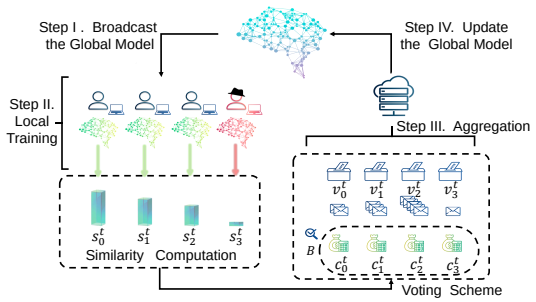

Figure 3: Overview of FEDQV algorithm.

attack that does not require knowledge of the training data distribution. Under this attack, the malicious parties train with clean data but with flipped labels. Specifically, we flip a label k as $K - k - 1$, where K is the total class number.

Gaussian Attack Zhao et al. (2022): This attack forges local model updates via Gaussian distribution on the malicious parties. malicious parties forge local model updates via Gaussian distribution.

Krum Attack Fang et al. (2020): Malicious parties craft poisoned local model updates opposite from benign ones, and enable them to circumvent the defence of Krum Blanchard et al. (2017).

Trim Attack Fang et al. (2020) The poisoned local model updates constructed by malicious parties are optimised for evading the Trim-mean and Median Yin et al. (2018).

Min-Max Attack Shejwalkar & Houmansadr (2021)In order to ensure that the malicious gradients closely align with the benign gradients within the clique, attackers strategically compute the malicious gradient. This computation is carried out to limit the maximum distance of the malicious gradient from any other gradient, which is constrained by the maximum distance observed between any two benign gradients.

Min-Sum Shejwalkar & Houmansadr (2021) The Min-Sum attack enforces an upper bound on the sum of squared distances between the malicious gradient and all the benign gradients. This upper bound is determined by the sum of squared distances between any one benign gradient and the rest of the benign gradients.

The targeted poisoning attacks include:

Backdoor Attack Gu et al. (2019) Malicious parties inject specific backdoor triggers into the training data and modify their labels to the attacker-chosen target label. Specifically, we use the same backdoor pattern trigger and attacker-chosen target label as in Bagdasaryan & Shmatikov (2021) as our trigger and set the attacker-chosen target label as 5.

the backdoor can be introduced into a model by an attacker who poisons the training data with specially crafted inputs. A backdoor transformation applied to any input causes the model to misclassify it to an attacker-chosen label The pattern must be applied by the attacker during local training, by modifying the digital image.

Scaling attack Bagdasaryan et al. (2020) The malicious parties generate poisoned local model updates by backdoor attack and only launch this attack during the last communication round after scaling these updates by a factor of N.

Neurotoxin attack Zhang et al. (2022) In this attack, the adversary starts by downloading the gradient from the previous round and employs it to approximate the benign gradient for the upcoming round. The attacker identifies the top-k% coordinates of the benign gradient and treats them as the constraint set. Over several epochs of Projected Gradient Descent (PGD), the attacker computes gradient updates on the manipulated dataset and projects this gradient onto the constraint set, which consists of the bottom-k% coordinates of the observed benign gradient. PGD is employed to approach the optimal solution within the span of the bottom-k% coordinates. We adopt the original parameter setting from the paper, where k is set to 0.1.

QV-Adaptive attack We introduce an adaptive attack, **QV-Adaptive**, tailored for FEDQV, utilising the Aggregation-agnostic optimizations Shejwalkar & Houmansadr (2021) within the LMP framework Fang et al. (2020). This attack manipulates both the similarity score and the local model, following the procedure below:

1) The malicious party i generates benign updates $\boldsymbol{w}_i^t$ using clean data $\mathcal{D}_i$ in round t and calculates the corresponding similarity score;

2) malicious parties (with counts of m) collectively normalise all the similarity scores and employ the Aggregation-agnostic Min-Max optimisation to select the optimal similarity score. This optimisation objective aims to increase the likelihood of the score being accepted by the server.

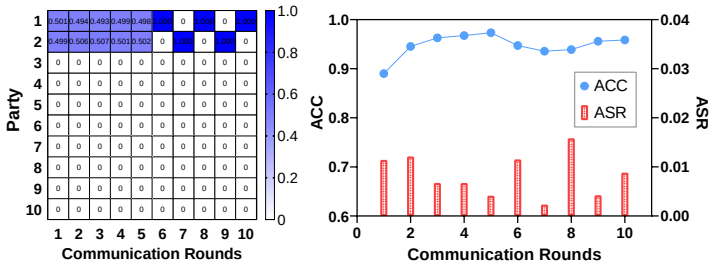

Figure 4: FEDQV aggregation weights of each party(left), ACC and ASR for global model(right), for 10 communication rounds in MNIST dataset under Backdoor attack

3) the adaptive attack focuses on local model poisoning to optimise the following problem:

$$\max \nu \tag{33}$$

$$\text{s.t.} \quad \boldsymbol{w}_{i\in m}^{t'} = \text{FedQV}(\boldsymbol{w}_1^t, \boldsymbol{w}_2^t, \ldots, \boldsymbol{w}_m^t) \tag{34}$$

$$\boldsymbol{w}_{i\in m}^{t'} = \boldsymbol{w}_i^t - \nu \hat{d} \tag{35}$$

Here, $\hat{d}$ represents a column vector encompassing the estimated changing directions of all global model parameters. The variables $\boldsymbol{w}_{i\in m}^t$ and $\boldsymbol{w}_{i\in m}^{t'}$ correspond to the local model before and after the attack. The parameter ν denotes the extent of the attack's impact on the model.

C.4 PRELIMINARY RESULTS

In FEDAVG, for example, if the malicious parties hold a substantial amount of local data and poison it, the accuracy of the global model would suffer owing to its aggregation rule. We use FEDQV to solve this dilemma.

To demonstrate how FEDQV constrain the influence of malicious parties, we consider two benign and one malicious party who conduct backdoor attacks with the amount of training data $\{1, 1, 2\}$. We train a multi-layer CNN for 10 rounds in the MNIST dataset same as in Section 5. The test accuracy is shown in Figure 1 in which the sides of the triangle correspond to the different parties and the position inside the triangle corresponds to their aggregation weights.

We observed that compared to FEDAVG with the weight $\{1, 1, 2\}$, QV, with the weight setup $\{1, 1, \sqrt{2}\}$, achieves higher accuracy. This suggests that QV can enhance performance by restraining the influence of attackers within FEDAVG. Consequently, when QV is integrated into FL with masked voting rules and a limited budget, as in FEDQV, it effectively excludes the malicious party and yields higher accuracy, represented by the weight configuration $\{1, 1, 0\}$.

To demonstrate how FEDQV compute the aggregation weights, consider the following scenario: there are 10 parties in the FL system, and 7 of them are attackers. The training consists of 10 communication rounds, during which attackers execute backdoor attacks. The rest of the settings are the same as the default. The result is shown in Figure 4. In the left of Figure 4, the first three parties are benign, and the rest are malicious. We observe that the aggregation weights of malicious parties are 0, implying that FEDQV succeed in eliminating their influence. As a result, ASR is quite low, and the accuracy of the global model is unaffected. This demonstrates that even if malicious parties dominate the majority, they do not prevail in damaging the global model.

C.5 NON-IID DEGREE

To concerning datasets with non-IID data across clients, our experiments incorporate datasets with non-IID characteristics, with a non-IID degree (ι) of 0.9. Moreover, we have examined the performance of FEDQV and FEDAVG across varying levels of non-IID data, spanning from 0.1 to 0.9, as depicted in Table 4.

Non-IID		0.1	0.3	0.5	0.7	0.9
FedQV	ACC(%)	84.94	86.01	83.88	81.37	75.96
	ASR(%)	3.39	4.55	17.64	20.59	24.18
FedAvg	ACC(%)	81.27	81.1	82.44	80.77	65.68
	ASR(%)	3.37	13.39	20.84	22.99	60.35

Table 4: Comparison of Accuracy (ACC) and Attack Success Rate (ASR) for FedQV and FedAvg under Backdoor Attack over 100 epochs with varying Non-IID Degrees on Fashion-MNIST Dataset.

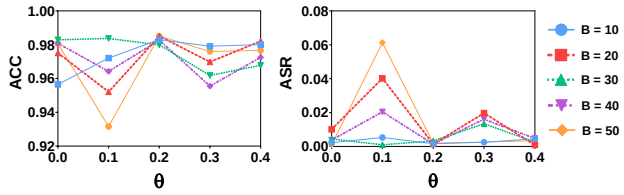

Figure 5: ACC and ASR as we vary the hyperparameters similarity threshold θ and budget B.

These results demonstrate that as the non-IID degree increases among the clients, the performance of the global model declines. Notably, FEDQV consistently maintains a superior performance compared to FEDAVG, even when confronted with different degrees of data heterogeneity under attack conditions.

C.6 IMPACT OF HYPERPARAMETERS

As noted, Theorem 4.1 provides general guidelines for tuning, and the findings from our grid search. As shown in Remark 4.3, the error rate is influenced by B and θ. To demonstrate the impact of these two hyper-parameters, we grid search B in $[10, 20, 30, 40, 50]$ and θ in $[0.1, 0.2, 0.3, 0.4, 0.5]$. The setup is the same as on the MNIST dataset under the backdoor attack with 30% malicious parties.

Figure 5 shows that the optimal values of B and θ are 30 and 0.2 respectively in this case. As B increases, there is a decline in ACC coupled with an increase in ASR. These results indicate that FEDQV's performance is not highly sensitive to the hyperparameters, as long as they are chosen in a reasonable range. The approach of combining theoretical guidelines with an exhaustive search to find optimal parameters is a commonly adopted strategy used in similar works.

We can see from Theorem 4.1, that the number of malicious devices m will affect the algorithm, and more malicious devices can lead to increased damage. However, this does mean the server needs to know the number of malicious devices to do the fine-tuning. We agree that determining optimal parameters can be challenging, especially in the absence of complete knowledge about the FL system.

A better tuning is possible if more information is available. For specific tasks, more information can indeed be collected from which practical parameter sets can be extracted either via exhaustive search or via simpler online algorithms using trial and error. We will add this to our future work and consider it when we study particular domain-specific problems using our method.

C.7 EXTRA RESULTS FOR INTEGRATION WITH BYZANTINE-ROBUST AGGREGATION

Table 5 demonstrates that when Multi-Krum are integrated with FedQV, its ACC increases by at least 28%, and its ASR decreases by at least 70%.

	MNIST		Fashion-MNIST	
	Multi-Krum	+ FEDQV	Multi-Krum	+ FEDQV
Backdoor				
ACC	70.20±9.99	**89.96±1.85**	33.24±13.24	**70.89±3.17**
ASR	32.03±11.20	**9.59±2.28**	68.87±17.77	**9.72±4.50**
Scaling				
ACC	68.35±16.76	**96.55±0.41**	59.43±14.22	**82.48±0.24**
ASR	33.65±19.15	**0.41±0.06**	33.64±19.08	**0.91±0.18**

Table 5: Comparison of Multi-Krum and Multi-Krum + FEDQV under targeted attacks with 30% malicious parties. The best results are in bold.

	Trimmed-Mean	Trimmed-Mean-QV	Trimmed-Mean	Trimmed-Mean-QV
Neurotoxin	ACC(%)	ACC(%)	ASR(%)	ASR(%)
1%	86.43	86.74	0.76	0.56
5%	84.96	86.34	0.92	0.72
10%	85.64	86.09	2.86	1.80
Backdoor				
1%	84.99	85.67	0.57	0.52
5%	84.83	85.66	0.93	0.46
10%	85.45	85.06	2.27	1.79

Table 6: Comparison of Trimmed-Mean and Trimmed-Mean Integrated with FedQV Methods under Targeted Attacks (Backdoor and Neurotoxin) Across Varying Percentages of Malicious Parties.