Technical Report of Federating from History in Streaming Federated Learning

A MORE DETAILS ABOUT FED-HIST FRAMEWORK

A.1 Knowledge Generation

Details of data distribution.

We elucidate what constitutes historical information advantageous for self-training. Assume the private data $D_{i,t}$ managed by client i at time t is composed of samples of various labels, i.e., $D_{i,t} = \bigcup_{l=1}^{L} D_{i,t}^{l}$, where L is the number of classes and $D_{i,t}^{l}$ refers to the subset containing samples of a single label l. The label distribution of the client i at time t is calculated as follows:

$$Y_{i,t} = \{y_{i,t}^{1}, y_{i,t}^{2}, \ldots, y_{i,t}^{L}\} \text{ where } y_{i,t}^{l} = \frac{\|D_{i,t}^{l}\|}{\|D_{i,t}\|} \tag{1}$$

Obviously $\sum_{l=1}^{L} y_{i,t}^{l} = 1$ and $y_{i,t}^{l} \in [0, 1]$. For a given dataset D, if its corresponding label distribution $\hat{Y}$ is similar to the label distribution $Y_{i,t}$ of client i at time t, indicating a minimal difference between $Y_{i,t}$ and $\hat{Y}$, we consider D to be advantageous for the training at time t of client i.

B A MAIN PROOF OF THE CONVERGENCE ANALYSIS

B.1 Complete Assumptions

Assumption 1. *(Smoothness). For each $i = 1, \ldots, N$, the function F_i is continuously differentiable. There exist constant L such that for each $i = 1, \ldots, N$: $\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)$ and $\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)$ is L-Lipschitz with respect to $w_{i,t}^O$ and $w_{i,t}^P$.*

Assumption 2. *(Bounded Variance). The stochastic gradients in Fed-HIST are unbiased and have bounded variance. That is, for all $w_{i,t}^O$ and $w_{i,t}^P$,*

$$\mathbf{E}[\tilde{\nabla}_O F_i(w_{i,t}^O, w_{i,t}^P)] = \nabla_O F_i(w_{i,t}^O, w_{i,t}^P)$$
$$\mathbf{E}[\tilde{\nabla}_P F_i(w_{i,t}^O, w_{i,t}^P)] = \nabla_P F_i(w_{i,t}^O, w_{i,t}^P)$$

Furthermore, there exists constant σ such that

$$\mathbf{E}[\|\tilde{\nabla}_O F_i(w_{i,t}^O, w_{i,t}^P) - \nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2] \le \sigma^2$$
$$\mathbf{E}[\|\tilde{\nabla}_P F_i(w_{i,t}^O, w_{i,t}^P) - \nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2] \le \sigma^2 \tag{2}$$

This is a standard bounded variance assumption on the per-device stochastic gradients.

Assumption 3. *(Bounded Knowledge Aggregation). The distance before and after the knowledge aggregation from the server is bounded by ζ:*

$$\|w_{i,t,0}^O - w_{i,t-1}^O\|^2 \le \zeta^2$$

Throughout this paper, we assume F is bounded below by F^* and denote $\Delta F_0 = F(w_{i,0}^O, w_{i,0}^P) - F^*$.

B.2 Convergence Analysis

Theorem 1. ***(Convergence of Fed-HIST).*** *Suppose Assumptions 1, 2 and 3 hold and the learning rate in Fed-HIST is chosen depending on the problem parameters L, σ^2, E, and the number of rounds T, we have (ignoring absolute constants),*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{E}[\|\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2] + \mathbf{E}[\|\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2]) \\
&\le \frac{L^3 \Delta F_0}{T} + \frac{(L(1+L^2)\sigma^2 \Delta F_0)^{\frac{1}{2}}}{T^{\frac{1}{2}}} \\
&+ \frac{(L^2(1+L^2)\sigma^2 \Delta F_0^2)^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{(L^3(1+L^2)^2\sigma^2 \Delta F_0^3)^{\frac{1}{4}}}{T^{\frac{3}{4}}}
\end{aligned} \tag{3}$$

Proof. We give the main proof here, where a complete one is coming soon.

Proof Outline. We use the smoothness of F_i to obtain

$$\begin{aligned}
&F_i(w_{i,t+1}^O, w_{i,t+1}^P) - F_i(w_{i,t+1}^O, w_{i,t+1}^P) \\
&\le \left\langle \nabla_O F_i(w_{i,t}^O, w_{i,t}^P), w_{i,t+1}^O - w_{i,t}^O \right\rangle + \frac{L(1+L^2)}{2}\|w_{i,t+1}^P - w_{i,t}^P\|^2 \\
&+ \left\langle \nabla_P F_i(w_{i,t}^O, w_{i,t}^P), w_{i,t+1}^P - w_{i,t}^P \right\rangle + \frac{L(1+L^2)}{2}\|w_{i,t+1}^O - w_{i,t}^O\|^2
\end{aligned} \tag{4}$$

We bound each term of Eq. 4 and invoke them to get the bound shown in Eq. 6. Note that we simplified some constants appearing on the gradient norm terms using Eq. 5.

$$\eta_i \le (EL(1+L^2))^{-1} \tag{5}$$

$$\begin{aligned}
&F_i(w_{i,t+1}^O, w_{i,t+1}^P) - F_i(w_{i,t+1}^O, w_{i,t+1}^P) \\
&\le -\frac{\eta_i E}{4}\|\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2 - \frac{\eta_i E}{4}\|\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2 \\
&+ \frac{L(1+L^2)+1}{2}\zeta^2 + (1+L^2)\eta_i^2 E^2 \sigma^2 \\
&+ \eta_i \sum_{e=1}^{E} \mathbf{E}[L^2\|\tilde{w_{i,t}^O} - w_{i,t}^O\|^2 + L^4\|\tilde{w_{i,t}^P} - w_{i,t}^P\|^2] \\
&+ \frac{2\eta_i + 3\eta_i^2 EL(1+L^2)}{2}\sum_{e=1}^{E}\mathbf{E}[L^4\|\tilde{w_{i,t}^O} - w_{i,t}^O\|^2 + L^2\|\tilde{w_{i,t}^P} - w_{i,t}^P\|^2]
\end{aligned} \tag{6}$$

Our next step is to bound the last two lines of Eq. 6. We define $\eta_i = \frac{c}{EL}$ where $c < 1/\sqrt{6}max\{1, L^{-2}\}$ and choose ζ value. Then, after plugging in the learning rates and further simplifying the

constants,

$$
\begin{aligned}
&F_i(w_{i,t+1}^O, w_{i,t+1}^P) - F_i(w_{i,t+1}^O, w_{i,t+1}^P) \\
&\leq -\frac{c}{4L}\|\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2 - \frac{c}{4L}\|\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2 \\
&+ \frac{L(1+L^2)+1}{2}\zeta^2 + c^2\frac{(L^2+1)\sigma^2}{L} + c^3(1-\frac{1}{E})\frac{8\sigma^2(1+L^2)}{L} \\
&+ c^4(1-\frac{1}{E})\frac{6\sigma^2(1+L^2)^2}{L} \qquad (7)
\end{aligned}
$$

Taking full expectation, telescoping the series over $t = 0, \ldots, T = 1$, rearranging the results terms and ignoring absolute constants, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{E}[\|\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2] + \mathbf{E}[\|\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2]) \\
&\leq \frac{\Delta F_0}{\eta_i T} + \eta_i L(1+L^2)\sigma^2 + \eta_i^2 L^2(1+L^2)\sigma^2 + \eta_i^3 L^3(1+L^2)^2\sigma^2 \qquad (8)
\end{aligned}
$$

Moreover, we further set (ignoring absolute constants)

$$
\begin{aligned}
\eta_i = &\left(\frac{\Delta F_0}{L(1+L^2)\sigma^2 T}\right)^{\frac{1}{2}} \bigwedge \left(\frac{\Delta F_0}{L^2(1+L^2)\sigma^2 T}\right)^{\frac{1}{3}} \\
&\bigwedge \left(\frac{\Delta F_0}{L^3(1+L^2)^2\sigma^2 T}\right)^{\frac{1}{4}} \bigwedge \frac{1}{\sqrt{6}L^3 E} \qquad (9)
\end{aligned}
$$

Then, we have, ignoring absolute constants,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{E}[\|\nabla_O F_i(w_{i,t}^O, w_{i,t}^P)\|^2] + \mathbf{E}[\|\nabla_P F_i(w_{i,t}^O, w_{i,t}^P)\|^2]) \\
&\leq \frac{L^3\Delta F_0}{T} + \frac{(L(1+L^2)\sigma^2\Delta F_0)^{\frac{1}{2}}}{T^{\frac{1}{2}}} \\
&+ \frac{(L^2(1+L^2)\sigma^2\Delta F_0^2)^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{(L^3(1+L^2)^2\sigma^2\Delta F_0^3)^{\frac{1}{4}}}{T^{\frac{3}{4}}} \qquad (10)
\end{aligned}
$$

□

C MOTIVATION AND EXPERIMENT DETAILS

C.1 Motivation Details

The dataset comprises hourly temperature records for Dublin, Ireland, and Wellington, New Zealand, collected from 1997 to 2023 by local clients using a streaming pattern. Specifically, we have amassed 650,000 hourly weather records from Dublin spanning 1990 to 2020, and 240,000 hourly weather records from Wellington covering the period from 1997 to 2023. These records include detailed data points such as timestamps, and minimum, maximum, and average temperatures, and are utilized for both model training and inference [6, 10].

C.2 Experiment Details

C.2.1 Details of Metrics. To evaluate the performance of our algorithm, we employ accuracy metrics, specifically Top-1 and Top-5 accuracy, as our primary indicators[1]. Top-1 accuracy refers to the proportion of times the model's highest probability prediction matches the actual label, essentially measuring the model's ability to correctly identify the exact class for a given input. Conversely, Top-5 accuracy quantifies the instances where the correct label is among the model's top five predictions with the highest probabilities, thereby offering a more lenient assessment of the model's predictive capacity by acknowledging near-correct classifications.

[1]See this website for more details.

C.2.2 Baselines. To assess the performance of Fed-HIST, we selected six benchmark algorithms from the existing literature on various FL approaches for comparison, including FedAvg [5], FedBN [4], FedProx [3], pFedMe [1], FedWeIT [9], AF-FCL [8], ODE-Stream [2], DRSR-Stream [7] and StandAlone.

The detailed introduction of each baseline algorithm is shown as follows:

- **FedAvg:** FedAvg involves clients training their models locally, sending updated weights to a central server for aggregation, and receiving an updated model for further training. This cycle repeats until convergence.
- **FedBN:** FedBN, adapting to data heterogeneity in federated learning, incorporates batch normalization (BN) layers in local models, excluding BN parameters from aggregation to maintain model updates akin to FedAvg.
- **FedProx:** Building on FedAvg, FedProx introduces a proximal term to mitigate data and device heterogeneity issues, allowing clients to adjust their training intensity based on computational resources.
- **pFedMe:** Aiming at personalization, pFedMe employs a Moreau Envelope-based optimization, merging local and global models to enhance adaptability to client traits.
- **FedWeIT:** Specializing in federated continual learning, FedWeIT enables clients to learn from sequential tasks, promoting knowledge sharing on similar tasks.
- **AF-FCL:** Introduced for experimental comparison, AF-FCL targets mitigating catastrophic forgetting in federated continual learning, showcasing the potential of Fed-HIST in such scenarios.
- **ODE-Stream:** An online data selection framework, which uses a data valuation metric to enhance model convergence and accuracy by coordinating networked devices to store valuable data samples.
- **DRSR-Stream:** The Dynamic Ratio Selective Replacement (DRSR) algorithm dynamically updates the local cache of each client by selectively replacing cached data to mitigate distribution discrepancies between long-term data distributions and local training datasets.
- **StandAlone:** A single method that each client trains on data streams with no collaboration.

REFERENCES

[1] Canh T. Dinh, Nguyen Hoang Tran, and Tuan Dung Nguyen. 2020. Personalized Federated Learning with Moreau Envelopes. *CoRR* abs/2006.08848 (2020).

[2] Chen Gong, Zhenzhe Zheng, Fan Wu, Yunfeng Shao, Bingshuai Li, and Guihai Chen. 2023. To Store or Not? Online Data Selection for Federated Learning with Limited Storage. In *WWW 2023*. ACM, 3044–3055.

[3] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated Optimization in Heterogeneous Networks. In *MLSys 2020*.

[4] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *ICLR 2021*.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS 2017 (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 1273–1282.
[6] NIWA. 2024. *New Zealand's National Climate Database.* https://cliflo.niwa.co.nz/ Dataset.
[7] Heqiang Wang, Jieming Bian, and Jie Xu. 2024. On the Local Cache Update Rules in Streaming Federated Learning. *IEEE Internet Things J.* 11, 6 (2024), 10808–10816.
[8] Abudukelimu Wuerkaixi, Sen Cui, Jingfeng Zhang, Kunda Yan, Bo Han, Gang Niu, Lei Fang, Changshui Zhang, and Masashi Sugiyama. 2024. Accurate Forgetting for Heterogeneous Federated Continual Learning. In *ICLR 2024.*
[9] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. 2020. Federated Continual Learning with Adaptive Parameter Communication. *CoRR* abs/2003.03196 (2020).
[10] Met Éireann. 2024. *Dublin Climate Database.* https://www.met.ie/climate/available-data/historical-data Dataset.