Technical Appendix: Exploring Metacognitive Features in Federated Learning

Anonymous submission

Appendix

1. Proof of Lemma 1

We will show that removing outliers reduces the variance for a set of points on a number line with scalar values. Let $\{a_i\}$ be a set where $a_i \in \mathbb{R}, i \in \mathbb{N}$ and $a_1 < a_2 < \ldots < a_N$. We consider one of those points, a_N, an outlier point a_o, meaning that a_o significantly deviates from the rest of the points. The mean $\bar{a}$ of $\{a_i\}$ is given as

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i. \tag{1}$$

If we exclude a_o, the new mean $\bar{a}'$ is

$$\bar{a}' = \frac{1}{N-1}\sum_{i=1}^{N-1} a_i. \tag{2}$$

But (8) can be rewritten as

$$\bar{a} = \frac{1}{N}\left(\sum_{i=1}^{N-1} a_i + a_o\right) \tag{3}$$

$$\bar{a} = \frac{N-1}{N}\bar{a}' + \frac{a_o}{N} \tag{4}$$

Equivalently,

$$\bar{a} - \bar{a}' = \frac{a_o - \bar{a}'}{N} \tag{5}$$

Variance σ^2 of the set without outlier removal:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2 \tag{6}$$

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N-1}(a_i - \bar{a})^2 + (a_o - \bar{a})^2\right) \tag{7}$$

Variance $(\sigma')^2$ of the set with a_o removed:

$$(\sigma')^2 = \frac{1}{N-1}\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 \tag{8}$$

The deviation of each term a_i around the mean $\bar{a}$ is

$$a_i - \bar{a} = a_i - \bar{a}' - (\bar{a} - \bar{a}') \tag{9}$$

Using (12):

$$a_i - \bar{a} = a_i - \bar{a}' - \frac{a_o - \bar{a}'}{N} \tag{10}$$

$$(a_i - \bar{a})^2 = \left(a_i - \bar{a}' - \frac{a_o - \bar{a}'}{N}\right)^2 \tag{11}$$

$$= (a_i - \bar{a}')^2 - 2(a_i - \bar{a}')\left(\frac{a_o - \bar{a}'}{N}\right) + \left(\frac{a_o - \bar{a}'}{N}\right)^2 \tag{12}$$

$$\sum_{i=1}^{N-1}(a_i - \bar{a})^2 = \sum_{i=1}^{N-1}(a_i - \bar{a}')^2 - 2\left(\frac{a_o - \bar{a}'}{N}\right)\sum_{i=1}^{N-1}(a_i - \bar{a}') + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2 \tag{13}$$

$\sum_{i=1}^{N-1}(a_i - \bar{a}') = 0$ due to sum of deviations around the mean being zero. Then (20) reduces to

$$\sum_{i=1}^{N-1}(a_i - \bar{a})^2 = \sum_{i=1}^{N-1}(a_i - \bar{a}')^2 + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2 \tag{14}$$

Plugging (21) into (14) we get

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2 + (a_o - \bar{a})^2\right] \tag{15}$$

Using (15):

$$\sigma^2 = \frac{N-1}{N}\sigma'^2 + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2 + (a_o - \bar{a})^2 \tag{16}$$

Given that a_o is sufficiently large, from (23) it follows that $\sigma^2 > \sigma'^2$.

2. Proof of Theorems 1.1, 1.2

According to lemma 1 (the inequality here is not strict because we might not remove any model weights at all):

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left(w_t^i-\mu_t^{\mathcal{G}}\right)^2 \leq \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\left(w_t^j-\mu_t^{\mathcal{A}}\right)^2 \tag{17}$$

Multiplying by $|\mathcal{G}|$ both sides and additionally multiplying the right side by $\frac{|\mathcal{A}|}{|\mathcal{A}|}$ yields:

$$\sum_{i\in\mathcal{G}}\left(w_t^i-\mu_t^{\mathcal{G}}\right)^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\left(w_t^j-\mu_t^{\mathcal{A}}\right)^2 \tag{18}$$

In vector notation using the Euclidean norm:

$$\left\|w_t^{\mathcal{G}}-\mu_t^{\mathcal{G}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2 \tag{19}$$

Because centroid $\mu_t^{\mathcal{A}}$ minimizes $\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2$:

$$\left\|w_t^{\mathcal{G}}-\mu_t^{\mathcal{G}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{G}}\right\|^2 \tag{20}$$

$$\lim_{t\to\infty}\left\|\mu_t^{\mathcal{G}}-w^*\right\| = 0$$

For $t \geq N$:

$$\left\|w_t^{\mathcal{G}}-w^*\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-w^*\right\|^2 \tag{21}$$

Finally,

$$\left\|w_t^{\mathcal{G}}-w^*\right\| \leq \sqrt{\frac{|\mathcal{G}|}{|\mathcal{A}|}}\left\|w_t^{\mathcal{A}}-w^*\right\| \qquad \blacksquare$$

2. Additional commentary to Theorems 1.1 and 1.2

If we further split $w_t^{\mathcal{A}}$ into "good" $w_t^{\mathcal{G}}$ and "bad" $w_t^{\mathcal{B}}$ clients ($\mathcal{B} = \{w_t^{i_1}, w_t^{i_2}, \ldots, w_t^{i_{|\mathcal{B}|}}\}$), we can derive the following, more detailed bound for the relation $\frac{\|w_t^{\mathcal{A}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$:

$$w_t^{\mathcal{A}} = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|}w_t^{\mathcal{G}} + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}w_t^{\mathcal{B}}$$

$$\left\|w_t^{\mathcal{A}}-w^*\right\| = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|}\left\|w_t^{\mathcal{G}}-w^*\right\| + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}\left\|w_t^{\mathcal{B}}-w^*\right\|$$

Dividing both sides by $\|w_t^{\mathcal{G}}-w^*\|$ yields

$$\frac{\left\|w_t^{\mathcal{A}}-w^*\right\|}{\left\|w_t^{\mathcal{G}}-w^*\right\|} = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|} + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}\frac{\left\|w_t^{\mathcal{B}}-w^*\right\|}{\left\|w_t^{\mathcal{G}}-w^*\right\|}$$

In comparison to Theorem 1.2, here we provide an equality, i.e. we can quantify the relation $\frac{\|w_t^{\mathcal{A}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$. However, since w^* in practice is unknown our approximation can only be based on μ_t. This would further increase the term $\frac{\|w_t^{\mathcal{B}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$ to $\frac{\|w_t^{\mathcal{B}}-\mu_t\|}{\|w_t^{\mathcal{G}}-\mu_t\|}$.

3. Code and Dataset Artifacts

Our code may be used for the reproduction and further reconfiguration of our experimental setup. Additionally, it provides the ability to collect and save metrics necessary for the further analysis. We also provide the datasets that we used to facilitate the reproduction of our empirical study experiments.

4. Related Work Overview

Various security mechanisms based on encryption and differential privacy have been proposed over the years. In (Liu et al. 2021) the authors proposed the FLAME framework, which employs randomized and encrypted gradient vectors sent to a shuffler to protect client identities. SplitFed (Thapa et al. 2022) combines split learning with FL to enhance data privacy and model robustness. The authors of Ensemble FL (Cao, Jia, and Gong 2021) utilize an ensemble approach to defend against malicious clients. A robust learning rate (RLR) mechanism is proposed in (Ozdayi, Kantarcioglu, and Gel 2021), which is regulated based on the sign information of the client updates. Certification is another effective technique in FL. CRFL (Xie et al. 2021) counters backdoor attacks by providing certified accuracy values. RoFL (Lycklama et al. 2023) is technique that allows the server to verify that clients' secret inputs satisfy a predefined constraint.

Advanced aggregation approaches have been a prominent research focus as well. The authors of FedInv (Zhao et al. 2022) proposed an aggregation technique to defend against stealthy data and model poisoning attacks. An aggregation mechanism based on trust bootstrapping through a small ground-truth dataset was presented in (Cao et al. 2020). Shapley values were utilized to perform more efficient client update aggregation in (Nagalapatti and Narayanam 2021). SVM has proved to be a useful aggregation tool in a recent study by (Wang et al. 2024). An aggregation technique designed specifically for data disparity and temporal unavailability was proposed by (von Wahl et al. 2024). (Ezzeldin et al. 2023) introduce FairFed motivated by the necessity of fairness-aware aggregation.

Adaptive learning mechanisms have also been extensively researched in recent years. FedAMP (Huang et al. 2021) presented a personalized FL approach based on attentive message passing, encouraging collaboration of similar clients. The authors of Self-Aware FL (Chen et al. 2022) proposed a bayesian approach to personalized FL that allows for atomatic model balancing. Ditto (Li et al. 2021) is another personalization approach that enhances fairness and robustness in FL. FedPLL (Yan and Guo 2024) is an adaptive regularization and data augmentation method that allows for more efficient partial learning. FedAC (Zang et al. 2024) is a three-step adaptive method that combines prospective momentum aggregation and fine-grained correction.

References

Cao, X.; Fang, M.; Liu, J.; and Gong, N. Z. 2020. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*.

Cao, X.; Jia, J.; and Gong, N. Z. 2021. Provably secure federated learning against malicious clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6885–6893.

Chen, H.; Ding, J.; Tramel, E. W.; Wu, S.; Sahu, A. K.; Avestimehr, S.; and Zhang, T. 2022. Self-aware personalized federated learning. *Advances in Neural Information Processing Systems*, 35: 20675–20688.

Ezzeldin, Y. H.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, A. S. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 7494–7502.

Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 7865–7873.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, 6357–6368. PMLR.

Liu, R.; Cao, Y.; Chen, H.; Guo, R.; and Yoshikawa, M. 2021. Flame: Differentially private federated learning in the shuffle model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8688–8696.

Lycklama, H.; Burkhalter, L.; Viand, A.; Küchler, N.; and Hithnawi, A. 2023. Rofl: Robustness of secure federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, 453–476. IEEE.

Nagalapatti, L.; and Narayanam, R. 2021. Game of gradients: Mitigating irrelevant clients in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9046–9054.

Ozdayi, M. S.; Kantarcioglu, M.; and Gel, Y. R. 2021. Defending against backdoors in federated learning with robust learning rate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9268–9276.

Thapa, C.; Arachchige, P. C. M.; Camtepe, S.; and Sun, L. 2022. Splitfed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8485–8493.

von Wahl, L.; Heidenreich, N.; Mitra, P.; Nolting, M.; and Tempelmeier, N. 2024. Data Disparity and Temporal Unavailability Aware Asynchronous Federated Learning for Predictive Maintenance on Transportation Fleets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15420–15428.

Wang, M.; Bodonhelyi, A.; Bozkir, E.; and Kasneci, E. 2024. TurboSVM-FL: Boosting Federated Learning through SVM Aggregation for Lazy Clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15546–15554.

Xie, C.; Chen, M.; Chen, P.-Y.; and Li, B. 2021. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, 11372–11382. PMLR.

Yan, Y.; and Guo, Y. 2024. Federated Partial Label Learning with Local-Adaptive Augmentation and Regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16272–16280.

Zang, Y.; Xue, Z.; Ou, S.; Chu, L.; Du, J.; and Long, Y. 2024. Efficient Asynchronous Federated Learning with Prospective Momentum Aggregation and Fine-Grained Correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16642–16650.

Zhao, B.; Sun, P.; Wang, T.; and Jiang, K. 2022. Fedinv: Byzantine-robust federated learning by inversing local model updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9171–9179.