

Information-Geometric Barycenters for Bayesian Federated Learning

Nour Jamoussi[†], Giuseppe Serra[†], Photios A. Stavrou[†], Marios Kountouris^{†‡}

[†]Communication Systems Department, EURECOM, France

[‡]Andalusian Institute of Data Science and Computational Intelligence (DaSCI)

Department of Computer Science and Artificial Intelligence, University of Granada, Spain

Abstract—Federated learning (FL) is a widely used and impactful distributed optimization framework that achieves consensus by averaging locally trained models. While effective, this approach may not align well with Bayesian inference, where the model space is more naturally represented as a distribution space. Taking an information-geometric perspective, we reinterpret FL aggregation as the problem of finding the barycenter of local posteriors using a predefined divergence metric, minimizing the average discrepancy across clients. This perspective provides a unifying framework that generalizes many existing methods and offers crisp insights into their theoretical underpinnings. We then propose BA-BFL, an algorithm that retains the convergence properties of Federated Averaging in non-convex settings. In non-independent and identically distributed scenarios, we conduct extensive comparisons with statistical aggregation techniques, showing that BA-BFL achieves performance comparable to state-of-the-art methods while also providing a geometric interpretation of the aggregation phase. Additionally, we extend our analysis to Hybrid Bayesian Deep Learning, exploring the impact of Bayesian layers on uncertainty quantification and model calibration.

Index Terms—Bayesian Federated Learning, Hybrid Bayesian Deep Learning, Uncertainty Quantification, Model Aggregation.

I. INTRODUCTION

Federated Learning (FL) has emerged as the de facto standard for decentralized learning, particularly in scenarios that demand strong privacy guarantees. As originally introduced in [1], an FL system consists of a central server that maintains a global model and interacts with multiple clients (end-user devices), each of which holds private local data. FL schemes typically operate in two phases. In the local learning phase, each client trains a model on its own private data. In the aggregation phase, the locally updated models are transmitted to the server and merged according to a predetermined rule. These phases repeat iteratively, with the global model from the previous iteration distributed to clients as the starting point for their local models in the current iteration.

Aggregation plays a central role in FL, allowing individual client contributions to be combined into a global model while preserving data privacy and ensuring communication efficiency. Although various aggregation strategies have been proposed (e.g., [10]), most aggregation methods rely on variants of weighted averaging. Notable examples include FedAvg [1] and FedProx [2], which aggregate local models by computing

a weighted average of their parameters. The choice of weights is an important design parameter, as it can encode auxiliary attributes such as the relative importance of each client’s model to the overall objective or reflect the amount of data available to each client.

A key challenge in FL, and generally in distributed learning systems, is the statistical heterogeneity among participating clients. In real-world scenarios, client datasets rarely satisfy the idealized assumption of independent and identically distributed (i.i.d.) data. Instead, they often exhibit significant heterogeneity and distributional shifts across clients. As reviewed in [2], [11], five major forms of heterogeneity are typically identified: *label distribution skew*: differences in the frequency of specific labels across clients (e.g., under- or over-represented classes); *feature distribution skew*: differences in the feature distributions associated with the same label; *concept drift*: cases where the same label is associated with different feature distributions across clients; *concept shift*: cases where identical samples receive different labels from different clients, and *quantity skew*: differences in the number of data samples held by each client. While these heterogeneities are of high practical relevance, addressing all of them simultaneously remains a significant challenge. As a result, most of the existing FL research is focused on only a subset of these challenges [12]–[14]. Given the growing applicability of FL in real-world settings, uncertainty quantification and model calibration are central to building trustworthy and reliable models. Nonetheless, these aspects remain largely underexplored in existing research on deterministic FL. A preliminary study on the topic, with a specific application to healthcare, is presented in [15]. It provides an overview of various uncertainty quantification methods for deterministic FL, which were later implemented in [16]. However, it is important to note that the techniques discussed are primarily inspired by Bayesian approaches, such as Bayesian ensembles and Monte Carlo dropout.

Bayesian learning excels at improving model reliability, as Bayesian methods enable more accurate uncertainty quantification and calibration, making them a compelling solution for FL in non-i.i.d. contexts. FedPPD [17] introduces an FL framework with built-in uncertainty quantification: in each round, each client estimates both the posterior distribution over its parameters and the posterior predictive distribution

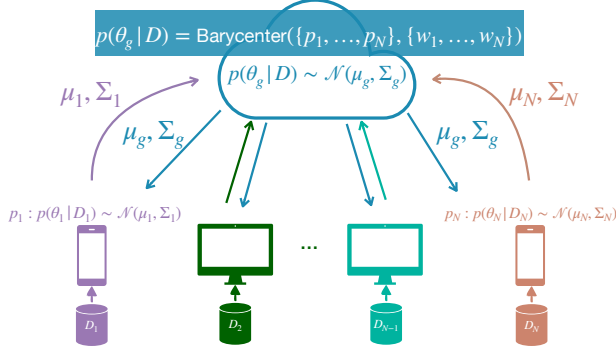


Fig. 1: The BA-BFL framework.

(PPD). The PPD is subsequently distilled into a single deep neural network, which is then sent to the server. pFedBayes [18] and Fedpop [19] are also Bayesian approaches with a focus on uncertainty quantification aspects, proposed in the context of personalized Bayesian Federated Learning (BFL). Nonetheless, we emphasize that many of the existing methods [17], [18], [20], [21] rely on variations of weighted averaging of the posteriors' parameters.

a) *Contributions*: This work introduces a unifying perspective on aggregation methods in BFL through the lens of barycentric aggregation (BA-BFL). Given a divergence metric, we interpret the aggregation process as a geometric problem, where the global model is identified as the barycenter of the local posteriors. Unlike existing methods that often rely on heuristic variations of parameter averaging, our approach is theoretically grounded: it explicitly minimizes the average divergence between the global posterior and the local posteriors. We show that this methodology generalizes several aggregation strategies previously proposed in the literature. Furthermore, the proposed methods preserve the convergence properties of FedAvg, even in non-convex settings (see Theorem III.1). We evaluate our approach against state-of-the-art Bayesian aggregation methods, comparing accuracy and uncertainty quantification in heterogeneous settings. The results demonstrate that our method achieves performance comparable to existing statistical aggregation techniques. To bridge gaps in the BFL literature and buildinsights from Hybrid Bayesian Deep Learning (HBDL) [22], we further examine how limiting the number of Bayesian layers affects the performance of different Bayesian aggregation methods.

b) *Notation*: Table I summarizes the key symbols and their meanings for clarity and convenience of reference.

II. BACKGROUND AND RELATED WORK

a) *Federated Learning*: An FL system [1] consists of a central server and N clients that engage in an iterative learning process through server-client communication. For each communication round, the k^{th} client trains its local model, parameterized by θ_k , on its private data $\mathcal{D}_k$. Subsequently, the model parameters θ_k are sent to the server, which aggregates them to obtain the global model. The updated global model is

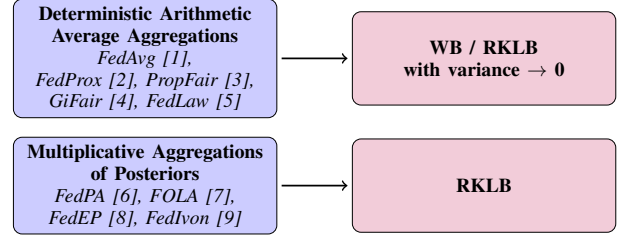


Fig. 2: Mapping of various aggregation methods to their corresponding barycenter formulations.

Symbol	Meaning
$\mathcal{D}_k$	Local dataset of client k
$\mathcal{D}$	Union of all datasets $\bigcup_{k=1}^N \mathcal{D}_k$
D_α	α -divergence
D_{KL}	Kullback–Leibler divergence
W_2^2	Squared 2-Wasserstein distance
w_k	Weight of client k
f_k	Local objective of client k
f	Weighted sum $\sum_{k=1}^N w_k f_k$
ψ_k	Posterior parameters of client k 's model
ψ_g	Posterior parameters of the global model

TABLE I: Summary of notation.

then distributed back to the clients to refine their local models in the next communication round. Through this process, FL aims to learn a global model θ^* on the aggregated dataset $\mathcal{D} = \bigcup_{k=1}^N \mathcal{D}_k$ from all participating clients.

In general, the objective function in FL takes the form

$$\min_{\theta} f(\theta) = \sum_{k=1}^N w_k f_k(\theta) \quad (1)$$

where $f_k(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_k} [\mathcal{L}(\theta; (x, y))]$ is the local objective function of the k^{th} client, and w_k is its associated weight, with $\sum_{k=1}^N w_k = 1$. At each communication round, minimizing $f_k(\theta)$ locally produces the client update θ_k .

b) *Hybrid Bayesian Deep Learning (HBDL)*: Despite its remarkable performance, deep learning does not address crucial challenges in realistic scenarios, such as reliability and uncertainty quantification. In a recent position paper [24], the authors propose Bayesian Deep Learning as a solution to the ethical, privacy, and safety challenges of modern deep learning. Acknowledging ongoing challenges with Bayesian DL, such as the additional computational cost of applying Bayesian methods to large-scale deep models, the authors envision the alternative framework of HBDL to preserve the efficiency and lower complexity of deep learning while retaining the reliability of Bayesian DL. HBDL is also discussed in [25] as *Bayesian inference applied only to the last (or last few) layers*. The core idea is to replace some of the layers in a Bayesian deep model with deterministic ones, thereby making the model closer to its classical deep learning counterpart. This partial

BFL Method	Local Bayesian Technique	Global Aggregation Method
FedPA [6]	MCMC sampling	Multiplicative aggregation of posteriors
pFedBayes [18]	Variational Inference	Weighted averaging of posteriors' statistics
FVBA [20]	Variational Inference	Weighted averaging of posteriors' statistics
FedEP [8]	Variational Inference	Multiplicative aggregation of posteriors
FedHB [23]	Variational Inference	Bayesian posterior inference
FOLA [7]	Laplace Approximation	Multiplicative aggregation of posteriors
Not Named [21]	Variational Inference & MC dropout	Weighted averaging of posteriors' statistics
FedPPD [17]	MCMC sampling	Weighted averaging of posteriors' statistics
FedIvon [9]	IVON	Multiplicative aggregation of posteriors

TABLE II: Categorization of parametric client-side Bayesian Federated Learning methods.

Bayesian formulation makes it possible to retain uncertainty quantification capabilities while reducing complexity relative to a fully Bayesian model.

c) Bayesian Federated Learning: BFL aims to incorporate the strengths of Bayesian deep learning into the FL framework and to provide a potential solution to the challenges outlined above. In this work, we focus primarily on parametric, client-side BFL methods [26]. We categorize these methods based on their strategies for global model construction. Table II provides a summary of this discussion.

- *Multiplicative Aggregation of Posteriors:* FedPA [6] employs Stochastic Gradient Markov Chain Monte Carlo (MCMC) for local posterior inference, aggregating client posteriors through a product of Gaussian distributions. FOLA [7] approximates local posteriors using Laplace approximation. A multivariate Gaussian product mechanism is used for global posterior construction, while prior distributions derived from the global posterior guide local training, thereby enabling a continual learning setting. FedEP [8] frames FL as a distributed variational inference problem, aligning the global posterior with local posteriors through multiplicative aggregation. FedIvon [9] adopts Improved Variational Online Newton (IVON) [27] to approximate local posteriors as Gaussians with diagonal covariance, efficiently updating both mean and variance with second-order information. The server aggregates these local posteriors via the weighted product of the local Gaussians. It is important to note that approximating the global posterior as the weighted product of local posteriors is equivalent to computing their geometric mean, which also corresponds to the RKL aggregation method introduced in this work, offering a geometric interpretation of this aggregation strategy.
- *Weighted Averaging of Posteriors' Statistics:* pFedBayes [18] employs variational inference to incorporate uncertainty into model parameters. From a continual learning perspective, it minimizes the KL divergence between global and local posterior distributions, balancing local reconstruction error with global alignment, which the paper presents mainly as a personalization aspect. FVBA [20] investigates aggregating variational Bayesian neural networks using five statistical aggregation schemes. In [21], the authors integrate Bayesian deep learning with FL, employing variational inference and Monte Carlo

Dropout for inference in local models. As in [20], different statistical aggregations are evaluated, highlighting the importance of the aggregation method chosen for model performance. FedPPD [17] leverages MCMC sampling for local posterior inference and distills posterior predictive distributions into individual deep neural networks via Stochastic Gradient Langevin Dynamics. It adopts either simple averaging or a distillation-based global aggregation approach.

- *Bayesian Posterior Inference:* FedHB [23] introduces a hierarchical Bayesian framework in which local model parameters are governed by a global latent variable. Variational inference is used to optimize the local and global posteriors through block-coordinate optimization.

III. PROPOSED METHOD

In this section, we introduce our problem formulation and present the main theoretical results. We start by formalizing the key technical aspects of the client-side BFL framework.

a) Client-Side BFL: The Bayesian view presents a different framework for FL. The goal is to estimate the posterior distribution of the global model's parameters, $p(\theta^*|\mathcal{D})$, given the posterior distributions of local models $p(\theta_k|\mathcal{D}_k)$.

Nevertheless, exact posterior inference is usually intractable, requiring the use of approximate inference methods instead. In this work, we consider variational inference [28], [29] to approximate the local posteriors given a common prior distribution $p(\theta)$ and the client likelihoods $p(\mathcal{D}_k|\theta_k)$.

For a parametric family $\mathcal{M}_\Psi$ parametrized by $\psi \in \Psi$, the optimization problem seeks to identify the distribution $q_\psi \in \mathcal{M}_\Psi$ that minimizes the KL divergence from the posterior distribution $p(\theta|\mathcal{D})$, i.e.,

$$\min_{\psi \in \Psi} D_{\text{KL}}(q_\psi(\theta) \| p(\theta|\mathcal{D})) \quad (2)$$

However, the minimization in (2) is not directly tractable and is commonly approached through the derivation of the *Negative Evidence Lower Bound* surrogate objective $\min_{\psi \in \Psi} \mathcal{L}(\psi, D)$, where

$$\mathcal{L}(\psi, D) = -\mathbb{E}_{q_\psi(\theta)}[\log p(\mathcal{D}|\theta)] + D_{\text{KL}}(q_\psi(\theta) \| p(\theta)). \quad (3)$$

The local models are trained by minimizing (3) to achieve their models' posterior distributions $p(\theta_k|\mathcal{D}_k)$, $\forall k \in \{1, \dots, N\}$. The local posteriors are then aggregated in order to get the global model's posterior $p(\theta^*|\mathcal{D})$. Given this setting,

Algorithm 1 BA-BFL: Barycentric Aggregation for Bayesian Federated Learning

Server's Input: number of communication rounds R , aggregation weights $\{w_i\}_{i=1}^N$, global distribution's initial parameters ψ_g^0 .

Client's k input: number of local training epochs T , local training set $\mathcal{D}_k$

```

1: for each round  $r = 1, \dots, R$  do
2:   Sample clients' subset  $\mathcal{S}_r \subset \{1, \dots, N\}$ 
3:   Communicate  $\psi_g^{r-1}$  to all clients  $k \in \mathcal{S}_r$ 
4:   for client  $k \in \mathcal{S}_r$  do
      /*  $T$  epochs of Gradient Descent
      (GD) starting at  $\psi_g^{r-1}$  */
5:    $\psi_k^r \leftarrow \mathbf{GD}(\mathcal{L}(\cdot, \mathcal{D}_k), \psi_g^{r-1})$ . // Eq. 3
6:   end for
      /* Aggregation and global update */
7:    $\psi_g^r \leftarrow \mathbf{D-Barycenter}(\{\psi_k^r\}_{k=1}^N, \{w_k\}_{k=1}^N)$ 
8: end for

```

we now introduce our main assumptions regarding the common prior $p(\theta)$ and the parametric family $\mathcal{M}$, which will stay valid throughout the rest of this paper.

Assumption 1. *For each client, we assume the prior distribution $p(\theta)$ to be a d -dimensional Gaussian with independent marginals, parameterized by a zero mean vector $\mathbf{0}_d$ and an identity covariance matrix $\mathbf{I}_d$.*

Assumption 2. (Mean-field Model) *The parametric family $\mathcal{M}_\Psi$ is composed of d -dimensional Gaussian distributions with independent marginals, i.e., $q_\psi = \mathcal{N}(\mu, \Sigma)$, with mean $\mu \in \mathbb{R}^d$ and diagonal covariance $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$.*

b) Bayesian Aggregation as Posteriors Barycenter: The main novelty of this work stands in the introduction of the general Barycentric Aggregation framework for BFL (BA-BFL), an aggregation method inspired by the geometric properties of the manifold to which the local posteriors $\{p(\theta_k|\mathcal{D}_k)\}_{k=1\dots N}$ belong. Given a divergence metric D , we propose as a global model the barycenter p_D^* of the set of clients' posteriors, i.e., the distribution that minimizes the weighted divergence from a given set. The following problem formalizes this interpretation of the aggregation process.

Problem 1. (D -barycenter) *Given a statistical manifold $\mathcal{M}$, a divergence function $D : \mathcal{M} \times \mathcal{M} \rightarrow [0, \infty)$, and a set of distributions $\mathcal{S} = \{p_k\}_{k=1\dots N} \subseteq \mathcal{M}$ with associated normalized weights $\{w_k\}_{k=1\dots N}$, i.e., $\sum_{k=1}^N w_k = 1$, the barycenter p_D^* of the set $\mathcal{S}$ is defined as:*

$$p_D^* = \arg \min_{q \in \mathcal{M}} \sum_{k=1}^N w_k D(p_k || q). \quad (4)$$

We now study Problem 1 under various assumptions on the distribution set $\mathcal{S}$ and the divergence metric D . First, we consider the general case where $D = D_\alpha$, namely,

the divergence belongs to the family of α -divergences for $\alpha \in \mathbb{R} \setminus \{0\}$, without any additional assumptions on $\mathcal{S}$. As shown in [30], [31], the corresponding barycenter $p_{D_\alpha}^*$ takes the following form:

$$p_{D_\alpha}^* = \frac{\left(\sum_{k=1}^N w_k p_k^\alpha\right)^{\frac{1}{\alpha}}}{\int \left(\sum_{k=1}^N w_k p_k^\alpha\right)^{\frac{1}{\alpha}} d\nu}. \quad (5)$$

Moreover, by taking the limit $\alpha \rightarrow 0$, i.e., corresponding to the *reverse Kullback-Leibler (RKL) divergence* $D_{RKL}(p||q)$, the barycenter p_{RKL}^* takes the form:

$$p_{RKL}^* = \frac{\prod_{k=1}^N p_k^{w_k}}{\int \prod_{k=1}^N p_k^{w_k} d\nu}. \quad (6)$$

We now focus on the case where all $p_k \in \mathcal{S}$ are d -dimensional Gaussian distributions, i.e., $p_k = \mathcal{N}(\mu_k, \Sigma_k)$, with mean μ_k and covariance matrix Σ_k . This setting derives from Assumption 2, where we assume that the parameters of each Bayesian layer are Gaussian distributed. For the same reasons, we are also interested in the cases where the resulting barycenter is itself Gaussian, to enforce that global and local models belong to the same family of distributions. Alas, this is not the case for the majority of α -divergences, as discussed in the following remark.

Remark 1. (On the α -barycenter of a set of Gaussians) *Given $\mathcal{S} = \{\mathcal{N}(\mu_k, \Sigma_k)\}_{k=1\dots N}$, the barycenter distribution $p_{D_\alpha}^*$ in (5) is not Gaussian. In fact, $(p_{D_\alpha}^*)^\alpha \propto \sum_{k=1}^N w_k p_k^\alpha$, showing that the resulting barycenter is related to the Gaussian mixture obtained from the weighted sum of the elements of $\mathcal{S}$. On the other hand, p_{RKL}^* is still Gaussian since the Gaussian family is closed under the product operation and (6) is the normalized product of unnormalized Gaussians.*

In light of the above technical remark, among the considered α -divergences we focus exclusively on the case of $\alpha \rightarrow 0$, i.e., p_{RKL}^* , as the barycenter naturally belongs to the Gaussian family, leaving the study of other α -divergences as future work. Given $\mathcal{S} = \{\mathcal{N}(\mu_k, \Sigma_k)\}_{k=1\dots N}$, the RKL barycenter is $p_{RKL}^* = \mathcal{N}(\mu_{RKL}, \Sigma_{RKL})$, where

$$\Sigma_{RKL} = \left(\sum_{k=1}^N w_k \Sigma_k^{-1}\right)^{-1}, \quad \mu_{RKL} = \Sigma_{RKL} \sum_{k=1}^N w_k \Sigma_k^{-1} \mu_k. \quad (7)$$

This result is well-established in the literature and has been derived using various approaches (e.g., see [32]).

Similarly to the D_{RKL} divergence, the barycenter of a set of Gaussians in the Wasserstein-2 distance belongs to the same family. In the general setting, the parameters of the barycenter are obtained using a set of fixed-point equations [33]. However, when the set of covariance matrices $\{\Sigma_k\}_{k=1\dots N}$ consists of diagonal matrices, i.e., $\Sigma_k = \text{diag}(\sigma_{k,1}^2, \dots, \sigma_{k,d}^2)$, analytic

expressions for the barycenter statistics can be derived, as shown in [33]:

$$\Sigma_{W_2^2} = \left(\sum_{k=1}^N w_k \Sigma_k^{\frac{1}{2}} \right)^2, \quad \mu_{W_2^2} = \sum_{k=1}^N w_k \mu_k. \quad (8)$$

In the sequel, we refer to the aggregation methods resulting from (7) and (8) with the acronyms RKL and WB, respectively. We discuss the applicability of the proposed methods in HBDL, focusing on cases where part of the architecture is deterministic. In such setting, the posterior distribution $p(\theta_{k,i}|\mathcal{D})$ for the i^{th} layer of the k^{th} client is constrained to be a point-mass located at $\mu_{k,i}$, i.e., $p(\theta_{k,i}|\mathcal{D}) = \delta_{(\theta_{k,i}=\mu_{k,i})}$ where δ_x is the Dirac distribution. We investigate the behavior of the proposed methods considering the posterior $p(\theta_{k,i}|\mathcal{D}) = \mathcal{N}(\mu_{k,i}, \epsilon)$ in the limit case of $\epsilon \rightarrow 0$. Both (7) and (8) can be shown to be well-defined in the limit, resulting in the barycenter distribution $p^*(\theta_i|\mathcal{D}) = \delta_{(\theta_i=\sum_{k=1}^N w_k \mu_{k,i})}$. Notably, this coincides with the arithmetic mean aggregation commonly used in deterministic FL, creating a seamless connection between deterministic and probabilistic aggregation approaches within our framework.

Compared to other state-of-the-art methodologies in parametric client-side BFL - the primary focus of this work - our barycentric perspective extends the widely used weighted multiplication of posteriors, a predominant aggregation method in the literature. We demonstrate that this approach coincides with the RKL, thereby reinforcing its theoretical foundation. The other baseline used in our comparative studies in Section IV, is the arithmetic mean of the local posteriors' statistics. In addition, we consider other statistical aggregation methods that, while explored in some comparative studies, have yet to be adopted in practical applications. More broadly, on the server side, BFL can leverage Bayesian ensembles [34], also referred to as Bayesian Model Averaging (BMA), which combines predictions from sampled models to produce a more robust global estimate. Notably, BMA can also be interpreted through the lens of KL barycenters, as highlighted in [Proposition 1.5, [35]]. This barycentric perspective not only provides a strong theoretical grounding for existing methods, as illustrated in Figure 2, but also opens the door to exploring alternative divergence measures to enhance the aggregation process, including the WB.

To conclude this section, we provide theoretical guarantees for the convergence of BA-BFL under both WB or RKL aggregation, as stated in the following theorem.

Theorem III.1. (Convergence) *Under Assumption 2, and using either RKL or WB aggregation, BA-BFL inherits and preserves the convergence properties of FedAvg, as shown in [36], for non-convex scenarios with both i.i.d. and non-i.i.d. data.*

Proof. The proof of convergence of BA-BFL is based on existing results on the convergence proof of FedAvg in the non-i.i.d. setup. This connection is possible by recognizing that BA-BFL can be seen as an instance of FedAvg on the parameter space

of the chosen parametric family of distributions, subject to a bijective transformation.

Considering a parametric family $\mathcal{M}_\Psi$ parametrized by $\psi \in \Psi$, let $F : \Psi \rightarrow \hat{\Psi}$ be an invertible mapping. Then, at the i^{th} user, there is no difference between optimizing the local objective on Ψ or a modified version optimized on $\hat{\Psi}$ via the inverse of F , i.e.,

$$\min_{\psi \in \Psi} f_i(\psi) = \min_{\hat{\psi} \in \hat{\Psi}} f_i(F^{-1}(\hat{\psi})) \quad (9)$$

where $f_i(\psi) = -\mathbb{E}_{q_\psi(\theta)}[\log p(\mathcal{D}|\theta)] + D_{KL}(q_\psi(\theta)||p(\theta))$.

We are interested in the case where there exists F such that the barycentric aggregation on Ψ induced by a divergence D is equivalent to an arithmetic mean on $\hat{\Psi}$, i.e.,

$$\psi_g = BA(\{\psi\}_{i=1}^N) = F^{-1} \left(\sum_{i=1}^N w_i F(\psi_i) \right). \quad (10)$$

Under the condition that such mapping exists, then the optimization dynamics of BA-BFL on Ψ is equivalent to FedAvg on $\hat{\Psi}$, i.e., *BA-BFL inherits the same convergence properties of FedAvg*¹.

Considering the family of d -dimensional mean-field Gaussian distributions, i.e., with diagonal covariance matrix $\Sigma = \text{diag}(\{\sigma_k^2\}_{k=1}^d)$, it can be parametrized by $\psi = [(\mu_1, \sigma_1^2), \dots, (\mu_d, \sigma_d^2)]$ with $\mu_k \in \mathbb{R}$ and $\sigma_k^2 \in \mathbb{R}^+$. Then, we can define the set of entry-wise invertible functions

$$F_{RKL}(\mu_k, \sigma_k^2) = \left(\frac{\mu_k}{\sigma_k^2}, \frac{1}{\sigma_k^2} \right), \quad F_{RKL}^{-1}(\nu_k, \psi_k) = F_{RKL}(\nu_k, \psi_k)$$

$$F_{W_2^2}(\mu_k, \sigma_k^2) = \left(\mu_k, \sqrt{\sigma_k^2} \right), \quad F_{W_2^2}^{-1}(\nu_k, \psi_k) = (\nu_k, \psi_k^2)$$

which satisfy (10) respectively for RKL and W_2^2 divergences. Therefore, given the previous result, BA-BFL under either RKL or W_2^2 divergences enjoys the same convergence properties of FedAvg, thus concluding the proof. $\square$

IV. EXPERIMENTS

We devote this section to the experimental investigation of the proposed BA-BFL. To this end, we conduct experimental studies on the FashionMNIST [37], SVHN [38] and CIFAR-10 [39] datasets, within a heterogeneous client setting. The datasets used exhibit varying levels of difficulty.

To compare the proposed methodologies, we consider the following baselines:

- for deterministic FL, FedAvg [1] aggregates the parameters of the clients' models through arithmetic weighted average, i.e., $\theta^* = \sum_{k=1}^N w_k \theta_k$.
- for BFL, [20], [21] propose different possible statistical aggregation methods detailed below:
 - Empirical Arithmetic Aggregation (EAA),

$$\mu_{EAA} = \sum_{k=1}^N w_k \mu_k, \quad \sigma_{EAA}^2 = \sum_{k=1}^N w_k \sigma_k^2.$$

¹For a detailed convergence proof of FedAvg in the non-i.i.d. scenario, we refer the reader to [36]

- Gaussian Arithmetic Aggregation (GAA),

$$\mu_{GAA} = \sum_{k=1}^N w_k \mu_k, \quad \sigma_{GAA}^2 = \sum_{k=1}^N w_k^2 \sigma_k^2.$$

- Arithmetic Aggregation with Log Variance (AALV),

$$\mu_{AALV} = \sum_{k=1}^N w_k \mu_k, \quad \sigma_{AALV}^2 = e^{\sum_{k=1}^N w_k \log \sigma_k^2}.$$

a) *Metrics*: We evaluate the performance of the considered FL algorithms based on three criteria: accuracy, uncertainty quantification, where lower NLL indicates a better model fit, and model calibration, where lower ECE indicates better alignment between predicted probabilities and actual outcomes.

b) *Experimental Setup*: To induce label shifts among the 10 clients participating in the FL scheme, we partition the samples of each label across the clients using a Dirichlet distribution as suggested in [20], [40]–[44].

We assign the aggregation weights to reflect the importance of each client in proportion to the volume of data locally owned, i.e., $w_k = \frac{|\mathcal{D}_k|}{|\mathcal{D}|}$ where $|\cdot|$ indicates the number of samples in the dataset.

The architecture of the global and local models consists of two convolutional layers and three fully connected layers. Following an HBDL approach, we implement the last $n = 0, 1, 2, 3$ layers as Bayesian fully connected layers, whereas the remaining layers are deterministic. In our comparative study, increasing n allows measuring the impact of additional Bayesian layers on the uncertainty quantification, model calibration, and the cost-effectiveness of the FL algorithm in time.

c) *Overview of the Results*: Table 1 summarizes the accuracy performance of BA-BFL using RKLB and WB, alongside the considered baseline methods. All aggregation methods are evaluated under the same model architecture, with an equal number of Bayesian layers. The results show that the proposed aggregation methods achieve performance comparable to the baselines in most scenarios. Furthermore, all Bayesian methods consistently outperform FedAvg across all evaluated datasets. Notably, the best accuracy is obtained with a single Bayesian layer for FashionMNIST and SVHN, and with two Bayesian layers for CIFAR-10. Interestingly, increasing the number of Bayesian layers does not always result in improved accuracy.

To assess whether the performance differences between aggregation methods are statistically significant, we employ the Bayesian signed-rank test as described in [45]. For each pair of aggregation methods, we compare their performance across all shared evaluation points, i.e., for each combination of dataset, number of Bayesian layers, and random seed, ensuring a paired analysis under identical experimental conditions. The Bayesian test computes the posterior probabilities that one method outperforms, underperforms, or performs similarly to the other, where similarity is defined within a Region of Practical Equivalence (ROPE). Rather than relying on a

TABLE III: Accuracy of the global models resulting from FedAvg, BA-BFL with RKL and W_2^2 barycentric aggregation (RKLB, WB), and BFL baseline aggregation methods (AALV, EAA, GAA). The methods are grouped based on the number of Bayesian layers (Nbl) used in the model architecture. The evaluation is conducted across three datasets (FashionMNIST, CIFAR-10, and SVHN). Bold values represent the highest performance for each dataset, while underlined values denote the best result for the specific dataset within each group of rows corresponding to a specific Nbl.

Nbl	Algorithm	FashionMNIST	SVHN	CIFAR-10
0	FedAvg	87.88 ± 0.79	86.06 ± 0.45	61.63 ± 3.11
1	AALV	88.22 ± 0.34	86.52 ± 0.29	63.42 ± 3.02
	EAA	88.07 ± 0.22	86.24 ± 0.20	63.69 ± 2.47
	GAA	88.15 ± 0.31	86.36 ± 0.28	63.66 ± 2.22
	RKLB	88.07 ± 0.36	86.26 ± 0.26	63.37 ± 2.62
	WB	88.34 ± 0.30	86.55 ± 0.37	63.91 ± 2.64
2	AALV	87.62 ± 0.45	85.46 ± 0.10	65.03 ± 2.92
	EAA	87.53 ± 0.57	<u>85.64 ± 0.33</u>	64.02 ± 1.99
	GAA	<u>87.82 ± 0.64</u>	85.54 ± 0.44	64.59 ± 3.51
	RKLB	87.59 ± 0.57	85.57 ± 0.45	65.20 ± 3.99
	WB	87.69 ± 0.74	85.57 ± 0.51	64.74 ± 3.29
3	AALV	<u>88.07 ± 0.58</u>	86.15 ± 0.80	63.71 ± 3.63
	EAA	87.81 ± 0.54	86.04 ± 0.62	64.45 ± 1.79
	GAA	88.02 ± 0.55	86.27 ± 1.02	64.40 ± 2.30
	RKLB	87.77 ± 0.80	<u>86.53 ± 1.03</u>	64.55 ± 2.97
	WB	87.54 ± 0.54	85.99 ± 0.68	64.30 ± 2.55

fixed ROPE, we adopt a data-driven approach, i.e., for each comparison, the ROPE is set to the 25th percentile of the absolute differences between the two methods. This threshold serves as a conservative estimate of what constitutes a practically negligible performance. When methods behave similarly, most observed differences fall within this region, otherwise, clear performance gaps emerge beyond it. Figure 3 presents the results based on the NLL metric, which is particularly informative as it captures both predictive accuracy and the quality of uncertainty estimation. As shown, the comparisons reveal no statistically significant advantage for any method, indicating broadly comparable performance across all evaluated scenarios.

Focusing on the reliability of BFL, we observe that, regardless of the aggregation technique used, the trends reported in Fig. 4 indicate that *increasing the number of Bayesian layers in local models improves both global model calibration and uncertainty quantification while reducing ECE and NLL scores*. However, this added Bayesian complexity often comes at the expense of reduced time efficiency. As the number of Bayesian layers increases, computational demand rises, leading to longer processing times per communication round. This trade-off between improved model reliability and increased computational cost must be carefully considered in practical applications.

Finally, in Table IV, we compare the performances of BA-

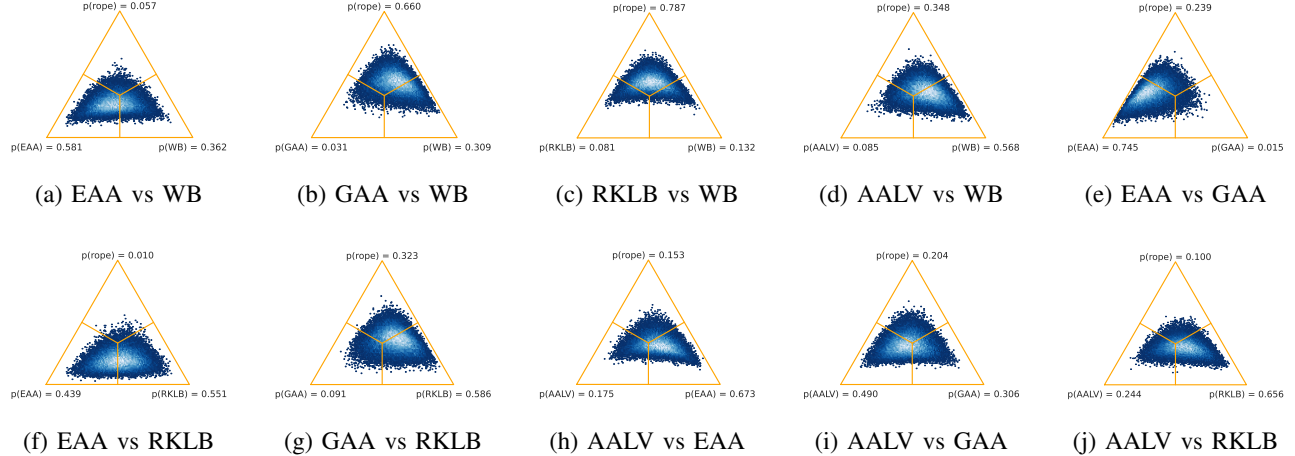


Fig. 3: Bayesian signed-rank test triangle plots comparing aggregation methods on the NLL metric. Each subfigure shows the posterior distribution over the relative performance between two methods.

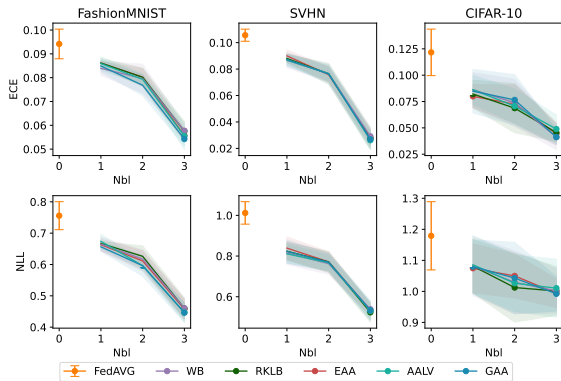


Fig. 4: Effect of Bayesian Layers on Uncertainty Quantification and Model Calibration.

BFL against two state-of-the-art parametric client-side BFL methods, pFedBayes [18] and pFedVem [46]. For the proposed method, we consider the setting with RKLB aggregation and three Bayesian layers, as it provides the best performance tradeoff. On the other hand, for pFedBayes and pFedVem, we consider the configurations detailed in their respective original papers. The results show comparable performance across all three methodologies. Our proposed approach achieves relatively better accuracy and NLL on the SVHN and CIFAR-10 datasets. In contrast, pFedVem and pFedBayes yield the best accuracy and NLL, respectively, on FashionMNIST.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced BA-BFL, a novel geometric interpretation of barycenters as a solution to the BFL aggregation problem. This approach provides an explainable aggregation method. The information-geometric view of the aggregation

Method	FashionMNIST	SVHN	CIFAR-10
pFedVem	89.50 $\pm$ 0.23	86.32 $\pm$ 0.22	60.88 $\pm$ 1.44
pFedBayes	88.02 $\pm$ 0.39	86.03 $\pm$ 0.41	63.86 $\pm$ 1.58
Ours	87.77 $\pm$ 0.80	86.53 $\pm$ 1.03	64.55 $\pm$ 2.97

(a) Comparison in Accuracy

Method	FashionMNIST	SVHN	CIFAR-10
pFedVem	0.45 $\pm$ 0.02	0.80 $\pm$ 0.03	2.44 $\pm$ 0.10
pFedBayes	0.34 $\pm$ 0.02	0.66 $\pm$ 0.04	1.25 $\pm$ 0.06
Ours	0.46 $\pm$ 0.03	0.52 $\pm$ 0.05	1.00 $\pm$ 0.08

(b) Comparison in NLL

TABLE IV: Comparison to state-of-the-art methods

step naturally enables operations such as clustering local posteriors directly on the statistical manifold, which has potential applications in hierarchical FL. Building on this concept, we recovered two aggregation techniques based on analytical results for Gaussian barycenters using two widely used divergences: the squared Wasserstein-2 distance and the reverse KL divergence. We demonstrated that BA-BFL retains the convergence properties of FedAvg for non-convex loss functions and performs robustly in both homogeneous and heterogeneous data scenarios. We experimentally evaluated the proposed methods in heterogeneous settings, showing improvements over state-of-the-art methods. We also examined the impact of varying the number of Bayesian layers in an HBDL context, evaluating their effects on accuracy, uncertainty quantification, model calibration, and cost-effectiveness. For future work, we envision several extensions, including expanding the family of distributions to include non-parametric ones and exploring alternative divergence measures. We also plan to address the personalization problem within the barycentric aggregation framework for BFL.

ACKNOWLEDGMENTS

The work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation programme (Grant agreement No.101003431), the IMT "Futur, Ruptures & Impacts" program, and the European Commission through the Horizon Europe/JU SNS project, ROBUST-6G (Grant Agreement No. 101139068).

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks From Decentralized Data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.
- [2] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated Learning on Non-iid Data Silos: An Experimental Study," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*. IEEE, 2022, pp. 965–978.
- [3] G. Zhang, S. Malekmohammadi, X. Chen, and Y. Yu, "Proportional Fairness in Federated Learning," *arXiv:2202.01666*, 2022.
- [4] X. Yue, M. Nouiehed, and R. Al Kontar, "GIFAIR-FL: A Framework for Group and Individual Fairness in Federated Learning," *INFORMS Journal on Data Science*, vol. 2, no. 1, pp. 10–23, 2023.
- [5] Z. Li, T. Lin, X. Shang, and C. Wu, "Revisiting Weighted Aggregation in Federated Learning with Neural Networks," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2023, pp. 19 767–19 788.
- [6] M. Al-Shedivat, J. Gillenwater, E. Xing, and A. Rostamizadeh, "Federated Learning via Posterior Averaging: A New Perspective and Practical Algorithms," *arXiv:2010.05273*, 2020.
- [7] L. Liu, X. Jiang, F. Zheng, H. Chen, G.-J. Qi, H. Huang *et al.*, "A Bayesian Federated Learning Framework with Online Laplace Approximation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
- [8] H. Guo, P. Greengard, H. Wang, A. Gelman, Y. Kim, and E. P. Xing, "Federated Learning As Variational Inference: A Scalable Expectation Propagation Approach," *arXiv:2302.04228*, 2023.
- [9] S. Pal, A. Gupta, S. Sarwar, and P. Rai, "Simple and Scalable Federated Learning with Uncertainty via Improved Variational Online Newton," in *OPT 2024: Optimization for Machine Learning*, 2024.
- [10] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, and F. Piccialli, "Model Aggregation Techniques in Federated Learning: A Comprehensive Survey," *Future Generation Computer Systems*, 2023.
- [11] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [12] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive Personalized Federated Learning," *arXiv:2003.13461*, 2020.
- [13] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized Federated Learning with Moreau Envelopes," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21 394–21 405, 2020.
- [14] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized Federated Learning: A Meta-learning Approach," *arXiv:2002.07948*, 2020.
- [15] Y. Zhang, T. Xia, A. Ghosh, and C. Mascolo, "Uncertainty Quantification in Federated Learning for Heterogeneous Health Data," in *International Workshop on Federated Learning for Distributed Data Mining*, 2023.
- [16] N. Koutsoubis, Y. Yilmaz, R. P. Ramachandran, M. Schabath, and G. Rasool, "Privacy Preserving Federated Learning in Medical Imaging with Uncertainty Estimation," *arXiv:2406.12815*, 2024.
- [17] S. Bhatt, A. Gupta, and P. Rai, "Federated Learning with Uncertainty via Distilled Predictive Distributions," in *Asian Conference on Machine Learning*. PMLR, 2024, pp. 153–168.
- [18] X. Zhang, Y. Li, W. Li, K. Guo, and Y. Shao, "Personalized Federated Learning via Variational Bayesian Inference," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2022, pp. 26 293–26 310.
- [19] N. Kotelevskii, M. Vono, A. Durmus, and E. Moulines, "Fedpop: A Bayesian Approach for Personalised Federated Learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 8687–8701, 2022.
- [20] A. Ozer, K. B. Buldu, A. Akgül, and G. Unal, "How to Combine Variational Bayesian Networks in Federated Learning," *arXiv:2206.10897*, 2022.
- [21] J. Fischer, M. Orescanin, J. Loomis, and P. McClure, "Federated Bayesian Deep Learning: The Application of Statistical Aggregation Methods to Bayesian Models," *arXiv:2403.15263*, 2024.
- [22] J. Zeng, A. Lesnikowski, and J. M. Alvarez, "The Relevance of Bayesian Layer Positioning to Model Uncertainty in Deep Bayesian Active Learning," *arXiv:1811.12535*, 2018.
- [23] M. Kim and T. Hospedales, "FedHB: Hierarchical Bayesian Federated Learning," *arXiv:2305.04979*, 2023.
- [24] T. Papamarkou, M. Skoularidou, K. Palla, L. Aitchison, J. Arbel, D. Dunson *et al.*, "Position Paper: Bayesian Deep Learning in the Age of Large-scale ai," *arXiv:2402.00809*, 2024.
- [25] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on Bayesian Neural Networks—a Tutorial for Deep Learning Users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022.
- [26] L. Cao, H. Chen, X. Fan, J. Gama, Y.-S. Ong, and V. Kumar, "Bayesian Federated Learning: A Survey," *arXiv:2304.13267*, 2023.
- [27] Y. Shen, N. Daheim, B. Cong, P. Nickl, G. M. Marconi, C. Bazan *et al.*, "Variational Learning Is Effective for Large Deep Networks," *arXiv:2402.17641*, 2024.
- [28] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine learning*, vol. 37, pp. 183–233, 1999.
- [29] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [30] R. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford university press, 1991.
- [31] G. Koliander, Y. El-Laham, P. M. Djurić, and F. Hlawatsch, "Fusion of Probability Density Functions," *Proceedings of the IEEE*, vol. 110, no. 4, pp. 404–453, 2022.
- [32] G. Battistelli and L. Chisci, "Kullback–Leibler Average, Consensus on Probability Densities, and Distributed State Estimation with Guaranteed Stability," *Automatica*, vol. 50, no. 3, pp. 707–718, 2014.
- [33] M. Agueh and G. Carlier, "Barycenters in the Wasserstein Space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.
- [34] H.-Y. Chen and W.-L. Chao, "FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning," *arXiv:2009.01974*, 2020.
- [35] J. Backhoff-Veraguas, J. Fontbona, G. Rios, and F. Tobar, "Bayesian Learning with Wasserstein Barycenters," *ESAIM: Probability and Statistics*, vol. 26, pp. 436–472, 2022.
- [36] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic Controlled Averaging for Federated Learning," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 5132–5143.
- [37] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," *arXiv:1708.07747*, 2017.
- [38] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading Digits in Natural Images with Unsupervised Feature Learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, no. 2. Granada, 2011, p. 4.
- [39] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *University of Toronto*, 2009.
- [40] Q. Li, B. He, and D. Song, "Practical One-shot Federated Learning for Cross-silo Setting," *arXiv:2010.01017*, 2020.
- [41] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian Nonparametric Federated Learning of Neural Networks," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2019, pp. 7252–7261.
- [42] H. Wang, M. Yurochkin, Y. Sun, D. Papaliopoulos, and Y. Khazaeni, "Federated Learning with Matched Averaging," *arXiv:2002.06440*, 2020.
- [43] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 7611–7623, 2020.
- [44] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble Distillation for Robust Model Fusion in Federated Learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 2351–2363, 2020.
- [45] J. Carrasco, S. García, M. del Mar Rueda, and F. Herrera, "rNPBST: An R Package Covering Non-parametric and Bayesian Statistical Tests," in *Hybrid Artificial Intelligent Systems: 12th International Conference, HAIS 2017, La Rioja, Spain, June 21-23, 2017, Proceedings 12*. Springer, 2017, pp. 281–292.
- [46] J. Zhu, X. Ma, and M. B. Blaschko, "Confidence-Aware Personalized Federated Learning via Variational Expectation Maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 24 542–24 551.