Variational Inference in Similarity Spaces: A Bayesian Approach to Personalized Federated Learning

Pedro H. Barros
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
pedro.barros@dcc.ufmg.br

Fabricio Murai
Worcester Polytechnic Institute
Worcester, USA
fmurai@wpi.edu

Amir Houmansadr
University of Massachusetts
Amherst, USA
amir@cs.umass.edu

Alejandro C. Frery
Victoria University of Wellington
Wellington, New Zealand
alejandro.frery@vuw.ac.nz

Heitor S. Ramos
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ramosh@dcc.ufmg.br

Abstract

Similarity space (or S-space) employs an encoder function, fed by labeled original pairwise data, to find a latent pairwise space with markers (prototypical) vector. It divides the space into regions where pairs of objects are either similar or dissimilar. This paper enhances S-space, equipping variational inference from personalized federated learning. The S-space representation aligns local representation spaces across clients, while variational inference improves generalization and reduces overfitting caused by data scarcity and client heterogeneity. Our theoretical analysis improved upper bounds on KL divergence between optimal local and optimal global variational models compared to traditional distributed Bayesian neural networks.

1 Introduction

Federated Learning (FL) has emerged as a methodology that enables learning from decentralized datasets without compromising data privacy [1]. Real-world FL applications often involve non-IID data, which decrease individual client performance due to heterogeneous local data distributions [1, 2]. Personalized Federated Learning (PFL) addresses this by tailoring models to better align with local data characteristics [3]. Various methods have been proposed to enhance PFL, including global model personalization [4, 5], federated meta-learning [6, 7], and parameter decoupling [3, 8].

Traditional Neural Networks (NN) are typically used in PFL, although they usually show poor calibration and overconfidence in predictions, particularly when faced with varying data distributions [9–11]. In contrast, Bayesian Neural Networks (BNNs) is a probabilistic approach that has been used in other contexts for modeling uncertainty and enabling models to learn continually by capturing past information [12], suggesting its potential use for PFL.

This paper proposes a novel framework for PFL with BNNs to address challenges in model overfitting due to limited local data (FL privacy constraints). Our approach leverages variational inference (VI) within an auxiliary representation space to enhance PFL model performance by quantifying

Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

weight uncertainty in NNs at client and server models. To achieve personalization, each client updates its local VI parameters by reusing the global distribution from the server and balancing the KL divergence between the local posterior distribution and the server variational parameters. This strategy improves the upper bounds on this KL divergence compared to traditional distributed BNNs [13, 14]. Finally, our experiments show promising results in five datasets in the literature.

2 Methodology

Federated Learning: Consider a client u_i wishing to train a machine learning model with their respective datasets $\mathcal{X}_i$. Traditional (centralized) machine learning training methods group all the data in the set $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_N$. In contrast, a FL system enables clients to collaboratively train a model with the same architecture across all clients without sharing their local datasets $\mathcal{X}_i$ with one another. However, FL introduces challenges due to the heterogeneity of local datasets.

(Neural Network-induced) Similarity Space: Let $\mathcal{X}$ be the set of $\boldsymbol{x}_i \in \mathcal{X}$ elements in an m-dimensional feature space, each associated with a label $y_i \in \mathcal{Y}$ and the function $\ell\colon \mathcal{X} \to \mathcal{Y}$, which maps the data into their respective labels. The original feature space $\mathcal{X}$ is transformed into a *latent feature space* $\mathcal{Z}$ by a representation function $f_\Theta\colon \mathcal{X} \to \mathcal{Z}$, where $\boldsymbol{z}_i = f_\Theta(\boldsymbol{x}_i) \in \mathbb{R}^d$ and Θ is the set of weights of a given NN.

We estimate an *auxiliary space*, referred to as *S-Space* [15], to refine the latent feature space with a mapping function $f^S\colon \mathcal{X} \times \mathcal{X} \to \mathcal{S}$, where if $\ell(\boldsymbol{x}_i) = \ell(\boldsymbol{x}_j)$, then $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a similar pair. On the other hand, we consider $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to be a dissimilar pair if $\ell(\boldsymbol{x}_i) \neq \ell(\boldsymbol{x}_j)$. Given a pair of elements $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the original feature space, f^S computes a similarity vector $\boldsymbol{s}_{ij}$ using the absolute element-wise difference between their latent space representations

$$\begin{aligned}\boldsymbol{s}_{ij} &= f^S(\boldsymbol{x}_i, \boldsymbol{x}_j)\\ &= |f_\Theta(\boldsymbol{x}_i) - f_\Theta(\boldsymbol{x}_j)|\\ &= |\boldsymbol{z}_i - \boldsymbol{z}_j|\\ &= (|z_{i,1} - z_{j,1}|, \ldots, |z_{i,d} - z_{j,d}|),\end{aligned}$$

where $\boldsymbol{s}_{ij} \in \mathbb{R}^d$ has the same dimension as $\boldsymbol{z}_i$, $z_{n,k}$ is the k-th feature of the n-th sample in the latent space representation $\mathcal{Z}$, and the absolute operation ensures symmetry $\boldsymbol{s}_{ij} = \boldsymbol{s}_{ji}$.

We define two disjoint sets in the S-Space to quantify the similarity between input pairs: similarity markers $(\mathcal{M}^+)$ and dissimilarity markers $(\mathcal{M}^-)$. Their union $\mathcal{M} = \mathcal{M}^+ \cup \mathcal{M}^-$ is the set of all markers (or prototypes) and $\mathcal{M}^+ \cap \mathcal{M}^- = \emptyset$. The (dis)similarity between input pairs is determined by the distance of the vector $\boldsymbol{s}_{ij}$ to the markers $\boldsymbol{m}_e \in \mathcal{M}$. We use a Cauchy kernel [16] to measure the similarity q_{ij}^e between the point $\boldsymbol{s}_{ij}$ and the marker $\boldsymbol{m}_e$ as

$$q_{ij}^e = \frac{\left(1 + \|\boldsymbol{s}_{ij} - \boldsymbol{m}_e\|_2^2\right)^{-1}}{\sum_{m_{e'} \in \mathcal{M}} \left(1 + \|\boldsymbol{s}_{ij} - \boldsymbol{m}_{e'}\|_2^2\right)^{-1}}.$$

Consider the pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we define the probability of s_{ij} vector being "similar" as $q_{ij}^+ = \sum_p q_{ij}^p$ for all $\boldsymbol{m}_p \in \mathcal{M}^+$. Analogously, the probability of s_{ij} vector being "dissimilar" is $q_{ij}^- = \sum_n q_{ij}^n$ for all $\boldsymbol{m}_n \in \mathcal{M}^-$. Notice that $q_{ij}^+ + q_{ij}^- = 1$ because the sets $\mathcal{M}^+$ and $\mathcal{M}^-$ are disjoint.

The binary pair cross-entropy loss function $J(\cdot)$ is

$$J(\mathcal{X}) = -\sum_{i \in \mathcal{X}} \sum_{J \in \mathcal{X}} \left[u_{ij} \log q_{ij}^+ + (1 - u_{ij}) \log q_{ij}^-\right], \tag{1}$$

where $u_{ij} = \mathbb{1}[\ell(\boldsymbol{x}_i) = \ell(\boldsymbol{x}_j)]$ is the indicator function, i.e., $\mathbb{1}[\cdot] = 1$ if $[\cdot]$ is true and 0 otherwise; We note that similar definitions exist for this auxiliary space [15]. However, as demonstrated in the following section, this is the first instance of applying VI within FL settings.

Variational inference: We exploit the S-Space by introducing VI techniques to estimate the markers $\boldsymbol{m}$. Our challenge is to make statistical inferences from the posterior distribution $p_\Theta(\mathcal{M} \mid \mathcal{X}_1, \ldots, \mathcal{X}_N)$ based on the NN parameters Θ without compromising data privacy. To tackle this, each

client minimizes the Kullback-Leibler (KL) divergence between each client's variational distribution q_{ϕ_k} and the true posterior, formulated as

$$\underset{q_{\phi_k}(\mathcal{M})\in\mathcal{Q}}{\arg\min}\Big\{\mathrm{KL}\big(q_{\phi_k}(\mathcal{M})\parallel p_\Theta(\mathcal{M}\mid\mathcal{X}_1,\ldots,\mathcal{X}_N)\big)\Big\}, \tag{2}$$

where $\mathcal{Q}$ is a variational family of distributions.

We assume that each client's marker distribution (variational parameters) in the S-Space follows a Gaussian distribution ($\mathcal{N}(\mu,\sigma^2)\in\mathcal{Q}$), representing the variational distribution q_{ϕ_k} associated with the client k as a product of normal distributions (mean-field approximation) [13, 17, 18]. The likelihood function $p_\theta(\mathcal{X}_k\mid\mathcal{M})\propto\exp(-J(\mathcal{X}_k)/\alpha)$ is defined using an exponential loss function (Boltzmann distribution) [19], where $J(\cdot)$ is the S-Space loss function (or energy function – Eq. 1) and $\alpha>0$ is a (temperature) scaling parameter [20, 21].

Denote as $\mathcal{X}_{\setminus k}=\{\mathcal{X}_1,\ldots,\mathcal{X}_{k-1},\mathcal{X}_{k+1},\ldots,\mathcal{X}_N\}$ the local datasets excluding the data from client u_k. Note that client u_k does not have access to $\mathcal{X}_{\setminus k}$ due to FL privacy constraints. We approximate the posterior distribution using Bayes' theorem and a server variational model as $p_\Theta(\mathcal{M}\mid\mathcal{X}_{\setminus k})\approx s(\mathcal{M})$. The KL divergence (Eq. 2) can be approximated as[1]

$$\mathrm{KL}\bigg(q_{\phi_k}(\mathcal{M})\,\Big\|\,s(\mathcal{M})\frac{p_\theta(\mathcal{X}_k\mid\mathcal{M})}{Z_k}\bigg)=\mathrm{KL}\big(q_{\phi_k}(\mathcal{M})\parallel s(\mathcal{M})\big)+\log Z_k+\frac{1}{\alpha}\mathbb{E}_{q_{\phi_k}}\left[J(\mathcal{X}_k)\right],$$

where Z_k is a normalization constant.

Following [23], we have adjusted the scale α to enhance numerical stability. We omit the normalization constant $\log Z_k$ from the optimization problem (Evidence Lower Bound) [13]. Our approach captures model performance on specific tasks and ensures regularization by minimizing divergence from the server model; cf. Appendix B for details. Our VI approach in FL is a dual optimization framework that enhances client-level personalization:

$$\text{Client: }\underset{q_{\phi_k}(\mathcal{M})\in\mathcal{Q}}{\arg\min}\Big\{F_k(s(\mathcal{M}))=\mathbb{E}_{q_{\phi_k}(\mathcal{M})}\left[J(\mathcal{X}_k)\right]+\alpha\mathrm{KL}\big(q_{\phi_k}(\mathcal{M})\parallel s(\mathcal{M})\big)\Big\}, \tag{3}$$

$$\text{Server: }\underset{s(\mathcal{M})}{\arg\min}\left\{\frac{1}{N}\sum_{k=1}^{N}F_k(s(\mathcal{M}))\right\}. \tag{4}$$

Theoretical Analysis: In this section, we present a theoretical discussion about the S-space. Moreover, we state the necessary Assumptions 1, 2, 3 and analyze equal-width BNNs as in [24, 25].

Definition 2.1. *(Optimal Variational Latent Space – OVLS) Consider all $\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{X}$ and a latent representation function $f_\Theta\colon\mathcal{X}\to\mathcal{Z}$. The transformation f_Θ generates an OVLS $\mathcal{Z}$ if (i) $\|\boldsymbol{s}_{ij}\|_2\sim\delta(0)\Rightarrow\ell(\boldsymbol{x}_i)=\ell(\boldsymbol{x}_j)$, and if (ii) $\ell(\boldsymbol{x}_i)\neq\ell(\boldsymbol{x}_j)\Rightarrow\mathbb{E}[\|\boldsymbol{s}_{ij}\|_2]>0$, where $\delta(.)$ is the Dirac delta function, and $\ell\colon\mathcal{X}\to\mathcal{Y}$ maps an unlabeled example $\boldsymbol{x}_i$ into its true label y_i.*

Definition 2.1 emphasizes the relationship between geometric proximity in the latent space $\mathcal{Z}$ and label identity. Unlike traditional deep metric learning methods [26, 27], our approach estimates a more flexible representation and captures more sophisticated data patterns (see Appendix A). Furthermore, a 1-nearest neighbor classifier in OVLS is optimal, emphasizing its capacity for perfect classification in theoretical applications.

Our algorithm induces a representation space as Definition 2.1. The loss function (Eq. 1) clusters samples by minimizing intra-group local distances ($\|\boldsymbol{s}_{ij}\|_2\sim\delta(0)$) and aligns a positive marker $\boldsymbol{m}^+\in\mathcal{M}^+$ near the origin ($\|\boldsymbol{m}^+\|_2\sim\delta(0)$), see Appendix D.2.

Corollary 1. *(based on Lemma C.1) Let $f_{i,\Theta}$ be a latent representation function that generates an OVLS, and $q^\star_{\phi_i}(\mathcal{M})$ denote the optimal variational distribution for the i-th user, estimated by a NN with weights Θ. If the markers are permuted based on the norm, i.e., the variational parameters $(\boldsymbol{\mu}^\star_{i,j},\boldsymbol{\sigma}^\star_{i,j})\sim q^\star_{\phi_i}(\mathcal{M})$ are organized into an ordered set according to the norm $\|\boldsymbol{\mu}^\star_{i,j}\|_2$, then the optimal server variational distribution $s^\star(\mathcal{M})$ admits the existence of an i such that $\|\boldsymbol{\mu}^{\star,s}_i\|_2\sim\delta(0)$.*

[1]The prior distribution is replaced with the global (server) distribution because the prior (for each client) is difficult to characterize in practice [13]. This approach avoids making assumptions about the prior distribution, leading to a better fit with the collected data. The global (server) distribution is also updated/recycled for each FL epoch [22].

Based on Corollary 1 (see Appendix C), we adapt the aggregation function proposed in [13] by ordering the set $\mathcal{M}$ using the norm $\|\cdot\|_2$ before transmitting the weights to the server. This permutation optimizes the aggregation function's performance, ensuring alignment with the criteria for OVLS.

Theorem 1. *If Assumptions 1, 2, 3 are true, then the following inequality holds*

$$KL\big(q^{\star}_{\phi_k}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})\big) \leq \frac{D-1}{D}(C' n r_n) < C'(n-1)r_n,$$

where D is the number of markers in $\mathcal{M}$, r_n and C' are constants.

Theorem 1 provides an upper limit for the KL divergence between the local (client) optimal variational solution $q^{\star}_{\phi_k}$ and the global (server) optimal variational solution $s^{\star}$. Our approach, thus, improves this FL theoretical results by using the variational S-space to estimate an optimal global VI distribution, with a tighter upper bound on divergence compared to traditional BNN approaches ($\mathrm{KL}(q^{\star}_{\phi_k}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})) \leq C' n r_n$), as documented in [13, 14]. We present the proof in Appendix C.2.

3 Experimentation

FL settings: We performed our experiments using the Flower framework [28] in a FL setting characterized by non-IID clients (quantity-based label imbalance) [29, 30]. For each dataset, we sorted the data by labels and divided it into N clients. Each client was assigned $\#S$ random non-overlapping subsets (shard), each containing an equal number of samples [29, 30].

Thus, we used a server and 100/200 clients in our experiment to evaluate our model, and we trained our method with two NVIDIA RTX 6000 Ada Generation ($48\,\mathrm{GB}$) for 1000 FL epochs. For each training round, the server selects $5\,\%$ of clients to train for five local epochs of the user model. We use the *F1-Score* commonly used in classification tasks, which can be directly computed from the confusion matrix.

Dataset description: We use five datasets for our evaluation: The MNIST/FMNIST datasets consist of 28x28 pixel grayscale images, with 60,000 training and 10,000 testing examples. The Malimg dataset comprises 9,339 samples from 25 malware families, with sample counts ranging from 80 to 2,949 per family. MaleVis features byte images of 25 malware types and one goodware class, totaling 14,226 images. CIFAR-10 includes 60,000 color images across ten different classes. Each malware binary code is visualized as a 64x64 grayscale image.

Results: We compared our proposal to six PFL techniques from the literature [3, 5, 13, 17, 31, 32], as well as the standard FedAvg (Global Model – GM) and Local model without client communication (baseline). The results, summarized in Table 1, show that our approach outperformed FedAvg across all five datasets. For example, on the CIFAR10 dataset, our method achieved an F1-score improvement of $27.43\,\%$ with four shards per client.

Our method also shows improvements across other datasets. For instance, on CIFAR10, our approach achieved the highest F1 scores of 0.7015 and 0.7202 with four and five shards per client, respectively. On the Malimg dataset, it scored 0.9324 with five shards per client, surpassing FedRep, which achieved the second-best result of 0.9250. However, on the MNIST dataset, our proposal achieved the third-best results for both shard values.

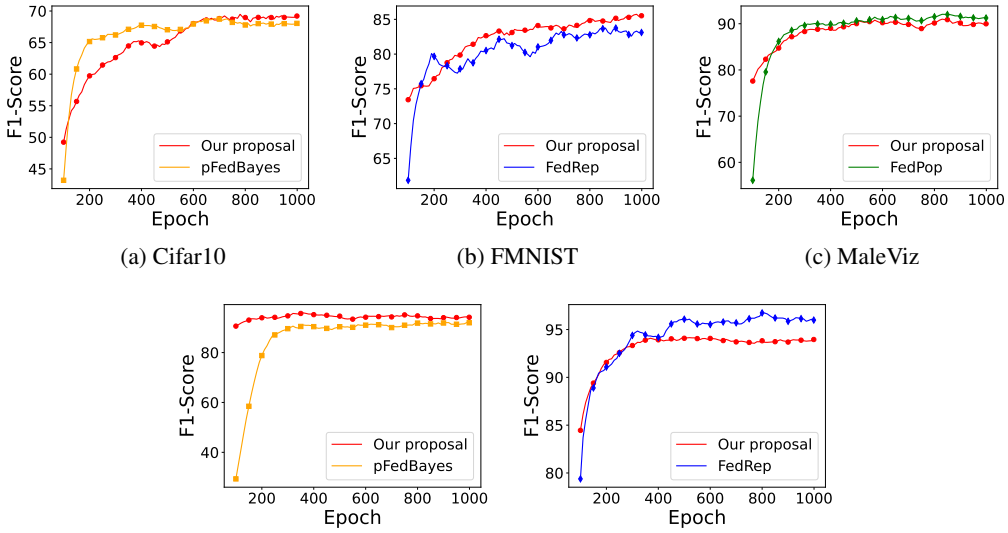
Figure 1 compares the F1-scores between our proposal and the best methods from the literature across five datasets with four shards per client over 1000 FL epochs. In summary, our approach achieved the best performance in six out of ten FL settings and the second-best in two additional settings. These findings are consistent with results obtained using 200 clients, as detailed in Table 2 (Appendix D.3).

4 Conclusion

In this work, we introduced the variational auxiliary similarity space within an FL environment, designed to enhance latent feature representation. The variational S-Space adapts to the unique data distributions of individual clients, mitigating the effects of data heterogeneity. Our evaluations, conducted across five datasets and ten different FL settings, demonstrate that our approach outperforms existing PFL methods. In addition, we have provided theoretical results showing that our approach achieves improved upper bounds compared to traditional distributed BNNs by applying

Table 1: F1-Score of various PFL approaches across five datasets with ***100*** clients. The best results for each dataset are highlighted in **bold**, and the second-best results are underlined.

Proposal	Datasets									
	MNIST		FMNIST		MaleViz		Malimg		CIFAR10	
	#S = 4	#S = 5	#S = 4	#S = 5	#S = 4	#S = 5	#S = 4	#S = 5	#S = 4	#S = 5
Local	0.9113	0.9025	0.8238	0.8047	0.8562	0.8254	0.8191	0.8086	0.4048	0.3768
FedAvg. (GM)	0.8275	0.8934	0.7026	0.7338	0.7009	0.7182	0.8497	0.8509	0.4224	0.4459
FedRep [31]	**0.9598**	0.9627	0.8351	**0.8616**	0.8439	0.8318	0.8874	0.9250	0.6891	0.7007
FedPer [3]	0.9553	**0.9664**	0.8296	0.8458	0.8988	0.9015	0.8974	0.9112	0.6599	0.6726
FedPop [17]	0.9180	0.9302	0.7734	0.8013	**0.9124**	0.9243	0.9033	0.9191	0.6211	0.6657
pFedSim [32]	0.8973	0.9406	0.7270	0.7670	0.9093	0.9135	0.8899	0.9017	0.6643	0.6975
pFedBayes [13]	0.8864	0.9143	0.8327	0.8452	0.8771	0.9025	0.9194	0.9236	0.6927	0.7098
DITTO [5]	0.9041	0.9392	0.8035	0.8243	0.8498	0.8826	0.8978	0.9128	0.6247	0.6536
Our proposal	0.9386	0.9525	**0.8476**	0.8564	0.9005	**0.9324**	**0.9275**	**0.9369**	**0.7015**	**0.7202**



(a) Cifar10 (b) FMNIST (c) MaleViz

(d) Malimg (e) MNIST

Figure 1: F1-Score results of our proposed method against the best approaches from the literature.

an aggregation function projected onto BNNs within an FL framework. A particularly promising direction is designing a specialized aggregation function optimized for the variational S-space.

Acknowledgment

This work was partially funded by São Paulo Research Foundation (grant #2023/00721-1), Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant #312682/2021-2), Fundação de apoio da UFMG – Fundep (grant # 29271) and Fundação de Amparo a Pesquisa do Estado de Minas Gerais (grant #APQ-00426-22). Additionally, travel support to the NeurIPS Workshop on Bayesian Decision-making and Uncertainty was generously provided by Google Research, with funds administered by Cornell University.

References

[1] B. McMahan *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7865–7873, May 2021.

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

[4] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 5132–5143.

[5] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 6357–6368.

[6] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3557–3568, 2020.

[7] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized federated learning with moreau envelopes," in *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2020.

[8] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," 2020.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *ACM Communication*, vol. 64, no. 3, p. 107–115, feb 2021.

[10] ——, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*, 2016.

[11] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig, "Laplace redux-effortless bayesian deep learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 089–20 103, 2021.

[12] A. Immer *et al.*, "Scalable marginal likelihood estimation for model selection in deep learning," in *International Conference on Machine Learning (ICML)*, vol. 139, 18–24 Jul 2021, pp. 4563–4573.

[13] X. Zhang, Y. Li, W. Li, K. Guo, and Y. Shao, "Personalized federated learning via variational bayesian inference," in *International Conference on Machine Learning (ICML)*, 2022, pp. 26 293–26 310.

[14] H. Chen, H. Liu, L. Cao, and T. Zhang, "Bayesian personalized federated learning with shared and personalized uncertainty representations," *arXiv preprint arXiv:2309.15499*, 2023.

[15] P. Barros, F. Queiroz, F. Figueiredo, J. A. D. Santos, and H. Ramos, "A new similarity space tailored for supervised deep metric learning," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, nov 2022.

[16] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research (JMLR)*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17] N. Kotelevskii, M. Vono, A. Durmus, and E. Moulines, "FedPop: A bayesian approach for personalised federated learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 8687–8701.

[18] J. Yao, Y. Yacoby, B. Coker, W. Pan, and F. Doshi-Velez, "An empirical analysis of the advantages of finite- vs. infinite-width bayesian neural networks," in *NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems*. NeurIPS, 2022. [Online]. Available: https://gp-seminar-series.github.io/neurips-2022/assets/camera_ready/58.pdf

[19] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538 – 557, 2012.

[20] Y. Wang, Y. Wang, J. Yang, and Z. Lin, "A unified contrastive energy-based model for understanding the generative ability of adversarial training," in *International Conference on Learning Representations*, 2021.

[21] B. Kim and J. C. Ye, "Energy-based contrastive learning of visual representations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4358–4369, 2022.

[22] T. G. J. Rudner, F. Bickford Smith, Q. Feng, Y. W. Teh, and Y. Gal, "Continual learning via sequential function-space variational inference," in *International Conference on Machine Learning (ICML)*, vol. 162, 17–23 Jul 2022, pp. 18 871–18 887.

[23] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.

[24] J. Bai, Q. Song, and G. Cheng, "Efficient variational inference for sparse deep learning with theoretical guarantee," *Advances in Neural Information Processing Systems*, vol. 33, pp. 466–476, 2020.

[25] N. G. Polson and V. Ročková, "Posterior concentration for sparse deep learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. USA: IEEE, 2006, pp. 1735–1742.

[27] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5022–5030.

[28] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[29] Z. He, Y. Li, D. Seo, and Z. Cai, "Fedcpd: Addressing label distribution skew in federated learning with class proxy decoupling and proxy regularization," *Information Fusion*, vol. 110, p. 102481, 2024.

[30] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, and C. Wu, "Federated learning with label distribution skew via logits calibration," in *International Conference on Machine Learning ICML*, vol. 162, 17–23 Jul 2022, pp. 26 311–26 329.

[31] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*. PMLR, 2021, pp. 2089–2099.

[32] J. Tan, Y. Zhou, G. Liu, J. H. Wang, and S. Yu, "pfedsim: Similarity-aware model aggregation towards personalized federated learning," *arXiv preprint arXiv:2305.15706*, 2023.

[33] J. Zhu, X. Ma, and M. B. Blaschko, "Confidence-aware personalized federated learning via variational expectation maximization," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 24 542–24 551.

[34] X. Liu, Y. Li, C. Wu, and C.-J. Hsieh, "Adv-BNN: Improved adversarial defense through robust bayesian neural network," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rk4Qso0cKm

[35] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning (ICML)*, vol. 32, no. 2, Bejing, China, 22–24 Jun 2014, pp. 1278–1286.

[36] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[37] 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.

[38] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *International Conference on Machine Learning*, 2015, p. 1613–1622.

A Revisiting Contrastive Loss

As introduced by [26], contrastive loss is a precursor approach for estimating latent representation based on pairs of items $(\boldsymbol{z}_i, \boldsymbol{z}_j)$, facilitating the discernment of their class relationships. Specifically, it is formulated as

$$L_c^{ij} = y_{ij}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 + (1 - y_{ij})\big[\max(0, \xi - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2)^2\big],$$

where $y_{ij} = \mathbb{1}[\ell'(\boldsymbol{z}_i) = \ell'(\boldsymbol{z}_j)]$ is the indicator function, i.e., $\mathbb{1}[\cdot] = 1$ if $[\cdot]$ is true and 0 otherwise and the function $\ell' \colon \mathcal{Z} \to \mathcal{Y}$, which maps the latent data $z_i = f_\Theta(x_i)$ into their respective labels. This method minimizes the loss function L_c^{ij} (also known as energy) by clustering similar points (same class) and separates points from different classes by at least a predefined margin ξ.

In contrast to this approach, we introduce the OVLS (Definition 2.1), a latent space representation projected for 1-nearest neighbor (1-NN) classification, achieving a classification accuracy of $100\,\%$. In an OVLS, similar pairs are collapsed, i.e., the distance between any two points within the same cluster is effectively zero ($\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2 = 0$), while ensuring that clusters of different classes maintain a minimum separation distance of $\xi > 0$.

Figure 2 is a toy representation of the OVLS, highlighting the distinct clustering of $4k$ points (k points per clustering) from two hypothetical classes, represented by red and green colors, with $k = 5$. The expected distance between pairs of elements $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ from different classes ($y_{ij} = 0$) is $\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2] = \xi$, leading to a contrastive loss of $L_c^{ij} = 0$ because $\max(0, \xi - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2)^2 = 0$. However, the expected distance between elements from the same class is $\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2] = \xi\sqrt{2}/2$ because we have two different clustering for each class, resulting in a non-zero contrastive loss that could potentially disrupt the latent space.

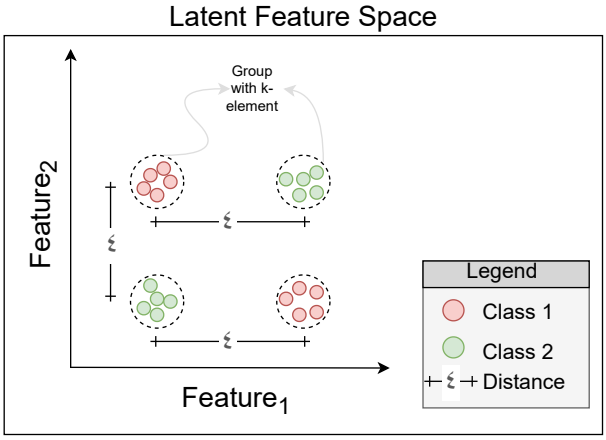


Figure 2: Illustration of the 1-NN optimal latent space for binary classification, derived from the encoder: distinct clusters for two classes (red and green points). Each class consists of 10 points. All points in the same cluster are collapsed to a single position (based [15]).

Our method advances the conventional contrastive loss framework by preserving the structural integrity of the OVLS. For the settings in Figure 2, if we introduce four 2D markers (in S-space) at positions $\mathcal{M}^+ = \{(0,0), (\xi,\xi)\}$ and $\mathcal{M}^- = \{(\xi,0), (0,\xi)\}$, we achieve a loss function equal to zero. This approach maintains the structured separation between classes within the OVLS, offering an alternative to the traditional contrastive loss model.

B Experimental Details

B.1 Parameters initialization and network architecture

The NN architecture used for all FL approaches, including our proposed method, is identical. Following [13, 33], the network dimensions are m-100-n for the MNIST and FMNIST datasets, where m represents the number of input features and d denotes the latent space representation

dimension. Additionally, for other datasets, we utilized a LeNet-5 architecture with the same latent space dimension ($d = 64$). We employed SGD with a learning rate of 0.01 for all experiments. All baseline models were configured using the hyperparameters recommended in their respective original publications.

B.2 Implementation details

This section describes the practical implementation of our proposed FL framework, emphasizing the optimization of similarity markers. Each client estimates a market set $\mathcal{M}$, where $\mathcal{M}^+$ represents similarity markers and $\mathcal{M}^-$ represents dissimilarity markers.

We hypothesize that the marker values for each client in the S-Space follow a Gaussian distribution. Therefore, the joint probability density function $q_\phi(\mathcal{M})$ is modeled as a product of normal densities:

$$q_\phi(\mathcal{M}) = \prod_{m_k \in \mathcal{M}} \mathcal{N}\big(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k = \text{Diag}(\boldsymbol{\sigma}_k^2)\big),$$

where $\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k \in \mathbb{R}^d$, Diag(.) denotes the diagonal matrix function, and mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ are the variational parameters. In addition, $\boldsymbol{L}$ is a diagonal matrix representing a mean-field parameterization. The variational parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ are initialized by sampling from the uniform distribution $\boldsymbol{\mu}_{prior} \sim \mathcal{U}(-d^{-1/2}, d^{-1/2})$, where d is the latent feature dimension, and the constant $\sigma_{\text{prior}} = 0.05$, as referenced in [34].

Furthermore, the variance parameters $\boldsymbol{\sigma}_k$ are reparameterized as $\boldsymbol{\sigma}_k = \exp(\boldsymbol{p}_k)$ to enable the application of gradient-based optimization techniques directly, resolving issues with the non-negativity constraint on standard deviations [35].

For sampling marker instances $\boldsymbol{m}_i$, we follow the methodology in [36], introducing a noise component $\boldsymbol{\epsilon} \in \mathbb{R}^d$ sampled from a standard normal distribution $\mathcal{N}(0, 1)$: we sample $\boldsymbol{m}_i \sim q_{\phi_k}(\mathcal{M})$ as

$$\boldsymbol{m}_i = \boldsymbol{\mu}_k + \exp(\boldsymbol{p}_k) \odot \boldsymbol{\epsilon},$$

where $\odot$ denotes element-wise multiplication. This formulation makes m_i differentiable, enabling the gradient backpropagation through the randomness introduced by $\boldsymbol{\epsilon}$.

We employ Monte Carlo sampling [37, 36] to approximate the objective function for client k (Eq. 3):

$$D_k^B = -\frac{n_k}{n_b}\frac{1}{K}\sum_{j=1}^{b}\sum_{i=1}^{K}\big[J(B_j) + \alpha \text{KL}(q_{\phi_k}(\mathcal{M}) \parallel s(\mathcal{M}))\big],$$

where $B \subset \mathcal{X}_k$ represents a minibatch of size n_b, n_k denotes the total number of data points in dataset $\mathcal{X}_k$, and $K = 10$ is the number of samples used in the Monte Carlo estimation [17, 38].

Finally, via the backpropagation algorithm, we update the variational model parameters using minibatch gradient descent, denoted by ΔD_k^B.

C Theoretical Analysis

C.1 External results

Assumption 1. *(Ref. [13, 24]) The activation function is 1-Lipschitz continuous.*

Assumption 2. *(Ref. [13, 24]) Consider a NN with T parameters, I hidden layers, n samples in the FL environment, an input dimension d, and M neurons per hidden layer. The parameters d, n, M, I are large enough such that*

$$\sigma_n^2 = \frac{T}{8n}A \leq B^2, \tag{5}$$

where $H = BM$ and

$$A = \log^{-1}(3dM)(2H)^{-2(I+1)}\left[\left(d + 1 + \frac{1}{H-1}\right)^2 + \frac{1}{(2H)^2 - 1} + \frac{2}{(2H-1)^2}\right]^{-1}, \tag{6}$$

As discussed in [13, 24], the parameter σ_n is constructed to facilitate the proof of Theorem 1, particularly in inequality 12. Given that the neural network parameters are bounded by B, their variance should be upper bounded by B^2.

Lemma C.1. *(Ref. [13]) Let $s^\star(\mathcal{M})$ be the optimal server variational distribution based on the following FL optimization problem*

$$s^\star(\mathcal{M}) = \arg\min_{s(\mathcal{M})} \frac{1}{N} \sum_{i=1}^{N} KL\big[q^\star_{\phi_i}(\mathcal{M}) \parallel s(\mathcal{M})\big],$$

where $q^\star_{\phi_i}(\mathcal{M})$ is the local optimal variational model for user u_i, the server variational distribution parameters of $s^\star(\mathcal{M})$ for marker $\boldsymbol{m}_j$ are $(\boldsymbol{\mu}_j^{\star,s}, \boldsymbol{\sigma}_j^{\star,s})$. We denote $\mu_j^{\star,s}|_n \in \mathbb{R}$ and $\sigma_j^{\star,s}|_n \in \mathbb{R}$ as the n-th components of the vectors $\boldsymbol{\mu}_j^{\star,s} \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_j^{\star,s} \in \mathbb{R}^d$, respectively.

Therefore, we have

$$\mu_a^{\star,s}|_n = \frac{1}{N} \sum_{i=1}^{N} \mu_{i,a}|_n,$$

and

$$(\sigma_a^{\star,s}|_n)^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(\sigma_{i,a}|_n)^2 + (\mu_{i,a}|_n)^2 - (\mu_a^{\star,s}|_n)^2\right], \tag{7}$$

where the variational distribution parameters of $q^\star_{\phi_i}(\mathcal{M})$ for the j-th marker are $(\boldsymbol{\mu}^\star_{i,j}, \boldsymbol{\sigma}^\star_{i,j})$. Futhermore, $\mu_{i,j}|_a \in \mathbb{R}$ and $\sigma_{i,j}|_a \in \mathbb{R}$ as the a-th components of the vectors $\boldsymbol{\mu}^\star_{i,j} \in \mathbb{R}^d$ and $\boldsymbol{\sigma}^\star_{i,j} \in \mathbb{R}^d$, respectively.

C.2 Theoretical results

Assumption 3. *Let the number of positive markers be $\#\mathcal{M}^+$, and the number of negative markers be $\#\mathcal{M}^-$ in the S-space. We assume $\#\mathcal{M}^+ = \#\mathcal{M}^-$ and $\#\mathcal{M}^+ < n/2$, where n represents the number of samples in the FL environment.*

Corollary 1. *Let $f_{i,\Theta}$ be a latent representation function that generates an OVLS, and $q^\star_{\phi_i}(\mathcal{M})$ denote the optimal variational distribution for the i-th user, estimated by a NN with weights Θ. If the markers are permuted based on the norm, i.e., the variational parameters $(\boldsymbol{\mu}^\star_{i,j}, \boldsymbol{\sigma}^\star_{i,j}) \sim q^\star_{\phi_i}(\mathcal{M})$ are organized into an ordered set according to the norm $\|\boldsymbol{\mu}^\star_{i,j}\|_2$, then the optimal server variational distribution $s^\star(\mathcal{M})$ admits the existence of an i such that $\|\boldsymbol{\mu}_i^{\star,s}\|_2 \sim \delta(0)$.*

Proof. Consider a FL system comprising N clients, each utilizing an OVLS solution (see Definition 2.1). These parameters are organized into an ordered set based on the norm $\|\boldsymbol{\mu}^\star_{i,j}\|_2$. Consequently, for expected positive markers with indices $1 \leq j \leq \#\mathcal{M}^+ - 1$, we have $\|\boldsymbol{\mu}^\star_{i,j}\|_2 \leq \|\boldsymbol{\mu}^\star_{i,j+1}\|_2$ for all clients in the FL environment.

From Definition 2.1 and the ordering, we conclude that for all client i, $(\mu^\star_{i,1}, \sigma^\star_{i,1}) \sim \delta(0)$. According to Lemma C.1, the server aggregated model's mean $\boldsymbol{\mu}_1^{\star,s}$ and variance $\boldsymbol{\sigma}_1^{\star,s}$ are given by

$$\mu_1^{\star,s}|_a = \frac{1}{N} \sum_{i=1}^{N} \mu^\star_{i,1}|_a, \text{ and } (\sigma_1^{\star,s}|_a)^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(\sigma_{i,1}|_a)^2 + (\mu_{i,1}|_a)^2 - (\mu_1^{\star,s}|_a)^2\right].$$

Therefore, we conclude than $(\mu^\star_{s,1}, \sigma^\star_{s,1}) \sim \delta(0)$. □

Theorem 1. *Suppose that the assumptions are true, then the following inequality holds*

$$KL\big(q^\star_{\phi_i}(\mathcal{M}) \parallel s^\star(\mathcal{M})\big) \leq \frac{D-1}{D}(C' n r_n) < C'(n-1) r_n,$$

where $D = (\#\mathcal{M}^+) + (\#\mathcal{M}^-)$ is the number of marker in $\mathcal{M}$ and C' and r_n are constants.

Proof. Considering the mean-field decomposition for $\boldsymbol{\mu}^{\star}_{i,m}, \boldsymbol{\mu}^{\star,s}_{m} \in \mathbb{R}^d$, we have

$$q^{\star}_{\phi_i}(\mathcal{M}) = \prod_{m=1}^{D} \mathcal{N}(\mu^{\star}_{i,m}, (\sigma^{\star}_{n})^2), \text{ and } s^{\star}(\mathcal{M}) = \prod_{m=1}^{D} \mathcal{N}(\mu^{\star,s}_{m}, (\sigma^{\star,s}_{m})^2)\cdot$$

Thus, the KL divergence between $q^{\star}_{\phi_i}(\mathcal{M})$ and $s^{\star}(\mathcal{M})$ can be decomposed as

$$\begin{aligned}\mathrm{KL}(q^{\star}_{\phi_i}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})) &= \mathrm{KL}\Big(\prod_{m=1}^{D} \mathcal{N}(\mu^{\star}_{i,m}, (\sigma^{\star}_{n})^2) \Big\| \prod_{m=1}^{D} \mathcal{N}(\mu^{\star,s}_{m}, (\sigma^{\star,s}_{m})^2) \Big) \\ &= \sum_{m=1}^{D} \mathrm{KL}\left(\mathcal{N}(\mu^{\star}_{i,m}, (\sigma^{\star}_{n})^2) \parallel \mathcal{N}(\mu^{\star,s}_{m}, (\sigma^{\star,s}_{m})^2) \right).\end{aligned}$$

For each marker m_m, the KL divergence between two Gaussian distributions is

$$\begin{aligned}&\mathrm{KL}\left(\mathcal{N}(\mu^{\star}_{i,m}, (\sigma^{\star}_{i,m})^2) \parallel \mathcal{N}(\mu^{\star,s}_{m}, (\sigma^{\star,s}_{m})^2) \right) = \\ &\qquad \frac{1}{2} \sum_{a=1}^{d} \left[\log \left(\frac{(\sigma^{\star,s}_{m}|_a)^2}{(\sigma^{\star}_{i,m}|_a)^2} \right) + \frac{(\sigma^{\star}_{i,m}|_a)^2 + (\mu^{\star}_{i,m}|_a - \mu^{\star,s}_{m}|_a)^2}{(\sigma^{\star,s}_{m}|_a)^2} - 1 \right]. \quad (8)\end{aligned}$$

By Corollary 1, we know that for variational parameters$(\mu^{\star}_{i,1}, \sigma^{\star}_{i,1}) \sim \delta(0)$ and $(\mu^{\star,s}_{1}, \sigma^{\star,s}_{1}) \sim \delta(0)$, i.e., for any client i with optimal variational latent space, the variation parameters for the first marker m_1 (marker with lowest $\| \cdot \|_2$ norm) is equal to the optimal server defined by optimization problem in Lemma C.1. Therefore, the KL divergence between those lowest norm markers resulted in zero (equal distribution). For the remaining markers $m \geq 2$, we have

$$\begin{aligned}\mathrm{KL}(q^{\star}_{\phi_i} \parallel s^{\star}) &= \sum_{m=2}^{D} \mathrm{KL}\left(\mathcal{N}(\mu^{\star}_{i,m}, (\sigma^{\star}_{n})^2) \parallel \mathcal{N}(\mu^{\star,s}_{m}, (\sigma^{\star,s}_{m})^2) \right) \\ &= \frac{1}{2} \sum_{a=1}^{d} \sum_{m=2}^{D} \left[\log \left(\frac{(\sigma^{\star,s}_{m})^2}{(\sigma^{\star}_{n})^2} \right) + \frac{(\sigma^{\star}_{n})^2 + (\mu^{\star}_{i,m} - \mu^{\star,s}_{m})^2}{(\sigma^{\star,s}_{m})^2} - 1 \right] \\ &= \frac{1}{2} \sum_{a=1}^{d} \sum_{m=2}^{D} \left[\log \left(\frac{(\sigma^{\star,s}_{m})^2}{(\sigma^{\star}_{n})^2} \right) \right] \quad (9) \\ &\leq \frac{1}{2} \sum_{a=1}^{d} \sum_{m=2}^{D} \left[\log \left(\frac{(\sigma^{\star}_{n})^2 + B^2}{(\sigma^{\star}_{n})^2} \right) \right], \quad (10)\end{aligned}$$

where we applied in Eq. (9) the bellow equality (based Lemma C.1 – Eq.A.14 in [13]) ensuring that

$$\frac{(\sigma^{\star}_{n})^2 + (\mu^{\star}_{i,m} - \mu^{\star,s}_{m})^2}{(\sigma^{\star,s}_{m})^2} = 1,$$

and the inequality applies Assumption 2 and Eq. (7) (as can see in) that

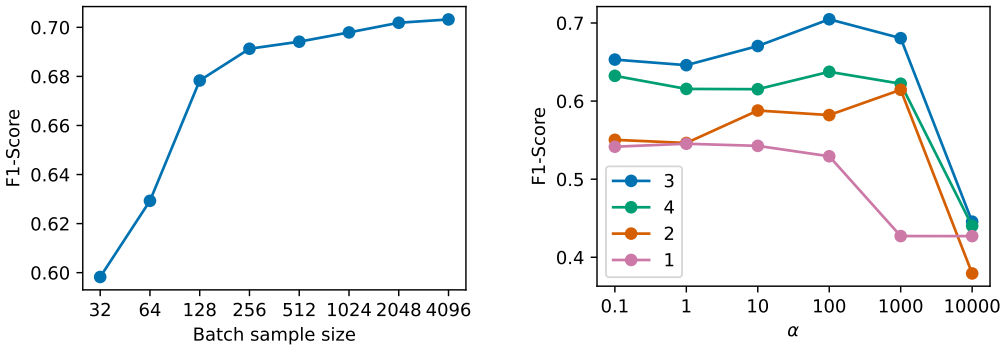
$$(\sigma^{\star,s}_{m})^2 = (\sigma^{*}_{n})^2 - (\mu^{\star}_{s,m})^2 + \frac{1}{N} \sum_{i=1}^{N} (\mu^{\star}_{i,m})^2 \leq (\sigma^{*}_{n})^2 + B^2.$$

By bounding the variance term using Assumption 2, we obtain

$$\mathrm{KL}(q^{\star}_{\phi_i}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})) \leq \frac{d(D-1)}{2} \log \left(\frac{2B^2}{(\sigma^{\star}_{n})^2} \right). \quad (11)$$

By Assumption 2, incorporating into Eq. (10) and following similar steps in Eq. A.19 from [13], we get

$$\log \left(\frac{2B^2}{(\sigma^{\star}_{n})^2} \right) \leq \frac{2}{dD} (C' n r_n), \quad (12)$$



(a) The effect of the number of pairs used in local training.

(b) The impact of the α parameter for different numbers of (dis)similar markers.

Figure 3: Performance of our proposed method across different hyperparameter settings on CIFAR10 datasets on the F1-Score.

which, combined with Eqs. (11) and (12), results in

$$\mathrm{KL}(q^{\star}_{\phi_i}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})) \leq \frac{d(D-1)}{2} \log\left(\frac{2B^2}{(\sigma^{\star}_n)^2}\right) \leq \frac{D-1}{D}(C' n r_n).$$

Therefore, by Assumption 3, we get

$$\frac{D-1}{D} < \frac{n-1}{n},$$

and

$$\mathrm{KL}\big(q^{\star}_{\phi_i}(\mathcal{M}) \parallel s^{\star}(\mathcal{M})\big) < C'(n-1) r_n.$$

□

D Additional experiments

D.1 Experimental Hyperparameter Settings

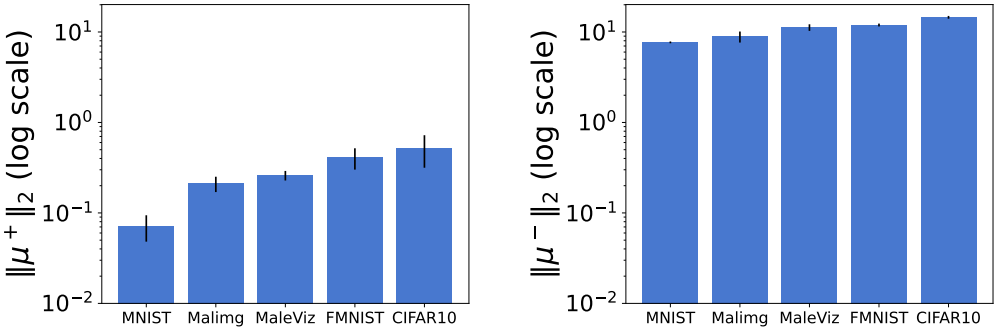
In this experiment, we evaluate the impact of various hyperparameter settings on the performance of our approach. The results of these experiments are summarized in Figure 3 with the CIFAR10 dataset.

Figure 3a shows the impact of the number of pairs used in the local training. As we increase the number of pairs, we gain more confidence about the pair distribution, improving model performance. However, this also increases the training time. Although the performance suggests that using a larger number of pairs can improve model performance, the F1-Score exhibits marginal improvements beyond 2048 pairs. For this short paper, we adopted 2048 pairs per epoch in the training step to balance performance and training efficiency.

Figure 3b presents the effect of the α parameter, which controls the influence of KL divergence regularization between the local and global variational parameters (Eq. (3)). We also evaluate our approach with a different number of similar/dissimilar markers. As expected, our experiment indicates that the F1-Score is sensitive to α. For lower values of alpha ($0.1 \leq \alpha \leq 10$), our model is stable. Furthermore, the number of markers affects our model's results, with more markers capturing the similarity structure more effectively. Therefore, our model achieves optimal performance with $\alpha = 100$ and three similar/dissimilar markers (six markers in total).

D.2 Markers analysis

This section analyzes the norms of the positive and negative expected markers, denoted as $\|\mu^{+}\|_2$ and $\|\mu^{-}\|_2$, respectively, across five datasets. Figure 4 shows the positive expected markers norm



(a) Positive Markers Norm.

(b) Negative Markers Norm

Figure 4: Comparison of Positive and Negative Marker Norms across five datasets.

$\|\mu^+\|_2$ (Figure 4a) and the negative expected markers norm $\|\mu^-\|_2$ (Figure 4b). For simplicity, each dataset utilizes one positive marker and one negative marker.

Definition 2.1 highlights a key feature of our algorithm: positive expected markers $\|\mu^+\|_2$ have a smaller norm than their negative marker $\|\mu^-\|_2$. Our loss function, aimed at clustering samples into groups with similar characteristics ($\|s_{ij}\|_2 \sim \delta(0)$), induces the model to align a positive marker $\mu^+ \in \mathcal{M}^+$ to the origin ($\|\mu^+\|_2 \sim \delta(0)$), as confirmed by our observations in Figure 4a. For example, in the MNIST dataset, we observe $\|\mu^+\|_2 = 0.071$ in contrast to $\|\mu^-\|_2 = 7.68$ with an expected ratio $(\|\mu^+\|_2/\|\mu^-\|_2)$ equal to $0.92\,\%$. We observed similar behavior for the Malimg and Maleviz datasets, with expected ratios of $2.37\,\%$ and $2.54\,\%$, respectively.

Additionally, we found a relation between positive expected markers norm $\|\mu^+\|_2$ and F1-Score. As seen in Table 1, a lower $\|\mu^+\|_2$ value correlates with a higher F1-Score. For instance, our model achieves F1-Scores of 0.9525, 0.8564, and 0.7202 corresponding to $\|\mu^+\|_2$ values of 0.071, 0.401, and 0.519 for the MNIST, FMNIST, and CIFAR10 datasets respectively. Therefore, this experiment found evidence that the positive norm position can be a proxy for the model's performance.

D.3 Quantitative results

We present additional experiments as discussed in Section 3. Specifically, we conducted an experiment with 200 clients to evaluate the impact of our proposed framework, as shown in Table 2. We compare the F1-Scores of various PFL approaches across five datasets described in Section 3. Similar to Table 1, our proposed method outperforms the other approaches in six FL settings (out of ten).

Table 2: F1-Score of various PFL approaches across five datasets with ***200*** clients. The best results for each dataset are highlighted in **bold**, and the second-best results are underlined.

Proposal	Datasets									
	MNIST		FMNIST		MaleViz		Malimg		CIFAR10	
	$\#S=4$	$\#S=5$	$\#S=4$	$\#S=5$	$\#S=4$	$\#S=5$	$\#S=4$	$\#S=5$	$\#S=4$	$\#S=5$
Local	0.8983	0.8821	0.7911	0.7803	0.8154	0.7714	0.8050	0.7807	0.3840	0.3615
FedAvg. (GM)	0.8131	0.8425	0.6951	0.7166	0.6682	0.6790	0.8069	0.8307	0.4275	0.4393
FedRep	0.9456	0.9601	0.8090	0.8271	0.7918	0.8112	0.8507	0.8638	0.5588	0.6213
FedPer	**0.9461**	**0.9692**	0.8039	0.8197	0.8315	0.8686	0.8816	0.9181	0.6301	0.6601
FedPop	0.9072	0.9323	0.7593	0.7824	**0.8624**	0.8733	0.8677	0.8993	0.5839	0.6420
pFedSim	0.8901	0.9115	0.7168	0.7492	0.8333	0.8537	0.8575	0.8947	0.6290	0.6437
pFedBayes	0.8715	0.9092	0.8195	0.8385	0.8253	0.8496	**0.9036**	0.9121	0.6496	0.6678
DITTO	0.8893	0.9288	0.7941	0.8065	0.7926	0.8341	0.8689	0.8767	0.6133	0.6366
Our proposal	0.9254	0.9448	**0.8247**	**0.8513**	0.8570	**0.8777**	0.9013	**0.9206**	**0.6561**	**0.6824**