Federated Variational Inference: Towards Improved Personalization and Generalization

Elahe Vedadi[1], Joshua V. Dillon[1], Philip Andrew Mansfield[1], Karan Singhal[1], Arash Afkanpour[2*], Warren Richard Morningstar[1]

[1]Google Research
[2]Vector Institute
{elahevedadi, jvdillon, memes, karansinghal, wmorning}@google.com, arash.afkanpour@vectorinstitute.ai

Abstract

Conventional federated learning algorithms train a single global model by leveraging all participating clients' data. However, due to heterogeneity in client generative distributions and predictive models, these approaches may not appropriately approximate the predictive process, converge to an optimal state, or generalize to new clients. We study personalization and generalization in stateless cross-device federated learning setups assuming heterogeneity in client data distributions and predictive models. We first propose a hierarchical generative model and formalize it using Bayesian Inference. We then approximate this process using Variational Inference to train our model efficiently. We call this algorithm *Federated Variational Inference (FedVI)*. We use PAC-Bayes analysis to provide generalization bounds for FedVI. We evaluate our model on FEMNIST and CIFAR-100 image classification and show that FedVI beats the state-of-the-art on both tasks.

Introduction

Federated Learning (FL) (McMahan et al. 2016) enables decentralized model training, avoiding the need to aggregate data on a central server due to privacy concerns. In FL, the central server oversees a global model distributed to clients who conduct local training, and the model updates are aggregated to iteratively improve the global model.

In simple and idealized settings, FL can approximate centralized training with similar theoretical guarantees, as seen in FedSGD (McMahan et al. 2016). However, real-world cross-device FL scenarios, such as those in (Reddi et al. 2020; Wang et al. 2021), often diverge from these ideal conditions. Practical FL implementations involve multiple local training steps to minimize communication overhead. Client participation is typically uneven, with some contributing more data and others not participating at all. Additionally, the non-Independently and Identically Distributed (non-IID) nature of client datasets, stemming from distinct data generation processes, challenges theoretical guarantees, leads to performance disparities between participating and non-participating clients (Yuan et al. 2022), and complicates training high-performing models in practical FL setups.

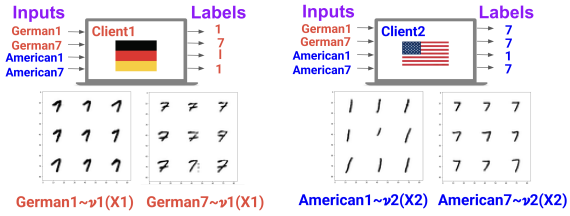


Figure 1: Illustration of diverse data generation and predictive models in cross-device FL.

Modern approaches address this challenge by either modifying the local loss to converge to a global solution (Li et al. 2020) or using personalized models to handle local distribution shifts (Zhang et al. 2022). Approaches for personalization have often focused on stateful FL setups, where clients are revisited throughout training and thus can update a locally stored model (Karimireddy et al. 2019; Wang et al. 2021). However, many production scenarios are effectively stateless, since individual clients only rarely contribute to training, and local models may be either stale or non-existent. Few studies have concentrated on personalization in this context. Those that have (Singhal et al. 2021), require clients to possess labeled examples for personalization.

This paper addresses personalization in stateless cross-device FL with FedVI, a Variational Inference (VI) algorithm enabling generalization and personalization across diverse client data, even for untrained clients. The key contributions encompass (i) proposing a hierarchical generative model rooted in mixed effects models for cross-device federated setups, (ii) offering generalization bounds through Probably Approximately Correct (PAC)-Bayes analysis, (iii) introducing FedVI algorithm, inspired by the theoretical approach, which provides a simplified experimental approximation and can be implemented by the existing FL frameworks, and (iv) demonstrating the superior performance of FedVI on two federated datasets, FEMNIST and CIFAR-100, compared to previous state-of-the-art methods.

Related Work

Bayesian FL: To tackle statistical heterogeneity in FL, various studies have employed Bayesian methods to incorporate domain knowledge and aid convergence. Early attempts

*Work done while at Google Research.

(Thorgeirsson and Gauterin 2020; Chen and Chao 2020) focused on model aggregation, either to retain uncertainty in model parameters, or to weight parameter updates proportional to performance. (Zhang et al. 2022) instead attempts to use a Bayesian Neural Network (BNN) approximated with VI to train a global model using a Kullback–Leibler (KL) regularizer which induces convergence similar to the proximal term in FedProx (Li et al. 2020). While their local models can, in principle, personalize by deviating from the global model, they realistically require stateful settings with significant labeled data on clients in order to do so. (Kotelevskii et al. 2022) casts personalized FL as mixed effects regression, and attempts to model the inherent heterogeneity in this setting explicitly using Stochastic Gradient Langevin Dynamics (Welling and Teh 2011). Our proposed method shares a similar generative process with (Kotelevskii et al. 2022), but leverages VI for efficient posterior inference and PAC-Bayes bounds (Germain et al. 2016) to promote generalization.

Stateful FL: There is a rich body of literature on personalization in FL (Corinzia, Beuret, and Buhmann 2019; Ghosh et al. 2021; Chen and Chao 2022; Collins et al. 2023; Deng, Kamani, and Mahdavi 2020; Li et al. 2021; Hassan, Salomone, and Mengersen 2023). Many previous methods focus on stateful settings, where a set of local parameters is stored on clients and is maintained throughout rounds of training. In contrast, we focus on stateless settings where it is not possible to maintain an up-to-date local state on each client. This is similar to the setting considered by (Marfoq et al. 2022), who uses K-nearest neighbors to account for client distributional shift. While this is a robust means of dealing with both input and output distributional shift, it requires clients to possess labeled examples for every class (which is unrealistic in real-world setups), and cannot be used outside of classification problems.

Meta Learning: Prior work extensively explores connections between personalized FL and Model-Agnostic Meta-Learning (MAML) (Finn, Abbeel, and Levine 2017; Singhal et al. 2021; Fallah, Mokhtari, and Ozdaglar 2020; Collins et al. 2023; Lin et al. 2023; Chen et al. 2019). These approaches focus on finding a global initialization that clients can rapidly adapt to their local data. Motivated by MAML, FedRecon (Singhal et al. 2021) uses a partially local FL setting, training only a subset of "global" parameters for fast local parameter reconstruction. Our work builds upon the principles established in FedRecon. Unlike this work, we also provide a means of reconstructing local parameters without access to labeled data.

Methods

Hierarchical Generative Model

Let us consider a stateless cross-device FL setup with multiple clients and a server, where client subsets are randomly selected each round. We categorize each client's model parameters as global (θ) and local (β_k for $k \in [c]$[1]) parameters, with c representing the total number of clients. Global parameters update at the server, while local parameters stay on-device. Global parameters are drawn from prior $t(\Theta)$, while

[1]In this paper we represent the set of $\{1, \ldots, c\}$ by $[c]$.

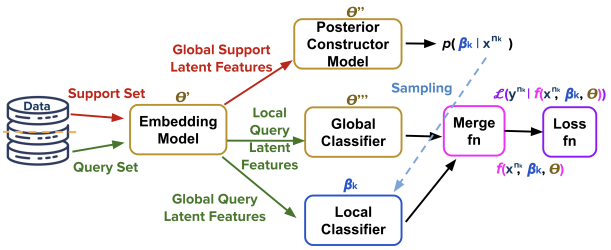


Figure 2: Our proposed model architecture implementing FedVI, where $\theta = \theta' \cup \theta'' \cup \theta'''$ represents global parameters.

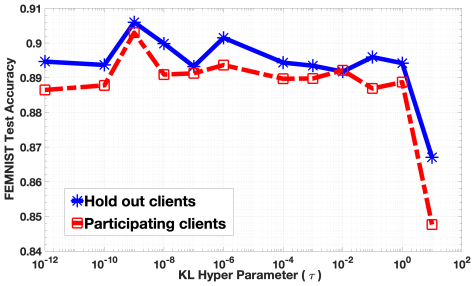
each client's local parameters are independent samples from the local prior $r(B_k)$. Data may be non-IID across clients ($x_{ik} \sim \nu_k(X_k)$ for $i \in [n_k]$), where n_k is the total number of data samples at client k. Moreover, each client may have a distinct predictive distributions. Although all clients share the same likelihood distribution family $\ell(Y_k|f(\theta, \beta_k, x_{ik}))$, the distribution varies based on β_k, making it different for each client.

The above setup is a prototypical example of a mixed effects model (Demidenko 2013), commonly employed for predicting a continuous random variable using multiple independent factors, including both random and fixed, and incorporating repeated measurements from the same observational unit. Framing our work within this well-established mixed effects model framework allows us to leverage existing theoretical insights from the field. To summarize, we propose the following hierarchical data generating process:

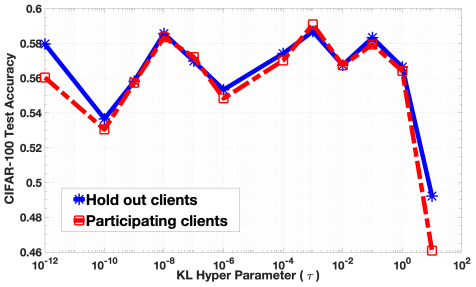
$$
\begin{aligned}
&\theta \sim t(\Theta) \\
&\text{for } k \in [c]: \\
&\quad \beta_k \sim r(B_k) \\
&\quad \text{for } i \in [n_k]: \\
&\quad\quad x_{ik} \sim \nu_k(X_k) \\
&\quad\quad y_{ik} \sim \ell(Y_k|f(\theta, \beta_k, x_{ik})),
\end{aligned}
\tag{1}
$$

where $f : \Phi \times \mathcal{B}_k \times \mathcal{X}_k \to \mathcal{Z}_k$ is a deterministic function (e.g., DNN) mapping what we know to the latent space $\mathcal{Z}_k$, which is the parameter space of our distribution over outcomes, $\ell(.)$.

For a more intuitive grasp of varying data generation processes and predictive distributions, consider the Federated EMNIST dataset (FEMNIST) in Figure 1. This figure highlights the diversity of data generation processes in FEMNIST dataset. Each client's unique handwriting style (e.g., how they write the numbers '1' or '7') influences their data. This leads to differing predictive distributions – one client might misinterpret another client's characters due to stylistic variations. Consequently, a single global model struggles to handle this diversity. To accurately model this scenario, algorithms need to incorporate local adjustments or personalization, recognizing that each client's data generation process may be distinct. Our proposed algorithm explicitly assumes this data generating process. Note that this assumption reduces in special cases to existing FL setups, such as IID predictive distributions ($r(B_k) = \delta(B_k - \beta)$), or IID data generating processes ($\nu_k(X_k) = \nu(X_k)$).



(a) FEMNIST



(b) CIFAR-100

Figure 3: Participating and non-participating test accuracy vs. KL hyperparameter τ.

Training Objective

This section defines the objective function for our training process. We will start by calculating the estimated probability density function of labels given input data, denoted as $\hat{p}(Y) \overset{\text{def}}{=} p(Y|X) = \int_\theta \int_{\beta_c} \cdots \int_{\beta_1} p(\theta, \beta^c|X, Y)\ell(Y|f(\theta, \beta^c, X))$, following a similar marginalization approach as (Watanabe 2018), where $\beta^c \overset{\text{def}}{=} \{\beta_k\}^c \overset{\text{def}}{=} \{\beta_k : k \in [c]\}$, $x^{n_k} \overset{\text{def}}{=} \{x_i : i \in [n_k]\}$, $X \overset{\text{def}}{=} \{x^{n_k}\}^c \overset{\text{def}}{=} \{x_{ik} : i \in [n_k], k \in [c]\}$, and y^{n_k} and Y are defined similar to x^{n_k} and X, respectively. Therefore, for calculating $\hat{p}(Y)$ it is required to calculate the posterior probability of model parameters given the training data which is equal to $p(\theta, \beta^c|X, Y) = \frac{p(\theta, \beta^c, Y|X)}{p(Y|X)}$.

Assuming that the prior distribution of the global parameters, $t(\theta)$, the prior distribution of the local parameters, $r(\beta_k)$, and the likelihood distribution of each client, $\ell(y^{n_k}|f(\theta, \beta_k, x^{n_k}))$, are independent we have $p(\theta, \beta^c, Y|X) = t(\theta)\prod_{k\in[c]} r(\beta_k)\prod_{k\in[c]}\prod_{i\in[n_k]}\ell(y_{ik}|f(\theta, \beta_k, x_{ik})) = t(\theta)r(\beta^c)\ell(Y|f(\theta, \beta^c, X))$. Moreover, $p(Y|X)$ can be written as $\int_\theta \int_{\beta_c} \cdots \int_{\beta_1} p(\theta, \beta^c, Y|X)$. Unfortunately this integral is not only infeasible to compute, but also mathematically intractable. Consequently, this makes the whole posterior intractable.

To handle the intractable posterior distribution, we approximate it with a tractable surrogate, $q(\theta, \beta^c|X, Y)$, using VI. We derive the evidence lower bound (ELBO), equivalent to the KL divergence between the posterior and surrogate ($D_{\text{KL}}(q(\theta, \beta^c|X, Y)\|p(\theta, \beta^c, Y|X))$). Minimizing the ELBO yields the best approximation for the intractable posterior. ELBO derivations is provided in Appendix A.

By asserting factorization, we define the surrogate as a parametric distribution as $q(\theta, \beta^c|X, Y) \overset{\text{def}}{=} q_\lambda(\theta|X, Y)\prod_{k\in[c]} q_\lambda(\beta_k|\theta, y^{n_k}, x^{n_k}) \overset{\text{def}}{=} q_\lambda(\theta|X, Y)q_\lambda(\beta^c|\theta, X, Y)$, where λ is the parameter set that uniquely defines the surrogate distribution. Therefore, the objective function for training the proposed hierarchical model is ELBO, which can be written as follows:

$$\mathcal{J}(\lambda; \gamma, \tau) = D_{\text{KL}}(q_\lambda(\theta, \beta^c|X, Y)\|p(\theta, \beta^c, Y|X)) =$$
$$\sum_{k\in[c]}\sum_{i\in[n_k]} \overbrace{\mathbb{E}_{q(\theta,\beta^c|\{y^{n_k}, x^{n_k}\}^c)}[-\log \ell(y_{ik}|f(\theta, \beta_k, x_{ik}))]}^{\text{Per Datum Expected Loss}}$$
$$+ \underbrace{\gamma D_{\text{KL}}(q_\lambda(\theta|X, Y)\|t(\theta))}_{\text{Global Regularizer}}$$
$$+ \sum_{k\in[c]} \tau \underbrace{\mathbb{E}_{q_\lambda(\theta|X,Y)}[D_{\text{KL}}(q_\lambda(\beta_k|\theta, y^{n_k}, x^{n_k})\|r(\beta_k))]}_{\text{Local Regularizer}}, \tag{2}$$

where γ, τ, $t(\theta), r(\beta_k)$, and the functional form of $q_\lambda(\theta, \beta^c|X, Y)$ are left as hyper parameters. The details of this derivation are provided in Appendix B.

Generalization Bounds

Our training objective is the ELBO, which minimizes training error, but our true goal is better generalization (minimizing error on unseen data). To achieve this, we leverage PAC-Bayes analysis and propose a slightly generalized version of Theorem 3 in (Germain et al. 2016), in the form of the following corollary, to compute a generalization bound for our model's true risk.

Corollary 1 *Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, a hypothesis set $\mathcal{F} = \{\theta, \beta^c\}$, a loss function $\ell : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, a prior distribution $\pi(\Theta, B^c) = t(\Theta)r(B^c)$ over $\mathcal{F}$, a $\delta \in (0, 1]$ and a real number $\eta > 0$, with probability at least $1-\delta$ over the choice of $(\{x^{n_k}\}^c, \{y^{n_k}\}^c) \overset{\text{def}}{=} (X, Y) \sim \mathcal{D}$, for any $q(.)$ on $\mathcal{F}$ we have:*

$$\overbrace{\mathbb{E}_{\mathcal{D}}[-\log\left(\mathbb{E}_{q(\theta,\beta^c|X,Y)}[\ell(Y|X, \theta, \beta^c)\right)]]}^{\textit{True risk}} \leq$$
$$\overbrace{\mathbb{E}_{X,Y}[\mathbb{E}_{q(\theta,\beta^c|X,Y)}[-\log(\ell(Y|X, \theta, \beta^c))]]}^{\textit{Empirical risk}}$$
$$+ \frac{1}{\eta}\Bigg[\overbrace{D_{\text{KL}}(q(\theta, \beta^c|X, Y)\|\pi(\theta, \beta^c))}^{\textit{KL divergence}}$$
$$+ \log\Big(\tfrac{1}{\delta}\mathbb{E}_{X,Y}\Big[\mathbb{E}_{\pi(\theta,\beta^c)}\Big[\exp\Big(\eta\mathbb{E}_{\mathcal{D}}[-\log(\ell(Y|X, \theta, \beta^c))]$$
$$\underbrace{-\eta\mathbb{E}_{X,Y}[-\log(\ell(Y|X, \theta, \beta^c))]\Big)\Big]\Big]\Big)}_{\textit{Slack term}}\Bigg], \tag{3}$$

Dataset	FedAvg	FedAvg+	ClusteredFL	DITTO	FedRep	APFL	KNN-Per	**FedVI**
FEMNIST	83.4/83.1	84.3/84.2	83.7/83.2	84.3/83.9	85.3/85.4	84.1/84.2	88.2/88.1	**90.3/90.6**
CIFAR-100	47.4/47.1	51.4/50.8	47.2/47.1	52.0/52.1	53.2/53.5	51.7/49.1	55.0/56.1	**59.1/58.7**

Table 1: Test accuracy of the participating/non-participating clients.

for $\mathbb{E}_{\mathcal{D}}[.]=\mathbb{E}_{(X,Y)\sim\mathcal{D}}[.]$ and $\mathbb{E}_{X,Y}[\log(\ell(Y|X,\theta,\beta^c))] = \frac{1}{\sum_{k=1}^{c} n_k} \sum_{k=1}^{c} \sum_{i=1}^{n_k} [\log(\ell(y_{ik}|x_{ik},\theta,\beta_k))]$.

Proof: This corollary's proof is provided in Appendix C.

Having obtained the generalization bound in Equation 3, we observe that it equals the ELBO (Equation 2) plus a constant slack term, unrelated to the surrogate or posterior distributions. Therefore, minimizing the ELBO also minimizes generalization error (assuming a finite slack term). This suggests that optimizing the ELBO will enhance the model's generalization capabilities.

Implementation and Experimental Evaluation

Distributions: For the prior distribution of the local parameters, we assume a normal distribution with zero mean and variance equal to that given by the initialization scheme (e.g. Glorot and Bengio 2010; Glorot, Bordes, and Bengio 2011; He et al. 2015). No assumptions are made on client data distributions. Our likelihood is a categorical distribution with logits generated by a deep neural network parameterized by θ and β_k. To simplify implementation, we use a point estimate for the global posterior (equivalent to $\gamma = 0$ in Equation 2). To ensure a finite KL divergence between the global posterior and the global prior, $D_{\mathrm{KL}}(q_\lambda(\theta|X,Y)\|t(\theta))$, we assume a narrow but finite global posterior, with the global prior remaining any finite function.

Tasks: We evaluate FedVI on FEMNIST (Caldas et al. 2019) (62-class, naturally diverse data generation per client) and CIFAR-100 (Krizhevsky 2009) (100-class, synthetically partitioned with LDA). FEMNIST highlights FedVI's ability to handle diverse data distributions, while CIFAR-100 showcases its performance on complex classification.

Model Architecture: There are infinitely many model architectures which could implement our method. The architecture that we chose in our experiments is illustrated in Figure 2 and explained in details in Appendix D.

Implementation: FedVI algorithm is implemented in TensorFlow Federated (TFF) and scaled to NVIDIA Tesla V100 GPUs for hyperparameter tuning. For FEMNIST (3400 clients), we hold out 20 clients for generalization testing. During training, 100 clients are randomly selected per round (with replacement across rounds). For CIFAR-100 (500 clients), we hold out 10 clients and randomly select 50 per round. Training involves 1500 rounds with mini-batches of 256 samples using mini-batch gradient descent. The details of training procedure are provided in Appendix D.

Evaluation Results and Discussion: We compare our proposed FedVI algorithm with the state-of-the-art personalized FL method, KNN-Per (Marfoq et al. 2022), as well as other methods including FedAvg (McMahan et al. 2016), FedAvg+ (Chen and Chao 2022), ClusteredFL (Ghosh et al. 2021), DITTO (Li et al. 2021), FedRep (Collins et al. 2023), and APFL (Deng, Kamani, and Mahdavi 2020), using the results reported in (Marfoq et al. 2022).

Table 1 presents weighted average accuracy on local test datasets (unseen during training) for both participating and non-participating clients using FedVI and other methods. For robustness, FedVI's test accuracy is averaged across the final 100 training rounds.

Figure 3a shows the average test accuracy over the last 100 FEMNIST training rounds for a range of KL hyperparameter τ, from 10^{-12} to 10 (As the horizontal axis of both figures in Figure 3 are semi-logarithmic, test accuracy results of $\tau = 0$ are shown at point $\tau = 10^{-12}$). Notably, $\tau = 10^{-9}$ outperforms others, achieving higher accuracy with a smaller generalization gap compared to $\tau = 0$.

Figure 3b displays the average test accuracy over the last 100 rounds in CIFAR-100, with varying KL hyperparameter τ. Notably, $\tau = 10^{-3}$ achieves the highest accuracy for both participating and non-participating clients. Comparing $\tau = 0$ to other values ($\tau \neq 0$) reveals that minimizing KL divergence reduces the gap in participation test accuracy, as anticipated. Furthermore, comparing this figure to Figure 3a, it's evident that the difference in test accuracy between $\tau = 0$ and $\tau = 10^{-9}$ in the FEMNIST experiment is significantly larger than the difference between $\tau = 0$ and $\tau = 10^{-3}$ in the CIFAR-100 experiment. This suggests that minimizing KL divergence is more critical for FEMNIST than for CIFAR-100. One possible explanation is that in FEMNIST, each client's data generation distribution naturally differs, while in CIFAR-100, data is synthetically partitioned and distributed among clients.

Conclusion and Future Work

This work addresses personalization in stateless cross-device federated setups through the introduction of FedVI, a novel algorithm grounded in mixed effects models and trained using VI. We establish generalization bounds for FedVI through PAC-Bayes analysis, present a novel architecture, and implement it. Evaluation on FEMNIST and CIFAR-100 datasets demonstrates that FedVI outperforms state-of-the-art methods in both cases. It is worth noting that in this paper, we employed a narrow normal distribution as the posterior for global parameters. However, in future research, we intend to explore more generalized distributions to enhance the modeling capabilities. Additionally, the model architecture presented in Figure 2 is just one of several possible architectures that align with our theoretical hierarchical model. In upcoming work we will focus on refining these architectures to optimize performance and explore their potential for achieving even better results.

References

Caldas, S.; Duddu, S. M. K.; Wu, P.; Li, T.; Konečný, J.; McMahan, H. B.; Smith, V.; and Talwalkar, A. 2019. LEAF: A Benchmark for Federated Settings. arXiv:1812.01097.

Chen, F.; Luo, M.; Dong, Z.; Li, Z.; and He, X. 2019. Federated Meta-Learning with Fast Convergence and Efficient Communication. arXiv:1802.07876.

Chen, H.-Y.; and Chao, W.-L. 2020. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*.

Chen, H.-Y.; and Chao, W.-L. 2022. On Bridging Generic and Personalized Federated Learning for Image Classification. arXiv:2107.00778.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2023. Exploiting Shared Representations for Personalized Federated Learning. arXiv:2102.07078.

Corinzia, L.; Beuret, A.; and Buhmann, J. M. 2019. Variational federated multi-task learning. *arXiv preprint arXiv:1906.06268*.

Demidenko, E. 2013. *Mixed models: theory and applications with R*. John Wiley & Sons.

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Adaptive Personalized Federated Learning. arXiv:2003.13461.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. E. 2020. Personalized Federated Learning: A Meta-Learning Approach. *CoRR*, abs/2002.07948.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *CoRR*, abs/1703.03400.

Germain, P.; Bach, F.; Lacoste, A.; and Lacoste-Julien, S. 2016. PAC-Bayesian Theory Meets Bayesian Inference. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2021. An Efficient Framework for Clustered Federated Learning. arXiv:2006.04088.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep Sparse Rectifier Neural Networks. In Gordon, G.; Dunson, D.; and Dudík, M., eds., *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, 315–323. Fort Lauderdale, FL, USA: PMLR.

Hassan, C.; Salomone, R.; and Mengersen, K. 2023. Federated Variational Inference Methods for Structured Latent Variable Models. arXiv:2302.03314.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026–1034.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2019. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning.

Kotelevskii, N.; Vono, M.; Moulines, E.; and Durmus, A. 2022. FedPop: A Bayesian Approach for Personalised Federated Learning.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. arXiv:2012.04221.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated Optimization in Heterogeneous Networks. arXiv:1812.06127.

Lin, Y.; Ren, P.; Chen, Z.; Ren, Z.; Yu, D.; Ma, J.; de Rijke, M.; and Cheng, X. 2023. Meta Matrix Factorization for Federated Rating Predictions. arXiv:1910.10086.

Marfoq, O.; Neglia, G.; Kameni, L.; and Vidal, R. 2022. Personalized Federated Learning through Local Memorization. arXiv:2111.09360.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive Federated Optimization.

Singhal, K.; Sidahmed, H.; Garrett, Z.; Wu, S.; Rush, K.; and Prakash, S. 2021. Federated Reconstruction: Partially Local Federated Learning.

Thorgeirsson, A. T.; and Gauterin, F. 2020. Probabilistic predictions with federated learning. *Entropy*, 23(1): 41.

Wang, J.; Charles, Z.; Xu, Z.; Joshi, G.; McMahan, H. B.; Al-Shedivat, M.; Andrew, G.; Avestimehr, S.; Daly, K.; Data, D.; et al. 2021. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*.

Watanabe, S. 2018. *Mathematical theory of Bayesian statistics*. CRC Press.

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 681–688.

Yuan, H.; Morningstar, W. R.; Ning, L.; and Singhal, K. 2022. What Do We Mean by Generalization in Federated Learning? In *International Conference on Learning Representations*.

Zhang, X.; Li, Y.; Li, W.; Guo, K.; and Shao, Y. 2022. Personalized Federated Learning via Variational Bayesian Inference. arXiv:2206.07977.