FedCal: Achieving Local and Global Calibration in Federated Learning via Aggregated Parameterized Scaler

Hongyi Peng [1] **Han Yu** [1] **Xiaoli Tang** [1] **Xiaoxiao Li** [2 3]

Abstract

Federated learning (FL) enables collaborative machine learning across distributed data owners. However, this approach poses a significant challenge for model calibration due to data heterogeneity. While prior work focused on improving accuracy for non-iid data, calibration remains under-explored. This study reveals existing FL aggregation approaches lead to sub-optimal calibration, and theoretical analysis shows despite constraining variance in clients' label distributions, global calibration error is still asymptotically lower bounded. To address this, we propose a novel Federated Calibration (FedCal) approach, emphasizing both local and global calibration. It leverages client-specific scalers for local calibration to effectively correct output misalignment without sacrificing prediction accuracy. These scalers are then aggregated via weight averaging to generate a global scaler, minimizing the global calibration error. Extensive experiments demonstrate that FedCal significantly outperforms the best-performing baseline, reducing global calibration error by 47.66% on average.

1. Introduction

Federated learning (FL) (McMahan et al., 2017; Liu et al., 2024) has emerged as a novel distributed machine learning paradigm, enabling clients to train models collaboratively. A fundamental challenge in this domain is the data distribution heterogeneity among FL clients (i.e., the data non-IIDness issue). Numerous studies (Kairouz et al., 2021; Li et al., 2020b; Zhao et al., 2018) have demonstrated that this issue can adversely affect the accuracy and convergence of FL models. In response, extensive research efforts on personalized federated learning (PFL) (Tan et al., 2023; Li et al., 2020b; Karimireddy et al., 2020; Wang et al., 2020; Reddi et al., 2020) have been directed towards mitigating this issue.

However, an aspect that has been largely overlooked in the current PFL literature is *reliability* (Lu & Kalpathy-Cramer, 2021; Plassier et al., 2023). As FL systems become increasingly embedded in high-stake decision-making processes, the importance of reliability, especially in mission-critical applications such as healthcare (Rieke et al., 2020; Dayan et al., 2021; Sheller et al., 2020), finance (Long et al., 2020; Dash et al., 2022) and autonomous driving systems (Li et al., 2021; Du et al., 2020), cannot be overstated.

Besides prediction accuracy, the reliability of an FL model also hinges on accurate uncertainty estimation, commonly referred to as *confidence* (Ovadia et al., 2019; Yu et al., 2022; Guo et al., 2017a). Consider a model for cancer diagnosis. It predicts that a patient has cancer with a confidence level of 0.1. This confidence score carries significant implications. For instance, a patient predicted with a 0.1 confidence might receive a more cautious treatment recommendation. In a well-calibrated model, if 10 patients are assigned a confidence of 0.1 for having cancer, approximately one of them should genuinely be diagnosed correctly with the disease. This process of aligning the model prediction confidence with the actual observed frequency of an event is referred to as *model calibration* (Guo et al., 2017b).

Model calibration is of critical importance to producing reliable machine learning models. While the topic has been extensively examined under centralized learning settings, it has been largely overlooked under FL settings. As highlighted in Figure 1, the prevailing FedAvg-based (McMahan et al., 2017) FL model aggregation approaches produce poorly calibrated models in the presence of data heterogeneity. Recognizing this gap, we delve into the unique challenges posed by FL, and contend that existing centralized calibration techniques are not directly applicable in the face of the distributed and data heterogeneous nature of FL.

Model calibration in FL settings faces two major challenges:

[1]College of Computing and Data Science, Nanyang Technological University, Singapore. [2]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. [3]Vector Institute, Canada. Correspondence to: Han Yu <han.yu@ntu.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

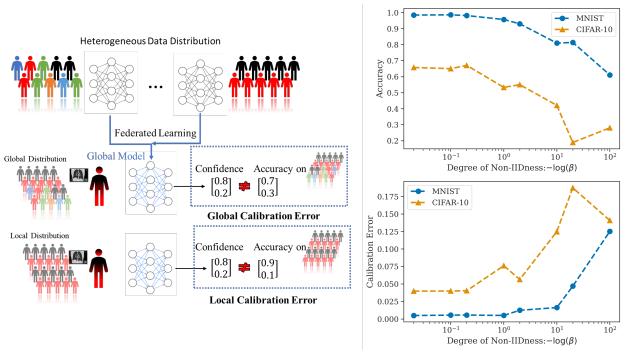

Figure 1. [Left] Impacts of Data Distribution Discrepancies on Model Calibration in Federated Learning. The presence of non-IID data across local nodes contributes to miscalibration issues in the aggregated model, influencing both local and global datasets. [Right] Impact of data heterogeneity on the accuracy and reliability of FL models. As the degree of non-IIDness (quantified by the Dirichlet distribution parameter β) increases, both accuracy and reliability of FL models trained on MNIST (MLP, 10 clients) and CIFAR-10 (ResNet-14, 10 clients) using FedAvg (McMahan et al., 2017) deteriorate.

1. **Need for Local and Global Calibration.** Our empirical observations reveal a dual challenge for calibration in data heterogeneous FL. Firstly, data heterogeneity negatively impacts the overall calibration of the global model. Secondly, the variance of calibration errors across clients increases with the degree of data heterogeneity. This necessitates a two-pronged approach: a personalized calibration step adjusting the model to individual client's data distribution, and a robust global calibration step for improved generalizability.

2. **Lack of Global Validation Datasets.** Given the distributed nature of FL, it is often impractical to have a comprehensive global validation set for global calibration due to privacy concerns and the challenges of data collection and maintenance.

To bridge this gap, our goal is to achieve both local and global model calibration via FL aggregation without relying on the existence of a global validation dataset. To achieve this goal, we propose a novel Federated Calibration (FedCal) approach. It involves training post-hoc scalers on local datasets for local calibration and subsequently aggregating them to achieve global calibration. Notably, the scalers in FedCal can be aggregated simply by averaging their parameters. As a post-hoc calibration method, FedCal can be easily integrated with existing FL methods.

Theoretical analysis shows that despite constraining the variance in clients' label distributions, the global model calibration error still asymptotically decreases. Extensive experimental evaluation based on four benchmark datasets reveals that FedCal significantly surpasses five state-of-the-art methods in both local and global calibration, regardless of the presence of global validation sets. Specifically, it reduces the global model calibration error by 47.66% on average compared to the best-performing baseline. Furthermore, we observe that ensembling global and local models can further enhance prediction accuracy.

2. Related Work

Calibration is an important research topic in centralized ML. A vast body of literature exists on the calibration of finely-tuned ML models. Notable methods include histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002), conformal prediction (Vovk et al., 2005), Platt scaling (Platt et al., 1999), and temperature scaling (Guo et al., 2017b). These techniques generally rely on the existence of a validation set for the post-processing of model predictions.

Recent research has started to focus on enhancing calibration in deep learning models. These include strategies such as augmentation-based training (Thulasidasan et al., 2019), calibration in neural machine translation (Kumar & Sarawagi, 2019), neural stochastic differential equations (Kong et al., 2020), self-supervised learning (Hendrycks et al., 2019) ensemble methods (Lakshminarayanan et al., 2017), and even providing statistical assurance for calibration in black-box models (Angelopoulos et al., 2021). Nevertheless, research addressing model calibration in FL settings remains limited.

Calibration in FL Settings While the importance of calibration in FL is being recognized, existing approaches primarily focus on performance improvement without considering reliability. Luo et al. (2021) demonstrated that classifier calibration significantly boosts FL model performance, albeit focusing on adjusting weights on IID data, a process different from aligning confidence with observed frequency. Zhang et al. (2022a) introduced FedLC to enhance model accuracy through a calibration-inspired cross-entropy loss. However, calibration in FedLC is defined as the error rate per class, which is different from our definition. Achituve et al. (2021) proposed pFedGP, a personalized approach that, although not designed for FL calibration, achieves some level of empirical calibration efficacy. But, its specialized nature hinders wider applicability. Closely related to our work is MD-TS (Yu et al., 2022). It employs domain-specific temperature scaling and a predictive linear regression model. Nevertheless, it relies on the existence of global validation sets, which might be difficult to prepare in practice and risk privacy leakage.

Orthogonal to our work, the field of PFL tailors models to

client-specific needs (Tan et al., 2023; Arivazhagan et al., 2019; Deng et al., 2020). While PFL addresses the issue of non-IID client data distributions, our focus lies in federated model calibration. Notably, FedCal can be integrated with existing PFL frameworks.

3. Preliminaries

3.1. Basics of Model Calibration

In a K-classification task with $\mathcal{Y} \equiv \{1, ..., K\}$, we aim to train a model $\boldsymbol{\theta} : \mathcal{X} \mapsto \mathbb{R}^K$ that predicts the labels $y \in \mathcal{Y}$ corresponding to the input $\boldsymbol{x} \in \mathcal{X}$. Here, $\boldsymbol{\theta}(\boldsymbol{x}_i)$ outputs the score, a quantity that represents a proper probabilistic estimation often obtained by applying an additional sigmoid layer σ as $f(\boldsymbol{x}_i) = \sigma(\boldsymbol{\theta}(\boldsymbol{x}_i))$, where $f : \mathcal{X} \mapsto \Delta^{K-1}$. Δ^{K-1} denotes a $(K-1)$ simplex such that for any class k, $f_k(\boldsymbol{x}_i) \in [0, 1]$ and $\sum_{k \in \mathcal{Y}} f_k(\boldsymbol{x}_i) = 1$.

Calibration (Guo et al., 2017b), in this framework, refers to the extent to which these predicted probabilities $f_k(\boldsymbol{x}_i)$ reflect the true conditional probability $p(y = k|\boldsymbol{x})$. While calibrating predictions for every class is an ideal goal, it can be challenging and impractical in real applications (Kumar et al., 2019). Often, the more achievable task of top-label calibration is prioritized, which focuses on ensuring the predicted probability of the most likely class aligns with its true probability. For a comprehensive overview of different calibration types and their nuances, please refer to (Kumar et al., 2019; Zhao et al., 2021).

We denote the predicted label $\hat{f}(\boldsymbol{x}_i)$ by a model as $\hat{f}(\boldsymbol{x}_i) \equiv \max f_k(\boldsymbol{x}_i)$ and $\hat{y}(\boldsymbol{x}) = \arg\max_{k \in \mathcal{Y}} f_k(\boldsymbol{x}_i)$. Then, the top-level calibration error can be defined as:

Definition 3.1. (Calibration Error). The calibration error $\text{CE}(f)$ of f is given by:

$$\text{CE}(f) = (\mathbb{E}[(\mathbb{P}(y = \hat{y}(\boldsymbol{x}_i)|\hat{f}(\boldsymbol{x}_i)) - \hat{f}(\boldsymbol{x}_i))^2])^{\frac{1}{2}}. \quad (1)$$

The lower the CE value, the better the calibration. A perfectly calibrated model achieves a CE value of 0. A widely adopted empirical measure of CE is the Expected Calibration Error (ECE) (Naeini et al., 2015). ECE measures calibration by averaging the difference between the predicted probabilities and the actual accuracy within each confidence bin. In our work, ECE is adopted as the metric for reporting calibration errors.

Definition 3.2. (Expected Calibration Error). Giving a partition c_m of the unit interval $[0, 1]$ and the buckets $B_m = \{i : c_{m-1} < \hat{f}(\boldsymbol{x}_i) \leq c_m\}$, ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{conf}_m - \text{acc}_m|, \quad (2)$$

where $\text{acc}_m = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}(\boldsymbol{x}_i) = y)$ and $\text{conf}_m = \frac{1}{|B_m|} \sum_{i \in B_m} f_k(\boldsymbol{x}_i)$.

3.2. Calibration by Scaling

Modern deep models tend to achieve poor calibration (Guo et al., 2017b). To address this issue, researchers have explored two main approaches: 1) post-hoc calibration, and 2) architecture/training modification for inherently calibrated models (Wu & Gales, 2021). We focus on the post-hoc calibration approach due to its flexibility and ease of integration with existing training paradigms. Post-hoc mapping methods involve employing a scaling function ϕ trained on an auxiliary dataset to adjust $f_k(\boldsymbol{x}_i)$. Here, we review a simple yet effective calibration method, Temperature scaling (Guo et al., 2017b), which resales model outputs using a singular temperature parameter T as:

$$\phi(T) \circ f(\boldsymbol{x}_i) \equiv \sigma\left(\frac{\boldsymbol{\theta}(\boldsymbol{x}_i)}{T}\right). \quad (3)$$

The optimal temperature value, T, is determined by minimizing the negative log-likelihood (NLL) on a validation set $\mathcal{D}_{val}$ as:

$$\begin{aligned} &\min_T \text{NLL}\left(y, \sigma\left(\frac{\boldsymbol{\theta}(\boldsymbol{x}_i)}{T}\right)\right) \\ &= -\sum_{i=1}^{|\mathcal{D}_{val}|} \sum_{j=1}^{K} \mathbb{1}(y_i = j) \cdot \log \sigma_j\left(\frac{\boldsymbol{\theta}(\boldsymbol{x}_i)}{T}\right). \end{aligned} \quad (4)$$

4. Theoretical Basis of FedCal

4.1. The Need for Local and Global Calibration

Calibrating FL models introduces distinct challenges due to the inherent non-IID nature of data across clients. Such heterogeneous data distributions significantly degrade the calibration performance of models evaluated on both local and global datasets (Figure 2). Moreover, the privacy requirements inherent to FL often preclude access to a centralized auxiliary validation dataset, rendering traditional calibration methods inapplicable. Figure 2 also highlights the discrepancies between calibration performance on clients' local datasets and the aggregated global dataset. This observation compels us to re-evaluate traditional notions of calibration within the context of FL. We posit that both local and global calibration are essential for building reliable FL models.

Consider the illustrative hospital example from the introduction, where treatment decisions hinge on the model's prediction and the associated confidence, interpreted as cancer probability. The decision minimizes the combined surgical and conservative treatment risks: $risk_{\text{surgery}} \times \mathbb{P}(\text{cancer}) + risk_{\text{conservative treatment}} \times \mathbb{P}(\text{benign tumor})$. As posited by Zhao et al. (2021), calibration directly impacts decision-making efficacy. Hospital A's specialized model, potentially

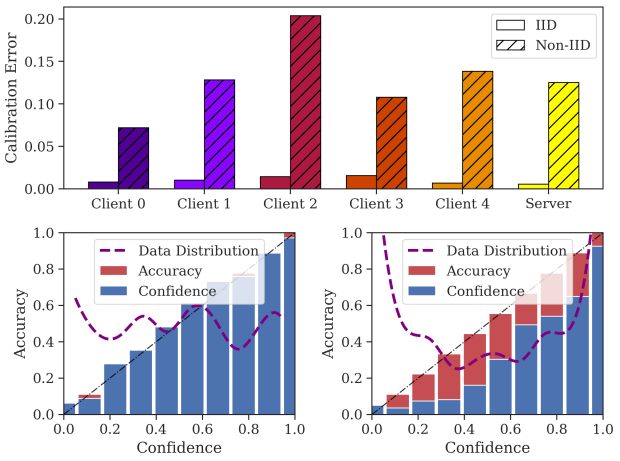

Figure 2. Impact of Non-IID Data Distribution on Client and Server Calibration in FL. The top plot shows the calibration error of five clients trained on the MNIST dataset with a Multilayer Perceptron (MLP) model under IID and non-IID distributions. We observe that the non-IID client exhibits significantly higher calibration error compared to the IID clients, and that calibration error can vary significantly across clients and even at the server due to the skewed data distribution. The two bottom plots depict the model reliability for a single client trained under IID and non-IID conditions. The purple dashed line represents the normalized class density of a client, and the gray dashed line represents perfect calibration (i.e., confidence aligns exactly with accuracy). In the IID case (left plot), the model is well-calibrated, with the confidence closely matching the accuracy throughout the range. In contrast, the non-IID case (right plot) reveals severe under-confidence.

well-calibrated for its specific patient population, might predict a higher cancer probability, leading to a more aggressive treatment approach. Conversely, Hospital B's general model, trained on a wider range of patients with different diseases, might predict a lower probability, favoring a more conservative approach. While local expertise and specialization hold value, neglecting global calibration across the FL network poses significant risks.

Maintaining consistent global calibration ensures that decisions are guided by comparable risk assessments across all participating clients. This, in turn, minimizes the potential for inequitable recommendations, especially as new clients with potentially divergent data distributions join FL. Here, we formalize the notions of local and global calibration.

Definition 4.1. (Local calibration error and global calibration error). Consider an FL system comprising C clients, each possessing a local dataset $\mathcal{D}_c = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N_c}$ drawn from a client-specific distribution $\mathbb{P}_c(\boldsymbol{x}, y)$. The local calibration error $\mathrm{CE}_c(f)$ is:

$$\mathrm{CE}_c(f) = \left(\mathop{\mathbb{E}}_{\mathbb{P}_c(\boldsymbol{x},y))} [(\mathbb{P}(y = \hat{y}(\boldsymbol{x}_i)|\hat{f}(\boldsymbol{x}_i)) - \hat{f}(\boldsymbol{x}_i))^2] \right)^{\frac{1}{2}}. \tag{5}$$

Similarly, the global calibration error $\overline{\mathrm{CE}}(f)$, conceptualized as the expected calibration error over unseen data sampled from a client-agnostic distribution $\mathbb{P}(\boldsymbol{x}, y)$, is:

$$\overline{\mathrm{CE}}(f) = \left(\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x},y)} [(\mathbb{P}(y = \hat{y}(\boldsymbol{x}_i)|\hat{f}(\boldsymbol{x}_i)) - \hat{f}(\boldsymbol{x}_i))^2] \right)^{\frac{1}{2}}. \tag{6}$$

Empirically, local calibration error can be approximated by evaluating ECE on the local dataset. Since the global data distribution is not directly observable, it is presumed that samples $(\boldsymbol{x}_i, y_i)$ in global dataset follows the same distribution of the pooled dataset $\cup_{c=1}^{C} \mathcal{D}_c$ and the global calibration error can be assessed using a reserved test set, sampled from the pooled dataset.

4.2. Label Skew Leads to Poor Calibration

In this paper, we focus on one of the most representative settings of non-IIDness, which is *label skew* (Luo et al., 2019; Lyu et al., 2022; Zhang et al., 2022b). Note that, while we focus on label skew, various forms of non-IID data exist (refer to (Hsieh et al., 2020) for more details).

Definition 4.2. (Label Skew). The label distribution across the clients is skewed. If the local distribution is rewritten as $\mathbb{P}_c(\boldsymbol{x}, y) = \mathbb{P}_c(\boldsymbol{x}|y)\mathbb{P}_c(y)$, label skew means, for any two clients c_1 and c_2, we have: (1) $\mathbb{P}_{c1}(y) \neq \mathbb{P}_{c2}(y)$ if $c_1 \neq c_2$; and (2) $\mathbb{P}_{c1}(\boldsymbol{x}|y) = \mathbb{P}_{c2}(\boldsymbol{x}|y)$.

We denote the empirical risk minimizer over the global distribution as $f^*(\boldsymbol{x}) \coloneqq \arg_f \min \xi(y, f(\boldsymbol{x}))$, where ξ is the cross entropy loss. The global model obtained after r rounds by averaging local updates is denoted as $f^r(\boldsymbol{x})$. Affected by the heterogeneous data, the local objectives of different clients are generally not identical and might not share the same risk minimizer. Consequently, even as all clients start from the same global model, local updates will steer the model towards the minima of local objectives (known as client drift (Charles & Konečnỳ, 2021)). This divergence implies that averaging local updates via FedAvg results in a model that is different from the global minimizer, i.e., $f^*(\boldsymbol{x}) \neq f^r(\boldsymbol{x})$ (Wang et al., 2020; Karimireddy et al., 2020). Expanding upon this fundamental understanding, we posit that, given a bounded distribution divergence between local and global distributions, client drift inherently imposes an asymptotic lower bound on the global calibration error.

Assumption 4.3. (Bounded discrepancy between local and global label distribution). We assume that the discrepancy

between the data distributions at each client and the global distribution, measured by the Kullback–Leibler divergence D_{KL}, is bounded. Specifically, the maximum divergence between the local data distribution $\mathbb{P}_c(y)$ and the global distribution $\mathbb{P}(y)$ does not exceed a value G. Mathematically, this is expressed as:

$$\sup_{c\in\{1,\cdots,C\}} \mathrm{D}_{KL}(\mathbb{P}_c(y) \parallel \mathbb{P}(y)) \leq G, \tag{7}$$

where C is the total number of clients.

For clarity and ease of understanding, we concentrate the discussion on the binary classification scenario. This does not reduce the generality of our findings as the problem of calibrating multi-class models can be effectively reduced to the binary case by letting the model output a probability corresponding to its top prediction, and the label represents whether the prediction is correct or not (Kumar et al., 2019).

Theorem 4.4. *(Lower bound of global calibration error). Consider the scenario where the discrepancy between the local and global label distributions is bounded by G, as stated in Assumption 4.2. Let R represent the number of FL communication rounds involving more than two clients. Jointly considering the established assumptions detailed in Appendix A, under these conditions, there exists a μ-convex objective function for which the resulting global model $f^R(\boldsymbol{x})$ after R rounds has a calibration error asymptotically bounded by:*

$$\overline{CE}(f^R) \geq \Omega\left(\frac{\sqrt{\frac{1}{2}G}}{\mu R^2}\right). \tag{8}$$

The formal proof of this theorem is presented in Appendix A. Here, we provide an illustrative sketch. Firstly, we establish that the global risk minimizer $f^*(\boldsymbol{x})$ also minimizes the calibration error (i.e., $\overline{\mathrm{CE}}(f^*(\boldsymbol{x})) = 0$). Consistent with the premises set forth in (Zhang et al., 2022b; Wang et al., 2021), we posit that the features extracted from the same class (inputs to the final layer) exhibit a high degree of similarity, a notion empirically observed and verified in (Wang et al., 2021). Based on this, we analyse how the bounded discrepancy between label distributions propagates to bounded gradient dissimilarity using Pinsker's Inequality. Lastly, based on Karimireddy et al. (2020) which relates the dissimilarity of gradient to the difference between risk minimizer $[f^R(\boldsymbol{x}) - f^*(\boldsymbol{x})]$, we draw the proof to its conclusion.

Theorem 4.4 elucidates that despite constraining the variance in clients' label distributions, the global model calibration error can still asymptotically lower bounded, which corroborates our empirical findings depicted in Figure 2 and Figure 1. While prevailing calibration techniques (e.g., temperature scaling) offer simplicity and effectiveness, they require access to a global validation dataset. This is not only impractical in many real-world scenarios, but also raises significant privacy concerns.

FedCal is a novel paradigm distinctly orthogonal to the conventional accuracy-centric FL approach. Formally, it is a task that aims to minimize both CE_i and $\overline{\mathrm{CE}}$ without access to $\mathcal{D}_{val}$ sampled from $\mathcal{D}$.

5. The Proposed FedCal Approach

While local calibration is relatively easy as the FL client has direct access to its local data making established calibration methods like temperature scaling applicable, achieving global calibration presents a significant technical challenge. To this end, we frame the problem of FedCal as aggregating local scalers into a global scaler, formulated as $\overline{\phi} = \mathrm{Agg}(\phi_{c_{i=1}}^{C})$. The success of $\overline{\phi}$ hinges on two key aspects: 1) the scaler architecture, and 2) the aggregation strategy. The following properties are desirable for these two components:

1. The scaler must possess robust generalization capabilities to handle potential discrepancies between local and global data distributions.
2. The processes of scaling and calibration should maintain model accuracy.
3. The scaler should also be "aggregatable", meaning that the aggregated version should perform well not only on a global scale, but also on the local scale.
4. The scaler aggregation strategy should not require direct access to local data distributions.

5.1. Scaler Architecture Design

To achieve Property 1, FedCal is equipped with a multilayer perceptron (MLP) with substantially more parameters than traditional minimal-parameter methods to enhance its generalization capability. The MLP processes the original output logits $\boldsymbol{\theta}(x)$, transforming them into accurate probabilities, particularly under conditions of significant local and global data distribution discrepancies.

However, a trade-off exists between Property 1 and Property 2. A more complex model, while capable of learning intricate mappings, risks overfitting the validation datasets. This can alter the original order of the model logits, leading to a decrease in top-k accuracy (Figure 3). Here order-preserving means that for any two classes j and k, if $f_j^R(\boldsymbol{x}_i) > f_k^R(\boldsymbol{x}_i)$, the scaled outputs must follow the same order (i.e., $\phi(f_j^R(\boldsymbol{x}_i)) > \phi(f_j^R(\boldsymbol{x}_i))$). To achieve order preservation, we incorporate the order-preserving technique from (Rahimi et al., 2020) into the scaler design. Formally,

the local scaler ϕ_c parameterized by an MLP $\boldsymbol{\pi}_c$ remaps the model output as:

$$\phi_c(\boldsymbol{\phi}) \circ f(\boldsymbol{x}_i) \equiv \phi_c(\boldsymbol{\theta}(\boldsymbol{x}_i); \boldsymbol{\pi}_c). \tag{9}$$

$\boldsymbol{\pi}_c$ is optimized by minimizing the NLL loss on a local dataset $\mathcal{D}_c$ similar to Eq. (4). Although training is required, the scaler can be effectively implemented in a post-hoc manner. Specifically, the local scaler undergoes training only after the FL model completes its local training phase.

5.2. Aggregation Strategy Design

One advantage of our MLP-based scaler design is its inherent permutation symmetries. These symmetries facilitate the aggregation of two distinct MLPs through linear mode connectivity (Ainsworth et al., 2022; Entezari et al., 2021; Nguyen et al., 2021). Linear mode connectivity implies that for two MLPs, $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, after aligning one of them (e.g., $\boldsymbol{\pi}_i$) through a weights permutation $\mathbf{M}(\boldsymbol{\pi}_i)$, they can be linearly combined in their parameter space as $\boldsymbol{\pi}^* = \lambda\boldsymbol{\pi}_i + (1-\lambda)\mathbf{M}(\boldsymbol{\pi}_j)$. The loss function of the new parameters $\mathcal{L}(\boldsymbol{\pi}^*)$ closely approximate a weighted sum of the individual losses as $\mathcal{L}(\boldsymbol{\pi}^*) \approx \lambda\mathcal{L}(\boldsymbol{\pi}_i) + (1-\lambda)\mathcal{L}(\boldsymbol{\pi}_j)$ for any $\lambda \in [0, 1]$. This ensures that our scaler achieves Property 3 and Property 4. The resulting global scaler, $\overline{\phi}(\boldsymbol{\pi})$, achieves a low local calibration error and is robust to the choice of aggregation weight λ. For practical implementation, we adopt the Weight Matching algorithm (Ainsworth et al., 2022) which does not require access to local data (Algorithm 2 in Appendix D).

5.3. The Combined Framework

FedCal (Algorithm 1) enhances the FedAvg procedure to improve model calibration. Initially, FL clients are provided with the global scaler parameters as a baseline for local scaler training. To establish a coherent relationship between their local and the global scaler, clients perform Weight Matching (Algorithm 2) to achieve optimal alignment. After local model updates, clients refine their scalers to minimize local calibration errors on their validation datasets. Upon receiving the updated models and scalers from the clients, the FL server performs model aggregation via FedAvg and averages the local scalers to form an updated global scaler.

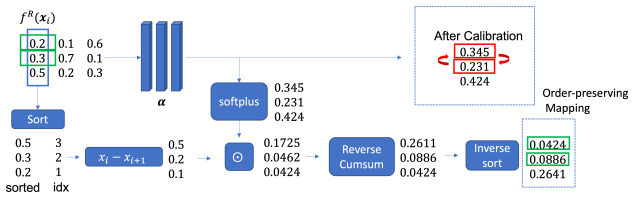

Figure 3. Impact of the Order-Preserving Network. Without order preservation, α can represent arbitrary mappings, potentially altering the predicted class ordering (highlighted in red). The working principles of order-preserving networks are in Appendix B.

Algorithm 1 FedCal

Input: C clients, local epochs E, learning rate η, global model weights $\boldsymbol{\theta}^0$, global scaler weights $\overline{\boldsymbol{\pi}}^0$
for each round $t = 1, 2, \ldots, R$ **do**
  Server selects a subset of m clients S_t
  for each client $c \in S_t$ **in parallel do**
    $\boldsymbol{\theta}_c^{t+1}, \boldsymbol{\pi}_c^{t+1} \leftarrow \text{ClientUpdate}(c, \boldsymbol{\theta}^0, \overline{\boldsymbol{\pi}}^0)$
  end for
  $\boldsymbol{\theta}^{t+1} \leftarrow \sum_{c=1}^{m} \frac{n_c}{N} \boldsymbol{\theta}_c^{t+1}$ {Aggregate updated models}
  $\overline{\boldsymbol{\pi}}^{t+1} \leftarrow \frac{1}{m} \sum_{c=1}^{m} \boldsymbol{\pi}_c^{t+1}$ {Aggregate updated scalers}
end for

procedure ClientUpdate($(c, \boldsymbol{\theta}^0, \overline{\boldsymbol{\pi}}^0)$):
$\mathbf{M} \leftarrow \text{WeightMatching}(\overline{\boldsymbol{\pi}}^0, \boldsymbol{\pi}_c)$
$\boldsymbol{\pi}_c \leftarrow \mathbf{M}(\boldsymbol{\pi}_c)$ {Permute to Align}
$\mathcal{B} \leftarrow$ (split $\mathcal{D}_c$ into batches of size B)
for each local epoch i from 1 to E **do**
  for batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w - \eta \nabla l(w; b)$ {Update model weights}
  end for
end for
$\boldsymbol{\pi}_c \leftarrow \text{TrainScaler}(\boldsymbol{\pi}_c, \mathcal{D}_c)$
return w, $\boldsymbol{\pi}_c$ to server

Notably, rather than adopting a purely post-hoc approach, where scaler training and aggregation occur only after the completion of FedAvg, FedCal opts for periodic synchronization of scaler updates. This strategy enhances the scalers' ability to learn more general mappings and boost overall aggregation efficacy. FedCal incurs additional communication overhead, equivalent to the scaler parameter count, and extra computational load due to weight matching. Nonetheless, given that the scaler parameter count is relatively modest compared to that of the primary classifier, the trade-off is considered acceptable.

Regarding privacy, FedCal necessitates the sharing of scaler parameters, which helps prevent the disclosure of specific data distribution details (such as logits, quantiles, etc.). Moreover, sharing parameters aligns with the standard paradigm in federated learning, making privacy-preserving techniques, such as Homomorphic Encryption (Zhang et al., 2020; Hardy et al., 2017) and Differential Privacy (Wei et al., 2020), could be integrated to augment privacy protection further. However, the incorporation and exploration of these techniques fall outside the scope of our current research.

6. Experimental Evaluation

To evaluate the performance of FedCal, we experimentally compare it with five baseline methods over four benchmark datasets with different degrees of non-IIDness.

6.1. Experiment Setup

We conduct our experiments on the following widely adopted benchmark datasets: 1) MNIST (Deng, 2012), 2) SVHN (Netzer et al., 2011), 3) CIFAR-10 and 4) CIFAR-100 (Krizhevsky et al., 2009). For MNIST and SVHN, we utilize the standard CNN as the base model, while for CIFAR-10 and CIFAR-100, we adopt ResNet-14 (He et al., 2016) and ResNet-32 (He et al., 2016) as the base model, respectively. Prior to distribution among FL clients, each dataset is pre-processed.

To replicate non-IID conditions typical in real-world settings, we follow the Distribution-based Label Skew method (Yurochkin et al., 2019; Li et al., 2020a; Zhang et al., 2022b), which uses a Dirichlet distribution ($p_c \sim \text{Dir}(\beta)$) to allocate class samples across clients. The parameter β modulates label skew, with higher values indicating more pronounced non-IIDness.

We carried out our experiments using PyTorch on a single NVIDIA A100 GPU, which has 40 GB of memory. In our FL setup, we include 20 FL clients. In each FL training round, we randomly select 5 of them to participate. Each local training round consists of 3 epochs. For the local updates, we adopt the SGD optimizer with a learning rate of 0.01, and a local batch size of 256. For the implementation of FedCal, the default configuration of the proposed scaler is an MLP with an activation structure of K-64-64-K, where K represents the number of classes. We set the maximum number of global epochs to 100.

6.2. Comparison Baselines

We compare FedCal with standard scaler designs and FL aggregation methods, including:

1. UNCAL.: FL without model calibration.
2. VAL. TS: uses a temperature scaler on a global validation set, considered as the performance upper bound.
3. ENS.: Implements Deep Ensemble's direct averaging of scaled probabilities. Despite its simplicity, it's known for robust uncertainty quantification (Lakshminarayanan et al., 2017; Lee et al., 2015).
4. AVGT: Extends ENS by averaging temperature parameters of individual scalers. Both ENS and AVGT are efficient, but have limitations when facing non-IID distributions (Rahaman et al., 2021; Abe et al., 2022).
5. LR-TS: Adapts MD-TS (Yu et al., 2022) for FL settings. It estimates the scaling temperature from model outputs, thereby achieving calibration without a shared validation dataset.

We adopt test accuracy and the global ECE as the evaluation metrics. For more details, please refer to Appendix C.

6.3. Results and Discussion

Table 6.1 reports the global ECE results. FedCal consistently achieves the lowest calibration error across all datasets. When compared to UNCAL and the second-best approach without requiring a global validation set ENS., FedCal significantly reduces the calibration error by 63.06% and 47.66%, respectively. It is also important to note that an increase in non-IIDness tends to worsen the calibration error, which corroborates our theoretical analysis. However, FedCal demonstrates stronger robustness compared to other baselines, even as non-IIDness increases. In most cases, ENS emerges as the second-best, reinforcing the notion that deep ensembles are effective in calibration, particularly when non-IIDness is moderate.

Under MNIST, we examine how calibration error changes with increasing non-IIDness (as indicated by $-\log\beta$). The results are illustrated in Figure 4, which also contains the results of parts of our ablation studies. The top left subplot indicates the average local calibration error, while the bottom left one shows the maximum local calibration error. It can be observed that despite the increase in non-IIDness, all scaling methods manage to maintain low calibration errors.

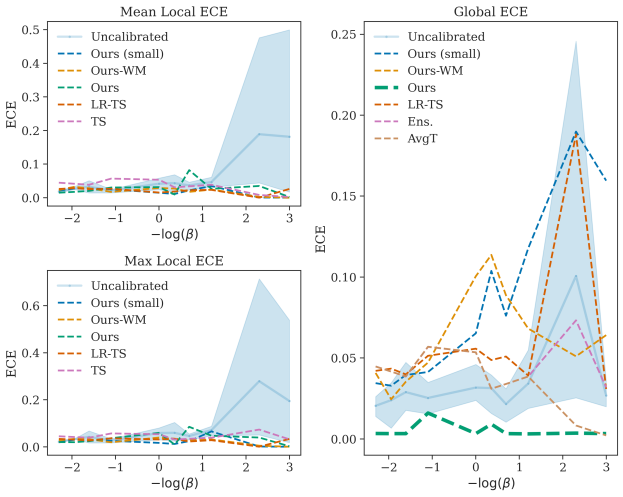

Figure 4. Local and global calibration errors as non-IIDness increases. [Top left]: the average local calibration errors. [Bottom left]: the maximum local calibration errors. [Right]: the global calibration error.

In contrast, the global calibration error Figure 4 (right sub-

DATASETS	SETTINGS		GLOBAL ECE %					
	β	ACC %	UNCAL.	VAL. TS	ENS.	AVGT.	LR-TS.	FedCal
MNIST-CNN	β =1	95.72 ± 2.3	0.5 ± 0.12	0.34 ± 0.08	0.43 ± 0.10	**0.41** ± 0.10	0.71 ± 0.17	0.45 ± 0.11
	β =0.5	93.03 ± 2.2	1.00 ± 0.24	0.79 ± 0.19	1.32 ± 0.32	0.72 ± 0.17	0.71 ± 0.17	**0.47** ± 0.11
	β =0.3	91.34 ± 2.2	1.6 ± 0.38	0.74 ± 0.18	1.34 ± 0.32	0.92 ± 0.22	1.57 ± 0.38	**0.61** ± 0.15
	β =0.1	81.01 ± 1.9	4.6 ± 1.10	0.74 ± 0.18	2.13 ± 0.51	3.17 ± 0.76	5.51 ± 1.32	**1.35** ± 0.32
SVNH-CNN	β=1	93.24 ± 2.2	1.12 ± 0.27	0.32 ± 0.08	0.43 ± 0.10	0.52 ± 0.12	0.69 ± 0.17	**0.44** ± 0.11
	β=0.5	85.13 ± 2.0	1.30 ± 0.31	0.51 ± 0.12	0.99 ± 0.24	1.03 ± 0.25	1.31 ± 0.31	**0.89** ± 0.21
	β=0.3	85.14 ± 2.0	4.56 ± 1.09	1.21 ± 0.29	1.58 ± 0.38	3.59 ± 0.86	3.51 ± 0.84	**0.77** ± 0.18
	β=0.1	79.23 ± 1.9	7.81 ± 1.87	1.25 ± 0.30	1.56 ± 0.37	6.12 ± 1.47	32.03 ± 7.69	**1.25** ± 0.30
CIFAR10-RESNET14	β=1	65.54 ± 1.6	7.61 ± 1.82	3.61 ± 0.87	5.42 ± 1.30	7.82 ± 1.88	3.42 ± 0.82	**3.71** ± 0.89
	β=0.5	60.21 ± 1.4	5.63 ± 1.35	4.12 ± 0.99	6.28 ± 1.51	8.34 ± 2.00	4.23 ± 1.01	**4.61** ± 1.10
	β=0.3	57.31 ± 1.4	9.81 ± 2.35	3.15 ± 0.76	8.16 ± 1.96	11.25 ± 2.70	11.12 ± 2.67	**4.91** ± 1.18
	β=0.1	48.05 ± 1.2	12.48 ± 2.99	3.51 ± 0.84	8.87 ± 2.13	13.34 ± 3.20	14.32 ± 3.44	**4.51** ± 1.08
CIFAR100-RESNET32	β=1	41.24 ± 1.0	22.45 ± 5.39	4.13 ± 0.99	16.34 ± 3.92	20.25 ± 4.86	19.21 ± 4.61	**7.41** ± 1.78
	β=0.5	30.01 ± 0.7	20.45 ± 4.91	4.19 ± 1.01	17.53 ± 4.21	29.21 ± 7.01	18.35 ± 4.40	**7.50** ± 1.80
	β=0.3	22.21 ± 0.5	25.71 ± 6.17	4.09 ± 0.98	18.93 ± 4.55	30.45 ± 7.31	22.92 ± 5.50	**8.91** ± 2.14
	β=0.1	20.8 ± 0.5	31.78 ± 7.63	4.82 ± 1.16	20.48 ± 4.92	30.25 ± 7.26	37.32 ± 8.96	**10.72** ± 2.57

Table 1. Comparison of global ECE across datasets with varying levels of non-IIDness.

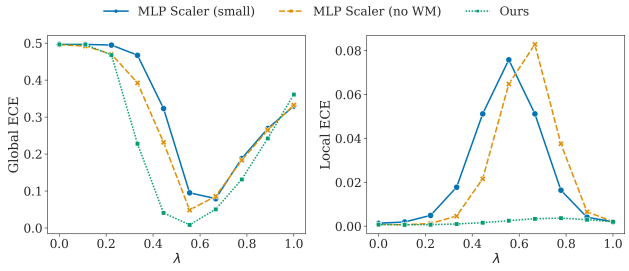

Figure 5. Local and global ECE vs. aggregation weights.

plot) behaves differently. It can be observed that FedCal consistently outperforms other methods without a substantial increase in global ECE as non-IIDness grows. However, without performing weight matching (OURS-WM), FedCal does not exhibit the same robustness. Similarly, reducing the MLP scaler size from 64 neurons to 8 (OURS-SMALL) results in a reduction in the generalization capabilities of FedCal. This underscores the importance of the synergy of the integrated scaler and the aggregation strategy in the FedCal design.

6.4. Ablation Studies

In our ablation studies, we demonstrate the effectiveness of MLP scalers and weight matching in FedCal (Figure 5). Using a synthetic dataset with high non-IIDness and two clients, we find that while individual local scalers are inadequate globally, their weighted aggregation significantly improves overall calibration. Increasing the MLP scaler size also enhances generalization. Directly applying global scalers locally leads to good individual calibration, but causes substantial local errors without a proper weight matching mechanism. The results highlight the importance of weight matching in federated calibration.

In FedCal, the order-preserving technique is a crucial component. We evaluate its significance through ablation studies conducted on the MNIST dataset, where β is fixed at 0.5. The results are shown in Table 6.4.

It can be observed that calibration methods like ENS, AVGT and LR-TS, which modify model logits, have no detrimental effect on top-3 accuracy. In contrast, removing the order-preserving element from FedCal (denoted as OURS W/O OP) results in a slight reduction in global ECE. Yet, this comes at the expense of significantly lower top-3 accuracy. This outcome suggests that, while FedCal can enhance test accuracy to a certain extent, excluding the order-preserving part of the design negatively affects the model's ability to accurately rank its top predictions. Conversely, while imposing order-preserving constraints does introduce some limitations, it only leads to a minor increase in global ECE.

	GLOBAL ECE	TOP-3 ACCURACY
WITHOUT SCALER	4.6%	92.3%
ENS.	2.13%	-
AVGT	3.17%	-
LR-TS	5.51%	-
OURS W/O OP	**1.32%**	87.91%
OURS	1.71%	**92.3%**

Table 2. Top-3 Accuracy and Calibration Error on the MNIST dataset with a CNN with $\beta = 0.5$.

We also evaluated FedCal in combination with methods designed to improve federated learning performance under non-IID data distributions. Specifically, we testedFedCal together with FedProx (Li et al., 2020b), which adds regularization to encourage local updates to stay closer to the global model under non-IID conditions. As shown in Ta-

ble 6.4, while FedProx improves accuracy, it still suffers from significant calibration errors. However, combining it with FedCal substantially reduced global calibration errors. This demonstrates that FedCal effectively complements existing non-IID approaches (e.g., FedProx) by enhancing calibration.

β	ACC %	GLOBAL ECE %	
		FEDPROX	FEDPROX + FedCal
β =1	94.32	0.38	0.40
β = 0.5	94.07	0.57	0.42
β = 0.3	90.25	1.32	0.58
β = 0.1	89.93	3.19	1.07

Table 3. Global calibration errors on the MNIST-CNN for FedAvg, FedProx, and FedProx + FedCal under varying non-IID levels

7. Conclusions and Future Work

In this paper, we provide both theoretical and empirical insights into the necessity of simultaneously achieving global calibration and local calibration in FL settings. The proposed FedCal approach is designed with a sophisticated multi-layer perceptron (MLP) scaler alongside the order-preserving technique to effectively handle the challenges posed by non-IID data distributions commonly encountered in real-world FL applications. Extensive experiments demonstrate that FedCal surpasses existing calibration methods by significantly reducing global model calibration error without compromising model accuracy.

In subsequent research, we plan to investigate the complex dynamics between order-preserving techniques and weight matching strategies. We will also develop theories about the relations between calibration effectiveness and model size.

Acknowledgements

This research is supported, in part, by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (No. AISG2-RP-2020-019); the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore; the Natural Sciences and Engineering Research Council of Canada (NSERC); and Compute Canada Research Platform.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abe, T., Buchanan, E. K., Pleiss, G., Zemel, R., and Cunningham, J. P. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660, 2022.

Achituve, I., Shamsian, A., Navon, A., Chechik, G., and Fetaya, E. Personalized federated learning with gaussian processes. *Advances in Neural Information Processing Systems*, 34:8392–8406, 2021.

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Charles, Z. and Konečnỳ, J. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Dash, B., Sharma, P., and Ali, A. Federated learning for privacy-preserving: A review of pii data analysis in fintech. *International Journal of Software Engineering & Applications (IJSEA)*, 13(4), 2022.

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C.-S., et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27 (10):1735–1743, 2021.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Du, Z., Wu, C., Yoshinaga, T., Yau, K.-L. A., Ji, Y., and Li, J. Federated learning for vehicular internet of things: Recent advances and open issues. *IEEE Open Journal of the Computer Society*, 1:45–61, 2020.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017a.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017b.

Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., and Thorne, B. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Kong, L., Sun, J., and Zhang, C. Sde-net: Equipping deep neural networks with uncertainty estimates. *arXiv preprint arXiv:2008.10546*, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Canada, 2009.

Kumar, A. and Sarawagi, S. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

Li, Q., Wen, Z., and He, B. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4642–4649, 2020a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Li, Y., Tao, X., Zhang, X., Liu, J., and Xu, J. Privacy-preserved federated learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8423–8434, 2021.

Liu, R., Xing, P., Deng, Z., Li, A., Guan, C., and Yu, H. Federated graph neural networks: Overview, techniques, and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Long, G., Tan, Y., Jiang, J., and Zhang, C. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pp. 240–254. Springer, 2020.

Lu, C. and Kalpathy-Cramer, J. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

Luo, J., Wu, X., Luo, Y., Huang, A., Huang, Y., Liu, Y., and Yang, Q. Real-world image datasets for federated learning. *arXiv preprint arXiv:1910.11089*, 2019.

Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., and Feng, J. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Philip, S. Y. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Nguyen, Q. N., Bréchet, P., and Mondelli, M. When are solutions connected in deep networks? *Advances in Neural Information Processing Systems*, 34:20956–20969, 2021.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. Conformal prediction for federated uncertainty quantification under label shift. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27907–27947. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/plassier23a.html.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Rahaman, R. et al. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.

Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33:13456–13467, 2020.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Wang, L., Xu, S., Wang, X., and Zhu, Q. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10165–10173, 2021.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Wu, X. and Gales, M. Should ensemble members be calibrated? *arXiv preprint arXiv:2101.05397*, 2021.

Yu, Y., Bates, S., Ma, Y., and Jordan, M. Robust calibration with multi-domain temperature scaling. *Advances in Neural Information Processing Systems*, 35:27510–27523, 2022.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pp. 7252–7261. PMLR, 2019.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 609–616, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., and Liu, Y. {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pp. 493–506, 2020.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022a.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022b.

Zhao, S., Kim, M., Sahoo, R., Ma, T., and Ermon, S. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

A. Proof of Theorem 4.4

Proof. First of all, we show that if f^* is the risk minimizer of NLL loss over $\mathbb{P}(x, y)$, it also minimizes the calibration error. if f^* is a risk minimizer of NLL, $f^*(x) = p(y = 1|x)$. Then $\mathbb{E}[y|f^*(x)] = f^*(x)$, thus f^* also minimize the the calibration error as $\text{CE}(f^*) = 0$.In binary classification, we can rewrite the calibration error as

$$\begin{aligned} \text{CE}(f) &= \left(\mathbb{E}_{\mathbb{P}(x,y)}\left[|\mathbb{E}[y|f(x)] - f(x)|^2\right]\right)^{\frac{1}{2}} \\ &= \mathbb{E}_{P(x)}\left[P(y=1|x)\cdot(1-f(x))^2 + P(y=0|x)\cdot(f(x))^2\right]. \end{aligned} \tag{A.1}$$

This equation relates the calibration error $\text{CE}(f^n) - \text{CE}(f^*) = \text{CE}(f^n)$ with $f^R(\boldsymbol{x}) - f^*(\boldsymbol{x})$.

Theorem A.1. *(Adopted from (Karimireddy et al., 2020).) When the gradients of the local loss function and global function have bounded gradient dissimilarity, stated as*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla\text{NLL}_c(f(\boldsymbol{x}_i), y_i)\|^2 \leq C_1^2 + C_2^2\|\nabla\text{NLL}(f(\boldsymbol{x}_i), y_i)\|^2\ ,\ \forall \boldsymbol{x}_i, y_i\ . \tag{A.2}$$

where C_1 and C_2 are constants s.t. $C_1 \geq 0$ and $C_2 \geq 1$, there exists μ-convex function for which FedAvg with more than two clients has an error

$$f^R(\boldsymbol{x}) - f^*(\boldsymbol{x}) \geq \Omega\left(\frac{C_1^2}{\mu R^2}\right). \tag{A.3}$$

To leverage Theorem A.1 to bound the deviation of $f^n(x) - f^*(x)$, denote the $\nabla\text{NLL}(f(\boldsymbol{x}, y))$ as $h(\boldsymbol{x}, y)$, we are required to show that

$$\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(\boldsymbol{x},y)}[h(\boldsymbol{x}, y)] \leq C_2^2 \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x},y)}[h(\boldsymbol{x}, y)] + C_1^2. \tag{A.4}$$

According to the Definition 4.2 of label skew, we can rewrite equation A.4 as

$$\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)} \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[h(\boldsymbol{x}, y)] \leq C_2^2 \mathop{\mathbb{E}}_{\mathbb{P}(y)} \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[h(\boldsymbol{x}, y)] + C_1^2. \tag{A.5}$$

Since we assume $\mathbb{P}(\boldsymbol{x}|y)$ is the same across clients, denote $\mathbb{E}_{\mathbb{P}(x|y)}[h(\boldsymbol{x}, y)]$ as $g(\boldsymbol{x})$, our objective is to show that

$$\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] \leq C_2^2 \mathop{\mathbb{E}}_{\mathbb{P}(y)}[h(x, y)] + C_1^2. \tag{A.6}$$

According to Assumption 4.2, we have

$$G \coloneqq \sup_i \mathrm{D}_{KL}(\mathbb{P}_i(y), \mathbb{P}(y)). \tag{A.7}$$

Suppose $g(\boldsymbol{x}, y)$ is bounded by M, which we will verify later, by total variation distance and Pinsker's Inequality, we have

$$|\mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] - \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]| \leq M \cdot D_{TV}(\mathbb{P}_i(y), \mathbb{P}(y)) \leq M\sqrt{\frac{1}{2}G}, \tag{A.8}$$

where D_{TV} is the total variation distance. Since, $w_i \geq 0$ and $\sum_{i=1}^{K} w_i = 1$, due to the linearity of expectation and triangular inequality, we have

$$|\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] - \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]| \leq \sum_{i=1}^{C} w_i |\mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] - \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]|. \tag{A.9}$$

and due to $\sum_{i=1}^{C} w_i$, the RHS of equation A.9

$$RHS \leq \sum_{i=1}^{C} w_i \cdot M\sqrt{\frac{1}{2}G} = M\sqrt{\frac{1}{2}G}. \tag{A.10}$$

By adding $\mathbb{E}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]|$ on LHS of equation A.9, we have

$$LHS + |\mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]| \leq M\sqrt{\frac{1}{2}G} + |\mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]|, \tag{A.11}$$

and again, by triangular inequality, we have

$$LHS + |\mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]| \geq |LHS + \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)]| = |\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)]|. \tag{A.12}$$

And since $g(\boldsymbol{x}, y)$ is the norm of gradient which is positive, we have

$$\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] \leq \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)] + M\sqrt{\frac{1}{2}G}. \tag{A.13}$$

which meets our object with $C_1^2 = M\sqrt{\frac{1}{2}G}$ and $C_2^2 = 1$. Now, let's verify that $g(\boldsymbol{x}, y)$ is indeed bounded by a constant M. Here, we make one additional assumption that

Assumption A.2. the extracted feature $z \coloneqq \boldsymbol{\theta}(\boldsymbol{x})$ of samples in the same class are similar. To be more specific, $\mathrm{Var}_{\mathbb{P}(\boldsymbol{x}|y)}(z) \leq S$ where S. This assumption is empirically verified in (Wang et al., 2021) and also used in (Zhang et al., 2022a).

This assumption implies that gradient difference among clients only happens at the last layer. So that we can explicitly write $g(\boldsymbol{x}, y)$ as

$$g(\boldsymbol{x}, y) = \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[|\nabla \mathrm{NLL}(\boldsymbol{x}, y)|^2] = \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[|(\sigma(w^\top z + b) - y) * z|^2]. \tag{A.14}$$

Note that the existence of variance in Assumption A.2 implies the existence of the mean $\overline{z}$. Now, we have

$$\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[|(\sigma(w^\top z + b) - y) * z|^2] \leq \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x}|y)}[|z|^2] = \mathrm{var}^2(z) + \overline{z}^2 \leq S^2 + \overline{z}^2. \tag{A.15}$$

Now replacing M inA.13 with $(\overline{z}^2 + S^2)$, we have

$$\sum_{i=1}^{C} w_i \mathop{\mathbb{E}}_{\mathbb{P}_i(y)}[g(\boldsymbol{x}, y)] \leq \mathop{\mathbb{E}}_{\mathbb{P}(y)}[g(\boldsymbol{x}, y)] + (\overline{z}^2 + S^2)\sqrt{\frac{1}{2}G}. \tag{A.16}$$

By Theorem A.1, we have

$$f^R(\boldsymbol{x}) - f^*(\boldsymbol{x}) \geq \Omega\left(\frac{(\overline{z}^2 + S^2)\sqrt{\frac{1}{2}G}}{\mu \boldsymbol{R}^2}\right). \tag{A.17}$$

Lastly, we investigate how this deviation propagates through the calibration error

$$\begin{aligned}
\overline{\mathrm{CE}}(f^n) &= \left(\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x},y)}[|\mathbb{P}(y = 1|\boldsymbol{x}) - f^R(\boldsymbol{x})|^2]\right)^{\frac{1}{2}} \\
&= \left(\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{x},y)}[|f^*(\boldsymbol{x}) - f^R(\boldsymbol{x})|^2]\right)^{\frac{1}{2}} \\
&\geq \mathop{\mathbb{E}}_{\mathbb{P}(x,y)}[|f^*(\boldsymbol{x}) - f^R(\boldsymbol{x})|] \qquad \text{Jensen's Inequality} \\
&\geq \Omega\left(\frac{(\overline{z}^2 + S^2)\sqrt{\frac{1}{2}G}}{\mu \boldsymbol{R}^2}\right).
\end{aligned} \tag{A.18}$$

□

B. Examples of Order Preserving Network

Theorem 1 in (Rahimi et al., 2020) states that a continuous function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is order-preserving, if and only if $f(\mathbf{x}) = S^{-1}(\mathbf{x})U\mathbf{w}(\mathbf{x})$ where U is an upper-triangular matrix of ones and $\mathbf{w} : \mathbb{R}^n \mapsto \mathbb{R}^n$ s.t. $\begin{cases} \mathbf{w}_i(\mathbf{x}) = 0, \mathbf{y}_i = \mathbf{y}_{i+1} \text{ and } i < n \\ \mathbf{w}_i(\mathbf{x}) > 0, \mathbf{y}_i > \mathbf{y}_{i+1} \text{ and } i < n \\ \mathbf{w}_n(\mathbf{x}) \text{ is arbitrary} \end{cases}$ where $\mathbf{y} = S(\mathbf{x})\mathbf{x}$ is the sorted version of $\mathbf{x}$.

The order-preserving network is a direct application of this theorem, where $S(\mathbf{x})$ is achieved using the sorting component and the element-wise production between activation after softplus and the $\mathbf{y}_i - \mathbf{y}_{i+1}$ ensures $\mathbf{w}_i(\mathbf{x}) \geq 0$ when $\mathbf{y}_i - \mathbf{y}_{i+1} \geq 0$. To verify this theorem, consider an unsorted input vector $\mathbf{x} = [3, 4, 2, 2]^\top$, where the correct order should be $\mathbf{x}_2 > \mathbf{x}_1 > \mathbf{x}_3 = \mathbf{x}_4$. In other words, $\mathbf{y} = S(\mathbf{x})\mathbf{x} = [\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4]^\top$ if we prefer $\mathbf{x}_3$ or $[\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_3]^\top$ otherwise. Then, $\mathbf{w} = [\mathbf{w}_1 > 0, \mathbf{w}_2 > 0, \mathbf{w}_3 = 0, \mathbf{w}_4]^\top$. $U\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 + \mathbf{w}_2 + \mathbf{w}_3 + \mathbf{w}_4 \\ \mathbf{w}_2 + \mathbf{w}_3 + \mathbf{w}_4 \\ \mathbf{w}_3 + \mathbf{w}_4 \\ \mathbf{w}_4 \end{bmatrix}$ is monotonically non-decreasing. Once we permute $U\mathbf{w}$ back by exchanging the index 1 and 2, $f(\mathbf{x}) = S^{-1}(x)U\mathbf{w}$ also has the ordering $f(\mathbf{x})_2 > f(\mathbf{x})_1 > f(\mathbf{x})_3 = f(\mathbf{x})_4$.

C. Discussion on Baselines

As part of our experimental design, we benchmark our framework against intuitive scaler designs and aggregation strategies, and highlight their correspondence in existing works.

The simplest aggregation strategy involves disregarding the parameterization of scalers and adopting a direct averaging of the model's scaled probability, which is equivalent to Deep Ensemble (Lakshminarayanan et al., 2017; Rahaman et al., 2021) and We denote the methods as ENS..Within our framework, the ensemble method is expressed as

$$\overline{\phi} \circ f^R(\boldsymbol{x}_i) = \frac{1}{C}\sum_{c=1}^{C} \phi_c \circ f^R(\boldsymbol{x}_i). \tag{C.1}$$

Despite the simplicity of ENS, it has been demonstrated to yield robust uncertainty quantification, often outperforming more elaborate methods (Lakshminarayanan et al., 2017; Lee et al., 2015).

An intuitive extension to ENS is to average the temperature parameter of individual temperature scalers to

$$\overline{\phi} \circ f^R(\boldsymbol{x}_i) = \sigma\left(\frac{\boldsymbol{\theta}^R(\boldsymbol{x}_i)}{\sum_{c=1}^{C} T_c}\right) \tag{C.2}$$

referred to asAVGT. AVGT and ENS share similar advantages and limitations. Both methods can be applied post-training, negating the need for communication overhead However, Rahaman et al. (2021) reveal that ENS only works when $\phi_c \circ f^R(\boldsymbol{x}_i)$ is overconfident and it does not meaningfully contribute to an ensemble's uncertainty quantification when $\mathbb{P}_c(x, y)$ is different from $\mathbb{P}(x, y)$ (Abe et al., 2022). The simplicity of these methods limits their generalization capabilities.

MD-TS (Yu et al., 2022) proposes to use a linear regression model to regression estimate the temperature T from the last layer's output $\boldsymbol{\theta}_i(x)$. This approach improves the model's calibration under distribution shifts. MD-TS aligns with our analysis as linear regressor offers better generalization capability and scaling through temperature does not affect accuracy. We adapt this concept for federated settings. Unlike the original approach, which requires data sharing and a validation dataset $\mathcal{D}_{val}$, we suggest leveraging FedAvg for deriving the linear model, bypassing the need for direct access to $\mathcal{D}_{val}$, we denote this method as LR-TS.

LR-TS can be formulated as: A client maintains a linear regression mode lr_c with the aim of predicting the T_c.

$lr_c(\boldsymbol{\theta}\boldsymbol{x}_i; W, b)$ is trained using

$$lr_c^* = \arg_{lr}\min \sum_{i=1}^{|\mathcal{D}_{val}|} (W\boldsymbol{\theta}^R\boldsymbol{x}_i + b) - T_c \tag{C.3}$$

Then the global calibration with scaler $\overline{\phi}$ can be formulated as

$$\overline{\phi} \circ f^R(\boldsymbol{x}_i) = \sigma\left(\frac{\boldsymbol{\theta}^R(\boldsymbol{x}_i)}{lr(\boldsymbol{\theta}^R(\boldsymbol{x}_i))}\right) \tag{C.4}$$

, where $lr \coloneqq \text{FedAvg}(\{lr_c\})$.

D. The Weight Matching Algorithm

For ease of reference, we list the Weight Matching algorithm from (Ainsworth et al., 2022) here.

Algorithm 2 WeightMatching (Ainsworth et al., 2022)

procedure WeightMatching($\overline{\boldsymbol{\pi}}^0, \boldsymbol{\pi}_c$)
Given:
Global scaler weights $\overline{\boldsymbol{\pi}}^0 = \{W_1^{(A)}, \ldots, W_L^{(A)}\}$ and local scaler weight $\boldsymbol{\pi}_c = \{W_1^{(B)}, \ldots, W_L^{(B)}\}$
Result: A permutation $\mathbf{M} = \{P_1, \ldots, P_{L-1}\}$ of $\boldsymbol{\pi}_c$ such that $\text{vec}(\overline{\boldsymbol{\pi}}^0) \cdot \text{vec}(\pi(\boldsymbol{\pi}_c))$ is approximately maximized.
Initialize: $P_1 \leftarrow I, \ldots, P_{L-1} \leftarrow I$
repeat
  for ℓ in RANDOMPERMUTATION$(1, \ldots, L-1)$ **do**
    $P_\ell \leftarrow \text{SOLVELAP}(W_\ell^{(A)} P_{\ell-1} (W_\ell^{(B)})^T + (W_{\ell+1}^{(A)})^T P_{\ell+1} W_{\ell+1}^{(B)})$
  end for
until convergence
return $\mathbf{M}$