



HAL
open science

Personalised Federated Learning On Heterogeneous Feature Spaces

Alain Rakotomamonjy, Maxime Vono, Hamlet Jesse Medina Ruiz, Liva Ralaivola

► **To cite this version:**

Alain Rakotomamonjy, Maxime Vono, Hamlet Jesse Medina Ruiz, Liva Ralaivola. Personalised Federated Learning On Heterogeneous Feature Spaces. 2023. hal-03962195

HAL Id: hal-03962195

<https://hal.science/hal-03962195>

Preprint submitted on 30 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Personalised Federated Learning On Heterogeneous Feature Spaces

Alain Rakotomamonjy^{*1} Maxime Vono^{*1} Hamlet Jesse Medina Ruiz¹ Liva Ralaivola¹

Abstract

Most personalised federated learning (FL) approaches assume that raw data of all clients are defined in a common subspace *i.e.* all clients store their data according to the same schema. For real-world applications, this assumption is restrictive as clients, having their own systems to collect and then store data, may use *heterogeneous* data representations. We aim at filling this gap. To this end, we propose a general framework coined **FLIC** that maps client’s data onto a common feature space via local embedding functions. The common feature space is learnt in a federated manner using Wasserstein barycenters while the local embedding functions are trained on each client via distribution alignment. We integrate this distribution alignment mechanism into a federated learning approach and provide the algorithmics of **FLIC**. We compare its performances against FL benchmarks involving heterogeneous input features spaces. In addition, we provide theoretical insights supporting the relevance of our methodology.

1. Introduction

Federated learning (FL) is a machine learning paradigm where models are trained from multiple isolated data sets owned by individual agents (coined *clients*), without requiring to move raw data into a central server, nor even share them in any way (Kairouz et al., 2021). This framework has lately gained a strong traction from both industry and academic research. Indeed, it avoids the communication costs entailed by data transfer, allows all clients to benefit from participating to the learning cohort, and finally, it fulfills first-order confidentiality guarantees, which can be further enhanced by resorting to so-called *privacy-enhancing technologies* such as differential privacy (Dwork and Roth, 2014) or secure multi-party com-

putation (Bonawitz et al., 2017). As core properties, FL ensures data ownership, and structurally incorporates the principle of data exchange minimisation by only transmitting the required updates of the models being learnt. Depending on the data partitioning and target applications, numerous FL approaches have been proposed, such as horizontal FL (McMahan et al., 2017) and vertical FL (Hardy et al., 2017; Yang et al., 2019). The latter paradigm considers that clients hold disjoint subsets of features corresponding to the same users while the former assumes that clients have data samples from different users. Recently, horizontal FL works have focused on *personalised* FL to tackle statistical heterogeneity by using local models to fit client-specific data (Hanzely and Richtárik, 2020; Jiang et al., 2019; Khodak et al., 2019; Tan et al., 2022).

Existing horizontal personalised FL works assume that the raw data on *all* clients share the same structure and are defined in a common feature space. Yet, in practice, data collected by clients may use differing structures. For instance, clients may not collect exactly the same information, some features may be missing or not stored, or some might have been transformed (*e.g.* via normalisation, scaling, or linear combinations). To address the key issue of implementing FL when the clients’ feature spaces are heterogeneous, in the sense that they have different dimensionalities or that the semantics of given vector coordinates are different, we introduce *the first* — to the best of our knowledge — personalised FL framework dedicated to this learning situation.

Proposed Approach. The framework and algorithm described in this paper rest on the idea that before performing efficient FL training, a key step is to map the raw data into a common subspace. This is a prior necessary step before FL since it allows to define a relevant aggregation scheme on the central server for model parameters (*e.g.* via weighted averaging) as the latter become comparable. Thus, we map clients’ raw data into a common low-dimensional latent space, via local and learnable feature embedding functions.

In order to ease subsequent learning steps, data related to the same semantic information (*e.g.* label) have to be embedded in the same region of the latent space. To ensure this property, we align clients’ embedded feature distributions via a latent *anchor distribution* that is shared across clients. The learning of this *anchor distribution* is per-

^{*}Equal contribution ¹Criteo AI Lab, Paris, France. Correspondence to: Maxime Vono <m.vono@criteo.com>, Alain Rakotomamonjy <a.rakotomamonjy@criteo.com>.

formed in a federated manner *i.e.* by updating it locally on each client before aggregation on the central server. More precisely, each client updates her local version of the anchor distribution by aligning it, *i.e.* making it closer, to the embedded feature distribution. Then, the central server aims at finding the mean element, *i.e.* barycenter, of these local anchor distributions (Banerjee et al., 2005; Veldhuis, 2002). Once this distribution alignment mechanism (based on local embedding functions and anchor distribution) is defined, it can be seamlessly integrated into a personalised FL framework; the personalisation part aiming at tackling residual statistical heterogeneity. In this paper, without loss of generality, we have embedded this alignment framework into a personalised FL approach similar to the one proposed in Collins et al. (2021).

Related Ideas. Ideas that we have built on for solving the task of FL from heterogeneous feature spaces have been partially explored in related literature. From the theoretical standpoint, works on the Gromov-Wasserstein distance or variants seek at comparing distributions from incomparable spaces in a (non-FL) centralised manner (Alaya et al., 2022; Bunne et al., 2019; Mémoli, 2011). Other methodological works on autoencoders (Xu et al., 2020), word embeddings (Alvarez-Melis and Jaakkola, 2018; Alvarez-Melis et al., 2019) or FL under high statistical heterogeneity (Luo et al., 2021; Makhija et al., 2022; Zhou et al., 2022) use similar ideas of distribution alignment for calibrating feature extractors and classifiers. A detailed literature review and comparison with the proposed methodology is postponed to Section 2.

Contributions. In order to help the reader better grasp the differences of our approach with respect to the existing literature, we spell out our contributions:

1. We are *the first* to formalise the problem of personalised horizontal FL on heterogeneous clients' feature spaces. In contrast to existing approaches, the proposed general framework, coined **FLIC**, allows each client to leverage other clients' data even though they do not have the same raw representation.
2. We introduce a distribution alignment framework and an algorithm that learns the feature embedding functions along with the latent anchor distribution in a local and global federated manner, respectively. We also show how those essential algorithmic pieces are integrated into a personalised FL algorithm, easing adoption by practitioners.
3. We provide algorithmic and theoretical support to the proposed methodology. In particular, we show that for an insightful simpler learning scenario, **FLIC** is able to recover the true latent subspace underlying the FL problem.

4. Experimental analyses on toy data sets and real-world problems illustrate the accuracy of our theory and show that **FLIC** provides better performance than competing FL approaches.

Conventions and Notations. The Euclidean norm on \mathbb{R}^d is $\|\cdot\|$, we use $|S|$ to denote the cardinality of the set S and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^*$, we refer to $\{1, \dots, n\}$ with the notation $[n]$. We denote by $\mathcal{N}(m, \Sigma)$ the Gaussian distribution with mean vector m and covariance matrix Σ and use the notation $X \sim \nu$ to denote that the random variable X has been drawn from the probability distribution ν . We define the Wasserstein distance of order 2 for any probability measures μ, ν on \mathbb{R}^d with finite 2-moment by $W_2(\mu, \nu) = (\inf_{\zeta \in \mathcal{T}(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \theta'\|^2 d\zeta(\theta, \theta'))^{1/2}$, where $\mathcal{T}(\mu, \nu)$ is the set of transference plans of μ and ν .

2. Proposed Methodology

Problem Formulation. We are considering a centralised and horizontal FL framework involving $b \in \mathbb{N}^*$ clients and a central entity (Kairouz et al., 2021; Yang et al., 2019). Under this paradigm, the central entity orchestrates the collaborative solving of a common machine learning problem by the b clients; without requiring raw data exchanges. For the sake of simplicity, we consider the setting where all clients want to solve a multi-class classification task with $C \in \mathbb{N}^*$ classes. In Appendix, we also highlight how regression tasks could be encompassed in the proposed framework. The b clients are assumed to possess local data sets $\{D_i\}_{i \in [b]}$ such that, for any $i \in [b]$, $D_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j \in [n_i]}$ where $x_i^{(j)}$ stands for a feature vector, $y_i^{(j)}$ is a label and $n_i = |D_i|$. A core assumption of FL is that the local data sets $\{D_i\}_{i \in [b]}$ are *statistically heterogeneous i.e.* for any $i \in [b]$ and $j \in [n_i]$, $(x_i^{(j)}, y_i^{(j)}) \stackrel{\text{i.i.d.}}{\sim} \nu_i$ where ν_i is a *local* probability measure defined on an appropriate measurable space. Existing horizontal FL approaches typically assume that the raw input features of the clients are defined on a common subspace so that their marginal distributions admit the same support.

In contrast, **we suppose here that these features live in heterogeneous spaces.** Our main goal is to cope with this new type of heterogeneity in horizontal FL. More precisely, for any $i \in [b]$ and $j \in [n_i]$, we assume that $x_i^{(j)} \in \mathcal{X}_i \subseteq \mathbb{R}^{k_i}$ such that $\{\mathcal{X}_i\}_{i \in [b]}$ are not part of a common ground metric. This setting is challenging since standard FL approaches (Li et al., 2020; McMahan et al., 2017) and even personalised FL ones (Collins et al., 2021; Hanzely et al., 2021) cannot be applied directly. In addition, we also assume a specific type of *prior probability shift* where, for any $i \in [b]$ and $j \in [n_i]$, $y_i^{(j)} \in \mathcal{Y}_i \subseteq [C]$. For instance, a client might only see digits 1 and 7 from the

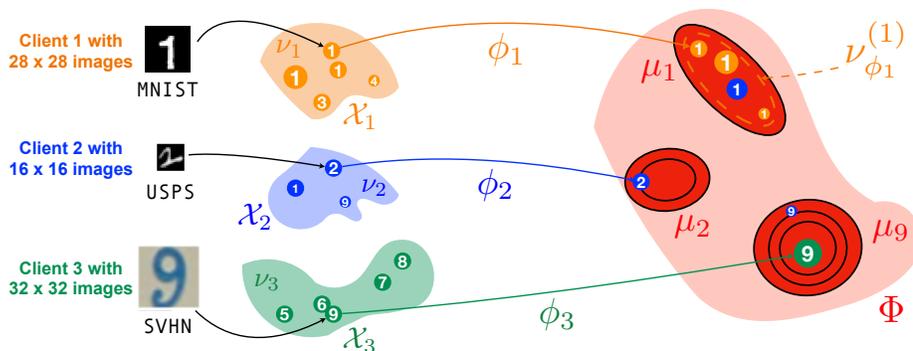


Figure 1. Illustration of part of the proposed methodology for $b = 3$ clients with *heterogeneous* digit images coming from three different data sets namely MNIST (Deng, 2012), USPS (Hull, 1994) and SVHN (Netzer et al., 2011). The circles with digits inside stand for a group of samples, of a given class, owned by a client and the size of the circles indicates their probability mass. In the subspace Φ , $\{\mu_i\}_{i \in [b]}$ (and their level sets) refer to some learnable reference measures to which we seek to align the transformed version ν_{ϕ_i} of ν_i . Personalised FL then occurs in the space Φ and aims at learning local models $\{\theta_i\}_{i \in [b]}$ for each clients as well as $\{\phi_i, \mu_i\}_{i \in [b]}$.

MNIST data set while another one only has access to USPS digits 1, 2 and 9, see Figure 1.

Methodology. To address the feature space heterogeneity issue, we propose to map clients’ features into a fixed-dimension common subspace $\Phi \subseteq \mathbb{R}^k$ by resorting to *local* embedding functions $\{\phi_i : \mathcal{X}_i \rightarrow \Phi\}_{i \in [b]}$ ¹. Our proposal for learning those local functions is illustrated in Figure 1. In order to preserve some semantical information (such as the class associated to a feature vector) on the original data distribution, we seek at learning the functions $\{\phi_i\}_{i \in [b]}$ such that they are aligned with (*i.e.* close to) some learnable latent anchor distribution that is shared across all clients. This anchor distribution must be seen as a universal “calibrator” for clients that avoids similar semantical information from different client being scattered across the subspace Φ , impeding then a proper subsequent federated learning procedure of the classification model. As depicted in Figure 1, we propose to learn the feature embedding functions by aligning their probability distributions conditioned on the class $c \in [C]$, denoted by $\nu_{\phi_i}^{(c)}$, via C learnable anchor measures $\{\mu_c\}_{c \in [C]}$ (Kollias et al., 2021; Tschannen et al., 2020; Xu et al., 2020; Zhou et al., 2021). In the literature, several approaches have been considered to align probability distributions ranging from mutual information maximisation (Tschannen et al., 2020), maximum mean discrepancy (Zellinger et al., 2017) to the usage of other probability distances such as Wasserstein or Kullback-Leibler ones (Shen et al., 2018).

Once data from the heterogeneous spaces are embedded in the same latent subspace Φ , we can deploy a federated

¹Note that we could also have considered push-forward operators acting on the marginals associated to the clients’ features, see Peyré and Cuturi (2019, Remark 2.5).

learning methodology for training from this novel representation space. At this step, we need to choose which of standard FL approaches, *e.g.* FedAvg (McMahan et al., 2017), or personalised one are more appropriate. Since the proposed distribution alignment training procedure via the use of an anchor distribution might not be perfect, some statistical heterogeneity may still appear in the common latent subspace Φ . Therefore, we aim at solving a *personalised* FL problem where each client has a local model tailored to her specific data distribution in Φ (Tan et al., 2022). By considering an empirical risk minimisation formulation, the resulting data-fitting term we want to minimise writes

$$f(\theta_{1:b}, \phi_{1:b}) = \sum_{i=1}^b \omega_i f_i(\theta_i, \phi_i), \quad (1)$$

where ϕ_i is the aforementioned local embedding function, $\theta_i \in \mathbb{R}^{d_i}$ stands for a local model parameter and $\{\omega_i\}_{i \in [b]}$ are non-negative weights associated to each client such that $\sum_{i=1}^b \omega_i = 1$; and for any $i \in [b]$,

$$f_i(\theta_i, \phi_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell \left(y_i^{(j)}, g_{\theta_i}^{(i)} \left[\phi_i \left(x_i^{(j)} \right) \right] \right). \quad (2)$$

In the local objective function defined in (2), $\ell(\cdot, \cdot)$ stands for a classification loss function between the true label $y_i^{(j)}$ and the predicted one $g_{\theta_i}^{(i)}[\phi_i(x_i^{(j)})]$ where $g_{\theta_i}^{(i)}$ is the local model that admits a personalised architecture parameterised by θ_i and taking as input an embedded feature vector $\phi_i(x_i^{(j)}) \in \Phi$.

Objective Function. At this stage, we are able to integrate the FL paradigm and the local embedding function learning into a global objective function we want to optimise, see (1). Remember that we want to learn the parameters $\{\theta_i\}_{i \in [b]}$ of personalised FL models, in conjunction

with some local embedding functions $\{\phi_i\}_{i \in [b]}$ and shared anchor distributions $\{\mu_c\}$. In particular, the latter have to be aligned with class-conditional distributions $\{\nu_{\phi_i}^{(c)}\}$. We propose to perform this alignment via a Wasserstein regularisation term leading to consider a regularised version of the empirical risk minimisation problem defined in (1), namely

$$\theta_{1:b}^*, \phi_{1:b}^*, \mu_{1:C}^* = \arg \min_{\theta_{1:b}, \phi_{1:b}, \mu_{1:C}} \sum_{i=1}^b F_i(\theta_i, \phi_i, \mu_{1:C}),$$

where for any $i \in [b]$,

$$F_i(\theta_i, \phi_i, \mu_{1:C}) = \omega_i f_i(\theta_i, \phi_i) + \lambda_1 \omega_i \sum_{c \in \mathcal{Y}_i} W_2^2(\mu_c, \nu_{\phi_i}^{(c)}) + \lambda_2 \omega_i \sum_{c \in \mathcal{Y}_i} \frac{1}{J} \sum_{j=1}^J \ell(c, g_{\theta_i}^{(i)}[Z_c^{(j)}]), \quad (3)$$

where $\{Z_c^{(j)}; j \in [J]\}_{c \in [C]}$ stand for samples drawn from $\{\mu_c\}_{c \in [C]}$, and $\lambda_1, \lambda_2 > 0$ are regularisation parameters. The second term in (3) aims at aligning the conditional probability measures of the transformed features. The third one is an optional term aspiring to calibrate the reference measures with the classifier in cases where two or more classes are still ambiguous after mapping onto the common feature space; it has also some benefits to tackle covariate shift in standard FL (Luo et al., 2021).

Design Choices and Justifications. In the sequel, we consider the Gaussian anchor measures $\mu_c = \mathcal{N}(v_c, \Sigma_c)$ where $v_c \in \mathbb{R}^k$ and $c \in [C]$. Note that, under this choice, the samples $\{Z_c^{(j)}; j \in [J]\}_{c \in [C]}$ can be written $Z_c^{(j)} = v_c + L_c \xi_c^{(j)}$ where $\xi_c^{(j)} \sim \mathcal{N}(0_k, \mathbf{I}_k)$ and $L_c \in \mathbb{R}^{k \times k}$ is such that $\Sigma_c = L_c L_c^\top$ by exploiting the positive semi-definite property of Σ_c . Invertibility of L_c is ensured by adding a diagonal matrix $\varepsilon \mathbf{I}_k$ with small positive diagonal elements. One of the key advantages of this Gaussian assumption is that, under mild assumptions, it guarantees the existence of a transport map $T^{(i)}$ such that $T_{\#}^{(i)}(\nu_i) = \mu$, owing to Brenier’s theorem (Santambrogio, 2015) as a mixture of Gaussians admits a density with respect to the Lebesgue measure. Hence, in our case, learning the local embedding functions boils down to approximating this transport map by ϕ_i . In addition, sampling from a Gaussian probability distribution can be performed efficiently (Gilavert et al., 2015; Parker and Fox, 2012; Vono et al., 2022), even in high dimension. We also consider approximating the conditional probability measures $\{\nu_{\phi_i}^{(c)}; c \in \mathcal{Y}_i\}_{i \in [b]}$ by using Gaussian measures $\{\hat{\nu}_{\phi_i}^{(c)} = \mathcal{N}(\hat{m}_i^{(c)}, \hat{\Sigma}_i^{(c)}); c \in \mathcal{Y}_i\}_{i \in [b]}$ such that for any $i \in [b]$ and $c \in [C]$, $\hat{m}_i^{(c)}$ and $\hat{\Sigma}_i^{(c)}$ stand for empirical mean vector and covariance matrix. The relevance of this approximation is detailed in Appendix S1.2.

These two Gaussian choices (for the anchor distribution and the class-conditional distributions) allow us to have a

closed-form expression for the Wasserstein distance of order 2 which appears in (3), see *e.g.* Dowson and Landau (1982); Gelbrich (1990). More precisely, we have for any $i \in [b]$ and $c \in [C]$,

$$W_2^2(\mu_c, \nu_{\phi_i}^{(c)}) = \|v_c - m_i^{(c)}\|^2 + \mathfrak{B}^2(\Sigma_c, \Sigma_i^{(c)}), \quad (4)$$

where $\mathfrak{B}(\cdot, \cdot)$ denotes the Bures distance between two positive definite matrices (Bhatia et al., 2019). In addition to yield the closed-form expression (4), the choice of the Wasserstein distance is motivated by two other important properties. First, it is always finite no matter how degenerate the Gaussian distributions are, contrary to other divergences such as the Kullback-Leibler one (Vilnis and McCallum, 2015). Being able to output a meaningful distance value when supports of distribution do not overlap is a key benefit of the Wasserstein distance, since when initialising ϕ_i , we do not have any guarantee on such overlapping (see illustration given in Figure S4). Second, its minimisation can be handled using efficient algorithms proposed in the optimal transport literature (Muzellec and Cuturi, 2018).

Related Work. As pointed out in Section 1, several existing works can be related to the proposed methodology. Loosely speaking, we can divide these related approaches into three categories namely (i) heterogeneous-architecture personalised FL, (ii) vertical FL and (iii) federated transfer learning.

Compared to traditional horizontal personalised FL (PFL) approaches, so-called *heterogeneous-architecture* ones are mostly motivated by local heterogeneity regarding resource capabilities of clients *e.g.* computation and storage (Collins et al., 2021; Diao et al., 2021; Hong et al., 2022; Makhija et al., 2022; Shamsian et al., 2021; Zhang et al., 2021). Nevertheless, they never consider features defined on heterogeneous subspaces, which is our main motivation. In vertical federated learning (VFL), clients hold disjoint subsets of features. However, a restrictive assumption is that a large number of users are common across the clients (Angelou et al., 2020; Hardy et al., 2017; Romanini et al., 2021; Yang et al., 2019). In addition, up to our knowledge, no vertical personalised FL approach has been proposed so far, which is restrictive if clients have different business objectives and/or tasks. Finally, some works have focused on adapting standard transfer learning approaches with heterogeneous feature domains under the FL paradigm. These *federated transfer learning* (FTL) approaches (Gao et al., 2019; Liu et al., 2020; Mori et al., 2022; Sharma et al., 2019) stand for FL variants of heterogeneous-feature transfer learning where there are b source clients and 1 target client with a target domain. However, these methods do not consider the same setting as ours and assume that clients share a common subset of features. We compare the most relevant approaches among the previous ones in Table 1.

Table 1. Related works. PFL refers to horizontal personalised FL, VFL to vertical FL and FTL to federated transfer learning.

METHOD	TYPE	≠ FEATURE SPACES	MULTI-PARTY	NO SHARED ID	NO SHARED FEATURE
(ZHANG ET AL., 2021)	PFL	✗	✓	✓	✗
(DIAO ET AL., 2021)	PFL	✗	✓	✓	✗
(COLLINS ET AL., 2021)	PFL	✗	✓	✓	✗
(SHAMSIAN ET AL., 2021)	PFL	✗	✓	✓	✗
(HONG ET AL., 2022)	PFL	✗	✓	✓	✗
(MAKHIJA ET AL., 2022)	PFL	✗	✓	✓	✓
FLIC (THIS PAPER)	PFL	✓	✓	✓	✓
(HARDY ET AL., 2017)	VFL	✓	✗	✗	✓
(YANG ET AL., 2019)	VFL	✓	✗	✗	✓
(GAO ET AL., 2019)	FTL	✓	✓	✓	✗
(SHARMA ET AL., 2019)	FTL	✗	✗	✓	✗
(LIU ET AL., 2020)	FTL	✓	✗	✗	✓
(MORI ET AL., 2022)	FTL	✓	✓	✗	✗

3. Algorithm

As detailed in Equation (3), we perform personalisation under the FL paradigm by considering local model architectures $\{g_{\theta_i}^{(i)}\}_{i \in [b]}$ and local weights $\theta_{1:b}$. As an example, we could resort to federated averaging with fine-tuning (e.g. FedAvg-FT, see Collins et al. (2022)), model interpolation (e.g. L2GD, see Hanzely and Richtárik (2020); Hanzely et al. (2020)) or partially local models (e.g. FedRep, see Collins et al. (2021); Oh et al. (2022); Singhal et al. (2021)). Table 2 details how these methods can be embedded into the proposed methodology.

Table 2. Current personalised FL techniques that can be embedded in the proposed framework. The parameters α, β_i stand for model weights while $\omega \in [0, 1]$.

Algorithm	Local model	Local weights
FedAvg-FT	$g_{\theta_i}^{(i)} = g_{\theta_i}$	θ_i
L2GD	$g_{\theta_i}^{(i)} = g_{\theta_i}$	$\theta_i = \omega\alpha + (1 - \omega)\beta_i$
FedRep	$g_{\theta_i}^{(i)} = g_{\beta_i}^{(i)} \circ g_{\alpha}$	$\theta_i = [\alpha, \beta_i]$

In Algorithm 1, we detail the pseudo-code associated to a specific instance of the proposed methodology when FedRep is resorted to learn model parameters $\{\theta_i = [\alpha, \beta_i]\}_{i \in [b]}$ under the FL paradigm. In this setting, α stand for the shared weights associated to the first layers of a neural network architecture and β_i for local ones aiming at performing personalised classification. Besides these two learnable parameters, the algorithm also learns the local embedding functions $\phi_{1:b}$ and the anchor distribution $\mu_{1:C}$. In practice, at a given epoch t of the algorithm, a subset $A_{t+1} \subseteq [b]$ of clients are selected to participate to the training process. Those clients receive the current latent anchor distribution $\mu_{1:C}^{(t)}$ and the current shared representation $\alpha^{(t)}$. Then, each client locally updates ϕ_i, β_i and her local ver-

sions of $\alpha^{(t)}$ and $\mu_{1:C}^{(t)}$. Afterwards, clients send back to the server an updated version of $\alpha^{(t)}$ and $\mu_{1:C}^{(t)}$. Updated global parameters $\alpha^{(t+1)}$ and $\mu_{1:C}^{(t+1)}$ are then obtained by weighted averaging of client updates on appropriate manifolds. The use of the Wasserstein loss in (3) naturally leads to perform averaging of the local anchor distributions via a Wasserstein barycenter; algorithmic details are provided in the next paragraph. In Algorithm 1, we use for the sake of simplicity the notation $\text{DescStep}(F_i^{(t,m)}, \cdot)$ to denote a (stochastic) gradient descent step on the function $F_i^{(t,m)} = F_i(\beta_i^{(t,m)}, \phi_i^{(t,m)}, \alpha^{(t)}, \mu_{1:C}^{(t)})$ with respect to a subset of parameters in $(\theta_i, \phi_i, \mu_{1:C})$. This subset is specified in the second argument of DescStep . An explicit version of Algorithm 1 is provided in Appendix, see Algorithm S2.

Note that we take into account key inherent challenges to federated learning namely *partial participation* and *communication bottleneck*. Indeed, we cope with the client/server upload communication issue by allowing each client to perform multiple steps (here $M \in \mathbb{N}^*$) so that communication is only required every M local steps. This allows us to consider updating global parameters, locally, via only one stochastic gradient descent step and hence avoiding the client drift phenomenon (Karimireddy et al., 2020).

Averaging Anchor Distributions. In this paragraph, we provide algorithmic details regarding steps 14 and 20 in Algorithm 1. For any $c \in [C]$, the anchor distribution μ_c involves two learnable parameters namely the mean vector v_c and the covariance matrix Σ_c . Regarding the former, step 14 stands for a (stochastic) gradient descent step aiming to obtain a local version of v_c denoted by $v_{i,c}^{(t+1)}$ and step 20 boils down to compute $v_c^{(t+1)} = (b/|A_{t+1}|) \sum_{i \in A_{t+1}} \omega_i v_{i,c}^{(t+1)}$. To enforce the positive

Algorithm 1 FLIC

Require: initialisation $\alpha^{(0)}, \mu_{1:C}^{(0)}, \phi_{1:b}^{(0,0)}, \beta_{1:b}^{(0,0)}$.

- 1: **for** $t = 0$ **to** $T - 1$ **do**
- 2: Sample a set of A_{t+1} of active clients.
- 3: **for** $i \in A_{t+1}$ **do**
- 4: The central server sends $\alpha^{(t)}$ and $\mu_{1:C}^{(t)}$ to A_{t+1} .
- 5: // Update local parameters
- 6: **for** $m = 0$ **to** $M - 1$ **do**
- 7: $\phi_i^{(t,m+1)} \leftarrow \text{DescStep} \left(F_i^{(t,m)}, \phi_i^{(t,m)} \right)$.
- 8: $\beta_i^{(t,m+1)} \leftarrow \text{DescStep} \left(F_i^{(t,m)}, \beta_i^{(t,m)} \right)$.
- 9: **end for**
- 10: $\phi_i^{(t+1,0)} = \phi_i^{(t,M)}$.
- 11: $\beta_i^{(t+1,0)} = \beta_i^{(t,M)}$.
- 12: // Update global parameters
- 13: $\alpha_i^{(t+1)} \leftarrow \text{DescStep} \left(F_i^{(t,M)}, \alpha^{(t)} \right)$.
- 14: $\mu_{i,1:C}^{(t+1)} \leftarrow \text{DescStep} \left(F_i^{(t,M)}, \mu_{1:C}^{(t)} \right)$.
- 15: // Communication with the server
- 16: Send $\alpha_i^{(t+1)}$ and $\mu_{i,1:C}^{(t+1)}$ to central server.
- 17: **end for**
- 18: // Averaging global parameters
- 19: $\alpha^{(t+1)} = \frac{b}{|A_{t+1}|} \sum_{i \in A_{t+1}} \omega_i \alpha_i^{(t+1)}$
- 20: $\mu_{1:C}^{(t+1)} \leftarrow \text{WassersteinBarycenter}(\{\mu_{i,1:C}^{(t+1)}\})$
- 21: **end for**

Ensure: parameters $\alpha^{(T)}, \mu_{1:C}^{(T)}, \phi_{1:b}^{(T,0)}, \beta_{1:b}^{(T,0)}$.

semi-definite constraint of the covariance matrix, we rewrite it as $\Sigma_c = L_c L_c^\top$ where $L_c \in \mathbb{R}^{k \times k}$ and optimise in step 14 with respect to the factor L_c instead of Σ_c . We can handle the gradient computation of the Bures distance in step 14 using the work of Muzellec and Cuturi (2018); and obtain a local factor $L_{i,c}^{(t+1)}$ at iteration t . In step 20, we compute $L_c^{(t+1)} = (b/|A_{t+1}|) \sum_{i \in A_{t+1}} \omega_i L_{i,c}^{(t+1)}$ and set $\Sigma_c^{(t+1)} = L_c^{(t+1)} [L_c^{(t+1)}]^\top$. When $\lambda_2 = 0$ in (3), these mean vector and covariance matrix updates exactly boil down to perform one stochastic (because of partial participation) gradient descent step to solve the Wasserstein barycenter problem $\arg \min_{\mu_c} \sum_{i=1}^b \omega_i W_2^2(\mu_c, \nu_{\phi_i}^{(c)})$.

4. Non-Asymptotic Convergence Guarantees in a Simplified Setting

Deriving non-asymptotic convergence bounds for Algorithm 1 in the general case is challenging since the considered C -class classification problem leads to jointly solving personalised FL and federated Wasserstein barycenter problems. Regarding the latter, obtaining non-asymptotic convergence results is still an active research area in the centralised learning framework (Altschuler et al., 2021). As such, we propose to analyse a simpler regression frame-

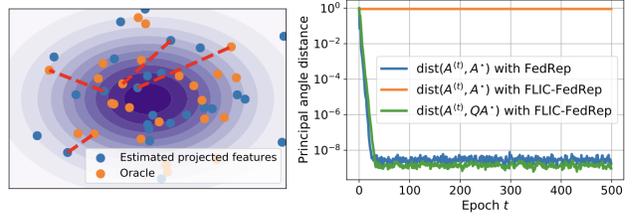


Figure 2. Red dashed line indicates that the two embedded features $\phi_i^*(x_i^{(j)})$ and $\hat{\phi}_i(x_i^{(j)})$ come from the same initial raw feature $x_i^{(j)}$. On test data, mean prediction errors for both FedRep operating on $\phi_i^*(x_i^{(j)})$ and Algorithm S3 (referred to as FLIC-FedRep) are similar ($\approx 4.98 \times 10^{-5}$).

work where the anchor distribution is known beforehand and not learnt under the FL paradigm.

More precisely, we assume that $x_i^{(j)} \sim \mathcal{N}(m_i, \Sigma_i)$ with $m_i \in \mathbb{R}^{k_i}$ and $\Sigma_i \in \mathbb{R}^{k_i \times k_i}$ for $i \in [b], j \in [n_i]$. In addition, we consider that the continuous scalar labels are generated via the oracle model $y_i^{(j)} = (A^* \beta_i^*)^\top \phi_i^*(x_i^{(j)})$ where $A^* \in \mathbb{R}^{k \times d}$, $\beta_i^* \in \mathbb{R}^d$ and $\phi_i^*(\cdot)$ are ground-truth parameters and feature transformation function, respectively. We make the following assumptions on the ground truth.

- H1.** (i) For any $i \in [b], j \in [n_i]$, embedded features $\phi_i^*(x_i^{(j)})$ are distributed according to $\mathcal{N}(0_k, I_k)$.
(ii) Ground-truth model parameters satisfy $\|\beta_i^*\|_2 = \sqrt{d}$ for $i \in [b]$ and A^* has orthonormal columns.
(iii) For any $t \in \{0, \dots, T-1\}, |A_{t+1}| = b'$ with $1 \leq b' \leq b$, and if we select b' clients, their ground-truth head parameters $\{\beta_i^*\}_{i \in A_{t+1}}$ span \mathbb{R}^d .
(iv) In (2), $\ell(\cdot, \cdot)$ is the ℓ_2 norm, $\omega_i = 1/b$, $\theta_i = [A, \beta_i]$ and $g_{\theta_i}^{(i)}(x) = (A\beta_i)^\top x$ for $x \in \mathbb{R}^k$.

Under H1-(i), Delon et al. (2022, Theorem 4.1) show that ϕ_i^* can be expressed as a non-unique affine map with closed-form expression. To recover the true latent distribution $\mu = \mathcal{N}(0_k, I_k)$, we propose to estimate $\hat{\phi}_i$ by leveraging this closed-form mapping between $\mathcal{N}(m_i, \Sigma_i)$ and μ . Because of the non-unicity of ϕ_i^* , we show in Theorem 1 that we can only recover it up to a matrix multiplication. Interestingly, Theorem 1 shows that the global representation $A^{(T)}$ learnt via FedRep (see Algorithm S3 in Appendix) is able to correct this feature mapping indetermination. Associated convergence behavior is illustrated in Figure 2 on a toy example whose details are postponed to Appendix S2.

Theorem 1. Assume H1. Then, for any $x_i \in \mathbb{R}^{k_i}$, we have $\hat{\phi}_i(x_i) = Q\phi_i^*(x_i)$ where $Q \in \mathbb{R}^{k \times k}$ is of the form $\text{diag}_k(\pm 1)$. Under additional technical assumptions detailed in Appendix S2, we have for any $t \in \{0, \dots, T-1\}$ and with high probability,

$$\text{dist}(A^{(t+1)}, QA^*) \leq (1 - \kappa)^{(t+1)/2} \text{dist}(A^{(0)}, QA^*),$$

Table 3. Performance over 3 runs of our **FLIC** model and the competitors on some real-data problems (*Digits* and *TextCaps* data set).

Data sets (setting)	Local	FedHeNN	FLIC -Class	FLIC -HL
Digits ($b = 100$, 3 Classes/client)	97.49	97.45	97.83	97.70
Digits ($b = 100$, 5 Classes/client)	96.16	96.15	96.46	96.31
Digits ($b = 200$, 3 Classes/client)	93.33	93.40	94.50	94.51
Digits ($b = 200$, 5 Classes/client)	86.50	87.22	90.66	90.63
TextCaps ($b = 100$, 2 Classes/client)	84.19	83.99	89.14	89.68
TextCaps ($b = 100$, 3 Classes/client)	76.04	75.39	81.27	81.50
TextCaps* ($b = 200$, 2 Classes/client)	83.78	83.89	87.73	87.74
TextCaps* ($b = 200$, 3 Classes/client)	74.95	74.77	79.08	78.49

where $\kappa \in (0, 1)$ is detailed explicitly in Theorem S3 and dist denotes the principal angle distance.

5. Numerical Experiments

In this section, we aim at illustrating how our algorithm **FLIC**, when using FedRep as FL approach, works in practice and showcasing its numerical performances. We consider several toy problems with different characteristics of heterogeneity; as well as experiments on real data namely a digit classification problem from images of different sizes and an object classification problem from either images or text captioning on clients.

Baselines. Since the problem we are addressing is novel, there exists no FL competitor that can serve as a baseline beyond local learning. However, we propose to modify the methodology proposed in Makhija et al. (2022) to make it applicable to clients with heterogeneous feature spaces. This latter approach can handle local representation models with different architectures and the key idea, coined Representation Alignment Dataset (RAD), is to calibrate those models by matching the latent representation of some fixed data inputs shared by the server to all clients. In our case, we can not use the same RAD across all clients due to the different dimensionality of the local models. A simple alternative that we consider in our experiments is to build a RAD given the largest dimension space among all clients and then prune it to obtain a lower-dimensional RAD suitable to each client. We refer to the corresponding algorithm as FedHeNN.

We are going to consider the same architecture networks for all baselines. As Makhija et al. (2022) considers all but the last layer of the network as the representation learning module, for a fair comparison, we also assume the same for our approach. Hence, in our case, the last layer is the classifier layer and the alignment with the latent reference distribution applies on the penultimate layer. This model is referred to as **FLIC**-Class, in which all weights are thus local (α is empty and β_i refers to the last layer). In addition,

we also have a model, coined **FLIC**-HL, which has an additional trainable global hidden layer, which α being the parameter of linear layer and β_i the parameter of the classification layer.

Data Sets. We consider three different classification problems to assess the performances of our approach. First, we are considering a toy classification problem with $C = 20$ classes and where each class-conditional distribution is a Gaussian with random mean. Covariance matrices of all classes are the same and considered fixed. Using this toy data set, we are considering two sub-experiments. For the first one, named *noisy features* (and labelled *toy NF* in figures), we consider a 5-dimensional problem ($k = 5$) and for each client add some random spurious features which dimensionality goes up to 10. Hence, in this case $k_i \in [5, 15]$. For the second sub-experiment, denoted as *linear mapping* (and labelled *toy LM* in figures), we apply a Gaussian random linear mapping to the original data which are of dimension 30. The output dimension of the mapping is uniformly drawn from 5 to 30 leading to $k_i \in [5, 30]$. More details are provided in Appendix.

The second problem we consider is a digit classification problem with the original MNIST and USPS data sets which are respectively of dimension 28×28 and 16×16 and we assume that a client hosts either a subset of MNIST or USPS dataset. Finally, the last classification problem is associated to a subset of the *TextCaps* data set (Sidorov et al., 2020), which is an image captioning data set, that we convert into a 4-class classification problem, with about 12,000 and 3,000 examples for respectively training and testing, either based on the caption text or the image. Some examples of image/caption pairs as well as more details on how the dataset has been obtained are shown in the Figure S5. The caption has been embedded into a 768-dimensional vector using a pre-trained Bert embedding and the image into a 512-dimensional ones using a pre-trained ResNet model. We further generated some heterogeneity by randomly pruning 10% of these features on each client. Again, we assume that a client hosts either some image

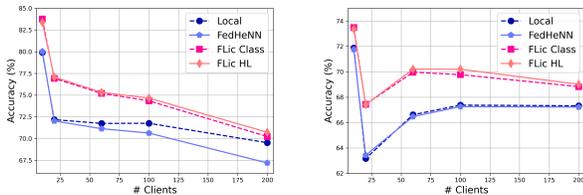


Figure 3. Performance of **FLIC** and competitors on the toy data sets with respect to the number of clients. (left) Gaussian classes in dimension $k = 5$ with added noisy feature. (right) Gaussian classes in dimension $k = 30$, transformed by a random map. Only 3 classes are present on each client among the 20 possible ones.

embeddings or text embeddings. For all simulations, we assume prior probability shift *e.g.* each client will have access to data of only specific classes.

Experimental Setting. For the experimental analysis, we use the codebase of Collins et al. (2021) with some modifications to meet our setting. For all experiments, we consider $T = 50$ communication rounds for all algorithms; and at each round, a client participation rate of $r = 0.1$. The number of local epochs for training has been set to $M = 10$. As optimisers, we have used an Adam optimiser with a learning rate of 0.001 for all problems and approaches. Further details are given in Appendix S3.3. For each component of the latent anchor distribution, we consider a Gaussian with learnable mean vectors and fixed Identity covariance matrix. As such, the Wasserstein barycenter computation boils down to simply average the mean of client updates and for computing the third term in Equation (3), we just sample from the Gaussian distribution. Accuracies are computed as the average accuracy over all clients after the last epoch in which all local models are trained.

Results on Toy Data Sets. Figure 3 depicts the performance, averaged over 5 runs, of the different algorithms with respect to the number of clients and when only 3 classes are present in each client. For both data sets, we can note that for the *noisy feature* setting, **FLIC** improves on FedHeNN of about 3% of accuracy across the setting and performs better than local learning. For the *linear mapping* setting, **FLIC** achieves better than other approaches with a gain of performance of about 4% while the gap tends to decrease as the number of clients increases. Interestingly, **FLIC-HL** performs slightly better than **FLIC-Class** showing the benefit of using a shared representation layer α .

Results on Digits and TextCaps Data Sets. Performance, averaged over 3 runs, of all algorithms on the real-word problems are reported in Table 3. For the *Digits* data set problem, we first remark that in all situations, FL algorithms performs a bit better than local learning. In addition, both variants of **FLIC** achieve better accuracy than

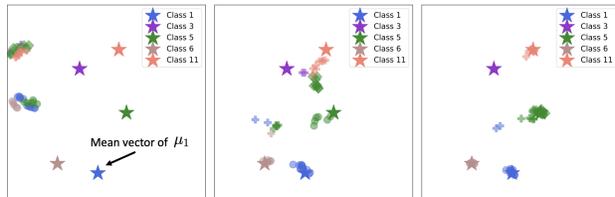


Figure 4. . 2D *t-sne* projection of 5 classes partially shared by 3 clients for the **toy LM** dataset after learning the local transformation functions for (left) 10 epochs, (middle) 50 epochs, (right) 100 epochs. The three different markers represent the different clients. the \star marker represents the class-conditional mean of the reference distribution. We note that training set converges towards those means.

competitors. Difference in performance in favor our **FLIC** reaches 3% for the most difficult problem. For the *TextCaps* data set, gains in performance of **FLIC-HL** reach about 4% across settings. While FedHeNN and **FLIC** algorithms follow the same underlying principle (alignment of representation in a latent space), we believe that our framework benefits from the use of the latent anchor distributions, avoiding the need of sampling from the original space. Instead, FedHeNN may fail as the sampling strategy of their RAD approach suffers from the curse of dimensionality and does not properly lead to a successful feature alignment.

Additional Experiments in Appendix. Due to the limited number of pages, additional experiments are postponed to the Appendix. In particular, we investigate the impact of pre-training the local embedding functions for a fixed reference distribution as in Section 4, before running the proposed algorithm detailed in Algorithm 1. The main message is that pre-training helps in enhancing performance but may lead to overfitting if too many epochs are considered. We also analyse the impact of client participation rate at each communication round reaching the conclusion that our model is robust to participation rate.

6. CONCLUSION

We have introduced a new framework for personalised FL when clients have heterogeneous feature spaces. We proposed a novel FL algorithm involving two key components: (i) a local feature embedding function; and (ii) a latent anchor distribution which allows to match similar semantical information from each client. Experiments on relevant data sets have shown that **FLIC** achieves better performances than competing approaches. Finally, we provided theoretical support to the proposed methodology, notably via a non-asymptotic convergence result.

REFERENCES

- Mokhtar Z Alaya, Maxime Bérar, Gilles Gasso, and Alain Rakotomamonjy. Theoretical guarantees for bridging metric measure embedding and optimal transport. *Neurocomputing*, 468:416–430, 2022.
- Jason Altschuler, Sinho Chewi, Patrik Robert Gerber, and Austin J Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. In *Advances in Neural Information Processing Systems*, 2021.
- David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.
- David Alvarez-Melis, Stefanie Jegelka, and Tommi S. Jaakkola. Towards Optimal Transport with Global Invariances. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1870–1879, 2019.
- Nick Angelou, Ayoub Benaissa, Bogdan Cebere, William Clark, Adam James Hall, Michael A Hoeh, Daniel Liu, Pavlos Papadopoulos, Robin Roehm, Robert Sandmann, et al. Asymmetric private set intersection with applications to contact tracing and private vertical federated machine learning. *arXiv preprint arXiv:2011.09350*, 2020.
- Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.
- Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.
- Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *Conference on Computer and Communications Security*, page 1175–1191, 2017.
- Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning Generative Models across Incomparable Spaces. In *International Conference on Machine Learning*, volume 97, pages 851–861, 2019.
- Wei-Ning Chen, Christopher A Choquette Choo, Peter Kairouz, and Ananda Theertha Suresh. The Fundamental Price of Secure Aggregation in Differentially Private Federated Learning. In *International Conference on Machine Learning*, volume 162, pages 3056–3089, 2022.
- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting Shared Representations for Personalized Federated Learning. In *International Conference on Machine Learning*, pages 2089–2099, 2021.
- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. FedAvg with Fine Tuning: Local Updates Lead to Representation Learning. In *Advances in Neural Information Processing Systems*, 2022.
- Julie Delon, Agnes Desolneux, and Antoine Salmona. Gromov-Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022.
- Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients. In *International Conference on Learning Representations*, 2021.
- D.C Dowson and B.V Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.
- Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. Multimodal and Multilingual Embeddings for Large-Scale Speech Mining. In *Advances in Neural Information Processing Systems*, volume 34, pages 15748–15761, 2021.
- Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. Privacy-preserving Heterogeneous Federated Transfer Learning. In *IEEE International Conference on Big Data (Big Data)*, pages 2552–2559, 2019.
- Matthias Gelbrich. On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.
- C. Gilavert, S. Moussaoui, and J. Idier. Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems Using MCMC. *IEEE Transactions on Signal Processing*, 63(1):70–80, 2015.
- Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

- Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint arXiv:2010.02372*, 2020.
- Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized Federated Learning: A Unified Framework and Universal Optimization Techniques. *arXiv preprint arXiv: 2102.09743*, 2021.
- Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.
- Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization. In *International Conference on Learning Representations*, 2022.
- J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.
- Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-Rank Matrix Completion Using Alternating Minimization. In *ACM Symposium on Theory of Computing*, page 665–674, 2013.
- Yihan Jiang, Jakub Konevcný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, K. A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G.L. D’Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konevcný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.
- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143, 2020.
- Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32:5917–5928, 2019.
- Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study. *arXiv preprint arxiv: 2105.03790*, 2021.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In *Machine Learning and Systems*, volume 2, pages 429–450, 2020.
- Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. A Secure Federated Transfer Learning Framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.
- Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No Fear of Heterogeneity: Classifier Calibration for Federated Learning with Non-IID Data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Disha Makhija, Xing Han, Nhat Ho, and Joydeep Ghosh. Architecture Agnostic Federated Learning for Neural Networks. In *International Conference on Machine Learning*, volume 162, pages 14860–14870, 2022.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282, 2017.
- Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.
- Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. PPFL: Privacy-Preserving Federated Learning with Trusted Execution Environments. In *International Conference on Mobile Systems, Applications, and Services*, page 94–108, 2021.
- Junki Mori, Isamu Teranishi, and Ryo Furukawa. Continual Horizontal Federated Learning for Heterogeneous Data. *arXiv preprint arXiv:2203.02108*, 2022.

- Boris Muzellec and Marco Cuturi. Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bis-sacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- Jaehoon Oh, SangMook Kim, and Se-Young Yun. Fed-BABU: Toward Enhanced Representation for Federated Image Classification. In *International Conference on Learning Representations*, 2022.
- A. Parker and C. Fox. Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients. *SIAM Journal on Scientific Computing*, 34(3):B312–B334, 2012.
- Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.
- Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.
- Daniele Romanini, Adam James Hall, Pavlos Papadopoulos, Tom Titcombe, Abbas Ismail, Tudor Cebere, Robert Sandmann, Robin Roehm, and Michael A Hoeh. PyVertical: A Vertical Federated Learning Framework for Multi-headed SplitNN. *arXiv preprint arXiv:2104.00489*, 2021.
- Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58–63):94, 2015.
- Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized Federated Learning using Hypernetworks. In *International Conference on Machine Learning*, volume 139, pages 9489–9502, 2021.
- Shreya Sharma, Chaoping Xing, Yang Liu, and Yan Kang. Secure and Efficient Federated Transfer Learning. In *IEEE International Conference on Big Data (Big Data)*, pages 2569–2576, 2019.
- Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In *Conference on Artificial Intelligence*, 2018.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758, 2020.
- Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated Reconstruction: Partially Local Federated Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 11220–11232, 2021.
- Alysa Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–17, 2022.
- Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. In *International Conference on Learning Representations*, 2020.
- R. Veldhuis. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Processing Letters*, 9(3): 96–99, 2002.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
- Cedric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.
- Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. In *International Conference on Learning Representations*, 2015.
- Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. High-dimensional Gaussian sampling: A review and a unifying approach based on a stochastic proximal point algorithm. *SIAM Review*, 64(1):3–56, 2022.
- Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, and Lawrence Carin. Learning Autoencoders with Relational Regularization. In *International Conference on Machine Learning*, volume 119, pages 10576–10586, 2020.
- Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated Machine Learning: Concept and Applications. *Transactions on Intelligent Systems and Technology*, 10(2), 2019.
- Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central Moment Discrepancy (CMD) for Domain-Invariant

Representation Learning. In *International Conference on Learning Representations*, 2017.

Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized Knowledge Transfer for Personalized Federated Learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Fan Zhou, Brahim Chaib-draa, and Boyu Wang. Multi-task Learning by Leveraging the Semantic Information. *Conference on Artificial Intelligence*, 35(12):11088–11096, 2021.

Tailin Zhou, Jun Zhang, and Danny Tsang. FedFA: Federated Learning with Feature Anchors to Align Feature and Classifier for Heterogeneous Data. *arXiv preprint arXiv: 22211.09299*, 2022.

Appendix

Table of Contents

S1 Algorithmic and Theoretical Insights	14
S1.1 Some Limited but Common Alternatives to Cope with Feature Space Heterogeneity	14
S1.2 Use of Wasserstein Losses Involving Empirical Probability Distributions	14
S1.3 Detailed Pseudo-Code for Algorithm 1	14
S1.4 Additional Algorithmic Insights	15
S2 Proof of Theorem 1	16
S2.1 Estimation of the Feature Transformation Functions	16
S2.2 Proof of Theorem 1	16
S2.3 Technical Lemmata	18
S3 Experimental Details	23
S3.1 Reference Distribution for Regression	23
S3.2 Data Sets	23
S3.3 Models and Learning Parameters	24
S3.4 Ablating Loss Curves	24
S3.5 On Importance of Alignment Pre-Training and Updates.	25
S3.6 On the Impact of the Participation Rate	25

SUPPLEMENTARY MATERIAL

Notations and conventions. We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ -field of \mathbb{R}^d , $\mathbb{M}(\mathbb{R}^d)$ the set of all Borel measurable functions f on \mathbb{R}^d and $\|\cdot\|$ the Euclidean norm on \mathbb{R}^d . For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{M}(\mathbb{R}^d)$ a μ -integrable function, denote by $\mu(f)$ the integral of f with respect to (w.r.t.) μ . Let μ and ν be two sigma-finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Denote by $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν and $d\mu/d\nu$ the associated density. We say that ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all measurable set A of \mathbb{R}^d , $\zeta(A \times \mathbb{R}^d) = \mu(A)$ and $\zeta(\mathbb{R}^d \times A) = \nu(A)$. We denote by $\mathcal{T}(\mu, \nu)$ the set of transference plans of μ and ν . In addition, we say that a couple of \mathbb{R}^d -random variables (X, Y) is a coupling of μ and ν if there exists $\zeta \in \mathcal{T}(\mu, \nu)$ such that (X, Y) are distributed according to ζ . We denote by $\mathcal{P}_1(\mathbb{R}^d)$ the set of probability measures with finite 1-moment: for all $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \|x\| d\mu(x) < \infty$. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite 2-moment: for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \|x\|^2 d\mu(x) < \infty$. We define the squared Wasserstein distance of order 2 associated with $\|\cdot\|$ for any probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$W_2^2(\mu, \nu) = \inf_{\zeta \in \mathcal{T}(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\zeta(x, y).$$

By Villani (2008, Theorem 4.1), for all μ, ν probability measures on \mathbb{R}^d , there exists a transference plan $\zeta^* \in \mathcal{T}(\mu, \nu)$ such that for any coupling (X, Y) distributed according to ζ^* , $W_2(\mu, \nu) = \mathbb{E}[\|x - y\|^2]^{1/2}$. This kind of transference plan (respectively coupling) will be called an optimal transference plan (respectively optimal coupling) associated with W_2 . By Villani (2008, Theorem 6.16), $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein distance W_2 is a complete separable metric space. For the sake of simplicity, with little abuse, we shall use the same notations for a probability distribution and its

associated probability density function. For $n \geq 1$, we refer to the set of integers between 1 and n with the notation $[n]$. The d -multidimensional Gaussian probability distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is denoted by $\mathcal{N}(\mu, \Sigma)$. Given two matrices $M, N \in \mathbb{R}^{k \times d}$, the principal angle distance between the subspaces spanned by the columns of M and N is given by $\text{dist}(M, N) = \|\hat{M}_\perp^\dagger \hat{N}\|_2 = \|\hat{N}_\perp^\dagger \hat{M}\|_2$ where \hat{M}, \hat{N} are orthonormal bases of $\text{Span}(M)$ and $\text{Span}(N)$, respectively. Similarly, $\hat{M}_\perp, \hat{N}_\perp$ are orthonormal bases of orthogonal complements $\text{Span}(M)^\perp$ and $\text{Span}(N)^\perp$, respectively. This principal angle distance is upper bounded by 1, see [Jain et al. \(2013, Definition 1\)](#).

Outline. This supplementary material aims at providing the interested reader with a further understanding of the statements pointed out in the main paper. More precisely, in Appendix S1, we support the proposed methodology **FLIC** with algorithmic and theoretical details. In Appendix S2, we prove the main results stated in the main paper. Finally, in Appendix S3, we provide further experimental design choices and show complementary numerical results.

S1. Algorithmic and Theoretical Insights

In this section, we highlight alternative but limited ways to cope with feature space heterogeneity; and justify the usage, in the objective function (3) of the main paper, of Wasserstein distances with empirical probability distributions instead of true ones. In addition, we detail the general steps depicted Algorithm 1.

S1.1. Some Limited but Common Alternatives to Cope with Feature Space Heterogeneity

Depending on the nature of the spaces $\{\mathcal{X}_i\}_{i \in [b]}$, the feature transformation functions $\{\phi_i\}_{i \in [b]}$ can be either known beforehand or more difficult to find. As an example, if for any $i \in [b]$, $\mathcal{X}_i \subseteq \mathcal{X}$, then we can set mask functions as feature transformation functions in order to only consider features that are shared across all the clients. Besides, we could consider multimodal embedding models to perform feature transformation on each client ([Duquenne et al., 2021](#)). For instance, if clients own either pre-processed images or text of titles, descriptions and tags, then we can use the Contrastive Language-Image Pre-Training (CLIP) model as feature transformation function ([Radford et al., 2021](#)). These two examples lead to the solving of a classical (personalised) FL problem which can be performed using existing state-of-the-art approaches. However, when the feature transformation functions cannot be easily found beforehand, solving the FL problem at stake becomes more challenging and has never been addressed in the federated learning literature so far, up to the authors' knowledge.

S1.2. Use of Wasserstein Losses Involving Empirical Probability Distributions

Since the true probability distributions $\{\nu_{\phi_i}^{(c)}; c \in \mathcal{Y}_i\}_{i \in [b]}$ are unknown a priori, we propose in the main paper to estimate the latter using $\{\hat{\nu}_{\phi_i}^{(c)}; c \in \mathcal{Y}_i\}_{i \in [b]}$ and to replace $W_2^2(\mu_c, \nu_{\phi_i}^{(c)})$ by $W_2^2(\mu_c, \hat{\nu}_{\phi_i}^{(c)})$ in the objective function (3) in the main paper. As shown in the following result, this assumption is theoretically grounded when the marginal distributions of the input features are Gaussian.

Theorem S2. *For any $i \in [b]$ and $c \in [C]$, let $n_i^{(c)} = |D_i^{(c)}|$ where $D_i^{(c)}$ denotes the subset of the local data set D_i only involving observations associated to the label c . Besides, assume that $\nu_{\phi_i}^{(c)}$ is Gaussian with mean vector $m_i^{(c)} \in \mathbb{R}^k$ and full-rank covariance matrix $\Sigma_i^{(c)} \in \mathbb{R}^{k \times k}$. Then, we have in the limiting case $n_i^{(c)} \rightarrow \infty$,*

$$\sqrt{n_i^{(c)}} \left(W_2^2(\mu_c, \hat{\nu}_{\phi_i}^{(c)}) - W_2^2(\mu_c, \nu_{\phi_i}^{(c)}) \right) \xrightarrow{\text{in distribution}} Z_i^{(c)},$$

where $Z_i^{(c)} \sim \mathcal{N}(0, s_i^{(c)})$ and $s_i^{(c)} = 4(m_i^{(c)} - v_c)^\top \Sigma_i^{(c)}(m_i^{(c)} - v_c) + 2\text{Tr}(\Sigma_i^{(c)}\Sigma_c) - 4\sum_{j=1}^k \kappa_j^{1/2} r_j^\top \Sigma_c^{-1/2} \Sigma_i^{(c)} \Sigma_c^{1/2} r_j$, with $\{\kappa_j, r_j\}_{j \in [k]}$ standing for (eigenvalue, eigenvector) pairs of the symmetric covariance matrix $\Sigma_i^{(c)}$.

Proof. The proof follows from [Rippl et al. \(2016, Theorem 2.1\)](#) with the specific choices $\mu_1 = \nu_{\phi_i}^{(c)}$, $\mu_2 = \mu_c$ and $\hat{\mu}_1 = \hat{\nu}_{\phi_i}^{(c)}$ which are defined in Section 2 in the main paper. \square

S1.3. Detailed Pseudo-Code for Algorithm 1

In Algorithm S2, we provide algorithmic support to Algorithm 1 in the main paper by detailing how to perform each step. Note that we use the decomposition $\Sigma = LL^\top$ to enforce the positive semi-definite constraint for the covariance matrix Σ .

Algorithm S2 Detailed version of **FLIC** when using FedRep

Require: initialisation $\alpha^{(0)}, \mu_{1:C}^{(0)} = [\Sigma_{1:C}^{(0)}, v_{1:C}^{(0)}]$ with $\Sigma_c^{(0)} = L_c^{(0)} [L_c^{(0)}]^\top$, $\phi_{1:b}^{(0,0)}, \beta_{1:b}^{(0,0)}$ and step-size $\eta \leq \bar{\eta}$ for some $\bar{\eta} > 0$.

- 1: **for** $t = 0$ **to** $T - 1$ **do**
- 2: Sample a set of A_{t+1} of active clients.
- 3: **for** $i \in A_{t+1}$ **do**
- 4: The central server sends $\alpha^{(t)}$ and $\mu_{1:C}^{(t)}$ to A_{t+1} .
- 5: *// Update local parameters*
- 6: **for** $m = 0$ **to** $M - 1$ **do**
- 7: Sample a fresh batch $I_{t+1}^{(i,m)}$ of n_i' samples with $n_i' \in [n_i]$.
- 8: Sample $Z_c^{(j,t,m)} \sim \mu_c^{(t)}$ for $j \in I_{t+1}^{(i,m)}$ and $c \in \mathcal{Y}_i$ via $Z_c^{(j,t,m)} = v_c^{(t)} + L_c^{(t)} \xi_i^{(t,m)}$ where $\xi_i^{(t,m)} \sim \mathcal{N}(0_k, \mathbf{I}_k)$.
- 9:
$$\phi_i^{(t,m+1)} = \phi_i^{(t,m)} - \eta \frac{n_i}{|I_{t+1}^{(i,m)}|} \sum_{j \in I_{t+1}^{(i,m)}} \nabla_{\phi_i} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,m)}]}^{(i)} \left[\phi_i^{(t,m)} \left(x_i^{(j)} \right) \right] \right) - \eta \lambda_1 \sum_{c \in \mathcal{Y}_i} \nabla_{\phi_i} \mathcal{W}_2^2 \left(\mu_c^{(t)}, \nu_{\phi_i^{(t,m)}}^{(c)} \right).$$
- 10:
$$\beta_i^{(t,m+1)} \leftarrow \beta_i^{(t,m)} - \eta \frac{n_i}{|I_{t+1}^{(i,m)}|} \sum_{j \in I_{t+1}^{(i,m)}} \left\{ \nabla_{\beta_i} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,m)}]}^{(i)} \left[\phi_i^{(t,m)} \left(x_i^{(j)} \right) \right] \right) - \eta \lambda_2 \sum_{c \in \mathcal{Y}_i} \nabla_{\beta_i} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,m)}]}^{(i)} \left[Z_c^{(j,t,m)} \right] \right) \right\}.$$
- 11: **end for**
- 12: $\phi_i^{(t+1,0)} = \phi_i^{(t,M)}$.
- 13: $\beta_i^{(t+1,0)} = \beta_i^{(t,M)}$.
- 14: *// Update global parameters*
- 15:
$$\alpha_i^{(t+1)} \leftarrow \alpha^{(t)} - \eta \frac{n_i}{|I_{t+1}^{(i,M)}|} \sum_{j \in I_{t+1}^{(i,M)}} \left\{ \nabla_{\alpha} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,M)}]}^{(i)} \left[\phi_i^{(t,M)} \left(x_i^{(j)} \right) \right] \right) - \eta \lambda_2 \sum_{c \in \mathcal{Y}_i} \nabla_{\alpha} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,M)}]}^{(i)} \left[Z_c^{(j,t,M)} \right] \right) \right\}.$$
- 16: **for** $c = 1$ **to** C **do**
- 17: Update $\hat{m}_i^{(c,t)}, \hat{\Sigma}_i^{(c,t)}$ using $\phi_i^{(t,M)}$.
- 18:
$$v_{i,c}^{(t+1)} = v_c^{(t)} - \eta \lambda_1 \nabla_{v_c} \left\| v_c^{(t)} - \hat{m}_i^{(c,t)} \right\|^2 - \eta \lambda_2 \sum_{c \in \mathcal{Y}_i} \frac{n_i}{|I_{t+1}^{(i,m)}|} \sum_{j \in I_{t+1}^{(i,m)}} \nabla_{v_c} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,M)}]}^{(i)} \left[Z_c^{(j,t,M)} \right] \right).$$
- 19:
$$L_{i,c}^{(t+1)} = L_c^{(t)} - \eta \lambda_1 \nabla_{L_c} \mathfrak{B}^2 \left(L_c^{(t)} [L_c^{(t)}]^\top, \hat{\Sigma}_i^{(c,t)} \right) - \eta \lambda_2 \sum_{c \in \mathcal{Y}_i} \frac{n_i}{|I_{t+1}^{(i,m)}|} \sum_{j \in I_{t+1}^{(i,m)}} \nabla_{L_c} \ell \left(y_i^{(j)}, g_{[\alpha^{(t)}, \beta_i^{(t,M)}]}^{(i)} \left[Z_c^{(j,t,M)} \right] \right).$$
- 20: **end for**
- 21: *// Communication with the server*
- 22: Send $\alpha_i^{(t+1)}, v_{i,1:C}^{(t+1)}$ and $L_{i,1:C}^{(t+1)}$ to central server.
- 23: **end for**
- 24: *// Averaging global parameters*
- 25:
$$\alpha^{(t+1)} = \frac{b}{|A_{t+1}|} \sum_{i \in A_{t+1}} w_i \alpha_i^{(t+1)}.$$
- 26: **for** $c = 1$ **to** C **do**
- 27:
$$v_c^{(t+1)} = (b/|A_{t+1}|) \sum_{i \in A_{t+1}} \omega_i v_{i,c}^{(t+1)}.$$
- 28:
$$L_c^{(t+1)} = (b/|A_{t+1}|) \sum_{i \in A_{t+1}} \omega_i L_{i,c}^{(t+1)}$$
 and set $\Sigma_c^{(t+1)} = L_c^{(t+1)} [L_c^{(t+1)}]^\top$.
- 29: **end for**
- 30: **end for**

Ensure: parameters $\alpha^{(T)}, \mu_{1:C}^{(T)}, \phi_{1:b}^{(T,0)}, \beta_{1:b}^{(T,0)}$.

S1.4. Additional Algorithmic Insights

Scalability. When the number of classes C is large, both local computation and communication costs are increased. In this setting, we propose to partition all the classes into $C_{\text{meta}} \ll C$ meta-classes and consider reference measures $\{\mu_c\}_{c \in [C_{\text{meta}}]}$ associated to these meta-classes. As an example, if we are considering a dataset made of features associated to animals, the meta-class refers to an animal (e.g. a dog) and the class refers to a specific breed (e.g. golden retriever).

Privacy Consideration. As other standard (personalised) FL algorithms, **FLIC** satisfies first-order privacy guarantees by not allowing raw data exchanges but rather exchanges of local Gaussian statistics. Note that **FLIC** stands for a post-hoc

approach and can be combined with other privacy/confidentiality techniques such as differential privacy (Dwork and Roth, 2014), secure aggregation via secure multi-party computation (Chen et al., 2022) or trusted execution environments (Mo et al., 2021).

Inference on New Clients. When a client who has not participated to the training procedure appears, there is no need to re-launch a potentially costly federated learning procedure. Instead, the server sends the shared parameters $\{\alpha^{(T)}, \mu_{1:C}^{(T)}\}$ to the new client and the latter only needs to learn the local parameters $\{\phi_i, \beta_i\}$.

S2. Proof of Theorem 1

This section aims at proving Theorem 1 in the main paper. To this end, we first provide in Appendix S2.1 a closed-form expression for the estimated embedded features based on the features embedded by the oracle. Then, in Appendix S2.3, we show technical lemmata that will be used in Appendix S2.2 to show Theorem 1.

To prove our results, we consider the following set of assumptions.

- H1.** (i) For any $i \in [b]$, $j \in [n_i]$, ground-truth embedded features $\phi_i^*(x_i^{(j)})$ are distributed according to $N(0_k, I_k)$.
(ii) Ground-truth model parameters satisfy $\|\beta_i^*\|_2 = \sqrt{d}$ for $i \in [b]$ and A^* has orthonormal columns.
(iii) For any $t \in \{0, \dots, T-1\}$, $|A_{t+1}| = \lfloor rb \rfloor$ with $1 \leq \lfloor rb \rfloor \leq b$, and if we select $\lfloor rb \rfloor$ clients, their ground-truth head parameters $\{\beta_i^*\}_{i \in A_{t+1}}$ span \mathbb{R}^d .
(iv) In (2) in the main paper, $\ell(\cdot, \cdot)$ is the ℓ_2 norm, $\omega_i = 1/b$, $\theta_i = [A, \beta_i]$ and $g_{\theta_i}^{(i)}(x) = (A\beta_i)^\top x$ for $x \in \mathbb{R}^k$.

S2.1. Estimation of the Feature Transformation Functions

As in Section 4 in the main paper, we assume that $x_i^{(j)} \sim N(m_i, \Sigma_i)$ with $m_i \in \mathbb{R}^{k_i}$ and $\Sigma_i \in \mathbb{R}^{k_i \times k_i}$ for $i \in [b]$, $j \in [n_i]$. In addition, we consider that the continuous scalar labels are generated via the oracle model $y_i^{(j)} = (A^* \beta_i^*)^\top \phi_i^*(x_i^{(j)})$ where $A^* \in \mathbb{R}^{k \times d}$, $\beta_i^* \in \mathbb{R}^d$ and $\phi_i^*(\cdot)$ are ground-truth parameters and feature transformation function, respectively. Under **H1-(i)**, the oracle feature transformation functions $\{\phi_i^*\}_{i \in [b]}$ are assumed to map k_i -dimensional Gaussian distributions $N(m_i, \Sigma_i)$ to a common k -dimension Gaussian $N(0_k, I_k)$. As shown in Delon et al. (2022, Theorem 4.1), there exist closed-form expressions for $\{\phi_i^*\}_{i \in [b]}$, which can be shown to stand for solutions of a Gromov-Wasserstein problem restricted to Gaussian transport plans. More precisely, these oracle feature transformation stand for affine maps and are of the form, for any $i \in [b]$,

$$\phi_i^*(x_i^{(j)}) = \left[\tilde{I}_k^{(i,*)} (D_i^{(k)})^{-1/2} \quad 0_{k, k_i - k} \right] \left(x_i^{(j)} - m_i \right),$$

where $\tilde{I}_k^{(i,*)} = \text{diag}_k(\pm 1)$ is a k -dimensional diagonal matrix with diagonal elements in $\{-1, 1\}$, $\Sigma_i = P_i D_i P_i^\top$ is the diagonalisation of Σ_i and $D_i^{(k)}$ stands for the restriction of D_i to the first k components. In the sequel, we assume that all oracle feature transformation functions share the same randomness, that is $\tilde{I}_k^{(i,*)} = \tilde{I}_k^* = \text{diag}_k(\pm 1)$.

For the sake of simplicity, we assume that we know the true latent distribution of $\phi_i^*(x_i^{(j)})$ and as such consider a pre-fixed reference latent distribution that equals the latter, that is $\mu = N(0_k, I_k)$. Since we know from Delon et al. (2022, Theorem 4.1) that there exist mappings between Gaussian distributions with supports associated to different metric spaces, we propose an estimate for the ground-truth feature transformation functions defined by for any $i \in [b]$,

$$\hat{\phi}_i(x_i^{(j)}) = \left[\tilde{I}_k (D_i^{(k)})^{-1/2} \quad 0_{k, k_i - k} \right] \left(x_i^{(j)} - m_i \right),$$

where $\tilde{I}_k = \text{diag}_k(\pm 1)$. By noting that $\tilde{I}_k = Q \tilde{I}_k^*$, where $Q \in \mathbb{R}^{k \times k}$ is a diagonal matrix of the form $\text{diag}_k(\pm 1)$, it follows that

$$\hat{\phi}_i(x_i^{(j)}) = Q \phi_i^*(x_i^{(j)}). \quad (\text{S1})$$

In Appendix S2.2, the equation (S1) will allow us to relate the ground-truth labels $y_i^{(j)} = (A^* \beta_i^*)^\top \phi_i^*(x_i^{(j)})$ with estimated predictions $\hat{y}_i^{(j)} = (A^{(T)} \beta_i^{(T)})^\top \hat{\phi}_i(x_i^{(j)})$ via Algorithm S3 starting from the same embedded features.

S2.2. Proof of Theorem 1

Algorithm S3 FLIC–FedRep for linear regression and Gaussian features

Require: step size η , number of outer iterations T , participation rate $r \in (0, 1)$, diagonalizations $\Sigma_i = P_i D_i P_i^\top$ sorting eigenvalues in decreasing order.

- 1: // *Estimation of embedded features*
 - 2: For each client $i \in [b]$, set $\hat{\phi}_i(x_i^{(j)}) = [\tilde{I}_k(D_i^{(k)})^{-1/2} \quad 0_{k, k_i - k}] (x_i^{(j)} - m_i)$.
 - 3: // *Initialisation $A^{(0)}$*
 - 4: Each client $i \in [b]$ sends $Z_i = (1/n_i) \sum_{j=1}^{n_i} (y_i^{(j)})^2 \hat{\phi}_i(x_i^{(j)}) [\hat{\phi}_i(x_i^{(j)})]^\top$ to the central server.
 - 5: The central server computes $UDU^\top \leftarrow \text{rank-}d \text{ SVD} \left((1/b) \sum_{i=1}^b Z_i \right)$.
 - 6: The central server initialises $A^{(0)} = U$.
 - 7: **for** $t = 0$ **to** $T - 1$ **do**
 - 8: Sample a set of A_{t+1} of active clients such that $|A_{t+1}| = \lfloor rb \rfloor$.
 - 9: **for** $i \in A_{t+1}$ **do**
 - 10: The central server sends $A^{(t)}$ to A_{t+1} .
 - 11: // *Update local parameters*
 - 12: $\beta_i^{(t+1)} = \arg \min_{\beta_i} \sum_{j=1}^{n_i} \left(y_i^{(j)} - \beta_i^\top [A^{(t)}]^\top \hat{\phi}_i(x_i^{(j)}) \right)^2$.
 - 13: // *Update global parameters*
 - 14: $A_i^{(t+1)} = A^{(t)} - \eta \nabla_A \sum_{j=1}^{n_i} \left(y_i^{(j)} - [\beta_i^{(t+1)}]^\top A^\top \hat{\phi}_i(x_i^{(j)}) \right)^2$.
 - 15: // *Communication with the server*
 - 16: Send $A_i^{(t+1)}$ to the central server.
 - 17: **end for**
 - 18: // *Averaging and orthogonalisation of global parameter*
 - 19: $\bar{A}^{(t+1)} = \frac{1}{\lfloor rb \rfloor} \sum_{i \in A_{t+1}} A_i^{(t+1)}$.
 - 20: $A^{(t+1)}, R^{(t+1)} \leftarrow \text{QR}(\bar{A}^{(t+1)})$.
 - 21: **end for**
- Ensure:** parameters $A^{(T)}, \beta_{1:b}^{(T)}$.

Let $B \in \mathbb{R}^{b \times d}$ the matrix having local model parameters $\{\beta_i\}_{i \in [b]}$ as columns and denote by $B_{A_{t+1}} \in \mathbb{R}^{\lfloor rb \rfloor \times d}$ its restriction to the row set defined by A_{t+1} where $|A_{t+1}| = \lfloor rb \rfloor$ for some $r \in (0, 1]$. For the sake of simplicity, we assume in the sequel that all clients have the same number of data points that is for any $i \in [b]$, $n_i = n$. For random batches of samples $\{(x_i^{(j)}, y_i^{(j)}), j \in [n]\}_{i \in [\lfloor rb \rfloor]}$, we define similarly to Collins et al. (2021); Jain et al. (2013), the random linear operator $\mathcal{A} : \mathbb{R}^{\lfloor rb \rfloor \times d} \rightarrow \mathbb{R}^{\lfloor rb \rfloor n}$ for any $M \in \mathbb{R}^{\lfloor rb \rfloor \times d}$ as $\mathcal{A}(M) = [\langle e_i(\phi_i^*(x_i^{(j)}))^\top, M \rangle]_{1 \leq i \leq \lfloor rb \rfloor, 1 \leq j \leq n}$, where e_i stands for the i -th standard vector of $\mathbb{R}^{\lfloor rb \rfloor}$. Using these notations, it follows from Algorithm S3 that for any $t \in \{0, \dots, T-1\}$, the model parameters $\theta_i^{(t+1)} = [A^{(t+1)}, \beta_i^{(t+1)}]$ are computed as follows:

$$B_{A_{t+1}}^{(t+1)} = \arg \min_{B_{A_{t+1}}} \frac{1}{\lfloor rb \rfloor n} \left\| \mathcal{A}^{(t+1)} \left(B_{A_{t+1}}^* [A^*]^\top - B_{A_{t+1}} [A^{(t)}]^\top Q \right) \right\|^2, \quad (\text{S2})$$

$$\begin{aligned} \bar{A}^{(t+1)} &= \bar{A}^{(t)} - \frac{\eta}{\lfloor rb \rfloor n} \left[(\mathcal{A}^{(t+1)})^\dagger \mathcal{A}^{(t+1)} \left(B_{A_{t+1}}^* [A^*]^\top - B_{A_{t+1}}^{(t+1)} [A^{(t)}]^\top Q \right) \right]^\top Q B_{A_{t+1}}^{(t+1)}, \\ A^{(t+1)}, R^{(t+1)} &\leftarrow \text{QR}(\bar{A}^{(t+1)}), \end{aligned} \quad (\text{S3})$$

where $\mathcal{A}^{(t+1)}$ stands for a specific instance of \mathcal{A} depending on the random subset of active clients available at each round and \mathcal{A}^\dagger is the adjoint operator of \mathcal{A} defined by $\mathcal{A}^\dagger(M) = \sum_{i \in [\lfloor rb \rfloor]} \sum_{j=1}^n [\langle e_i(\phi_i^*(x_i^{(j)}))^\top, M \rangle] e_i(\phi_i^*(x_i^{(j)}))$.

The update in (S2) admits a closed-form expression as shown in the following lemma.

Lemma S1. For any $t \in \dots, 0, \dots, T-1$, we have

$$B_{A_{t+1}}^{(t+1)} = B_{A_{t+1}}^* [A^*]^\top Q A^{(t)} - F^{(t)},$$

where $F^{(t)}$ is defined in (S12), $A^{(t)}$ is defined in (S3) and $B_{A_t}^{(t)}$ is defined in (S2).

Proof. The proof follows from the same steps as in Collins et al. (2021, Proof of Lemma 1) using (S2). \square

Under **H1**, we have the following non-asymptotic convergence result.

Theorem S3. *Assume **H1**. Then, for any $x_i \in \mathbb{R}^{k_i}$, we have $\hat{\phi}_i(x_i) = Q\phi_i^*(x_i)$ where $Q \in \mathbb{R}^{k \times k}$ is of the form $\text{diag}_k(\pm 1)$. Define $E_0 = \text{dist}(A^{(0)}, QA^*)$. Assume that $n \geq c(d^3 \log(\lfloor rb \rfloor))/E_0^2 + d^2 k/(E_0^2 \lfloor rb \rfloor)$ for some absolute constant $c > 0$. Then, for any $t \in \{0, \dots, T-1\}$, $\eta \leq 1/(4\bar{\sigma}_{\max, \star}^2)$ and with high probability at least $1 - e^{-110k} - e^{-110d^2 \log(\lfloor rb \rfloor)}$, we have*

$$\text{dist}(A^{(t+1)}, QA^*) \leq (1 - \kappa)^{(t+1)/2} \text{dist}(A^{(0)}, QA^*),$$

where $A^{(t)}$ is computed via Algorithm S3, dist denotes the principal angle distance and $\kappa \in (0, 1)$ is defined as

$$\kappa = 1 - \eta E_0 \bar{\sigma}_{\min, \star}^2 / 2.$$

Proof. The proof follows first by plugging Lemma S3, Lemma S8 and Lemma S9 into Lemma S2. Then, we use the same technical arguments and steps as in Collins et al. (2021, Proof of Lemma 6). \square

S2.3. Technical Lemmata

In this section, we provide a set of useful technical lemmata to prove our main result in Appendix S2.2.

Notations. We begin by defining some notations that will be used in the sequel. For any $t \in \{0, \dots, T-1\}$, we define

$$Z^{(t+1)} = B_{A_{t+1}}^{(t+1)} [A^{(t)}]^\top Q - B_{A_{t+1}}^* [A^*]^\top. \quad (\text{S4})$$

In addition, let

$$G^{(t)} = \begin{bmatrix} G_{11}^{(t)} & \cdots & G_{1d}^{(t)} \\ \vdots & \ddots & \vdots \\ G_{d1}^{(t)} & \cdots & G_{dd}^{(t)} \end{bmatrix}, C^{(t)} = \begin{bmatrix} C_{11}^{(t)} & \cdots & C_{1d}^{(t)} \\ \vdots & \ddots & \vdots \\ C_{d1}^{(t)} & \cdots & C_{dd}^{(t)} \end{bmatrix}, D^{(t)} = \begin{bmatrix} D_{11}^{(t)} & \cdots & D_{1d}^{(t)} \\ \vdots & \ddots & \vdots \\ D_{d1}^{(t)} & \cdots & D_{dd}^{(t)} \end{bmatrix},$$

where for $p, q \in [d]$,

$$G_{pq}^{(t)} = \frac{1}{n} \sum_{i \in A_{t+1}} \sum_{j=1}^n e_i \left(\phi_i^*(x_i^{(j)}) \right)^\top Q a_p^{(t)} [a_q^{(t)}]^\top Q \phi_i^*(x_i^{(j)}) e_i^\top, \quad (\text{S5})$$

$$C_{pq}^{(t)} = \frac{1}{n} \sum_{i \in A_{t+1}} \sum_{j=1}^n e_i \left(\phi_i^*(x_i^{(j)}) \right)^\top Q a_p^{(t)} [a_q^*]^\top Q \phi_i^*(x_i^{(j)}) e_i^\top, \quad (\text{S6})$$

$$D_{pq}^{(t)} = \langle a_p^{(t)}, a_q^* \rangle \mathbf{I}_{\lfloor rb \rfloor}, \quad (\text{S7})$$

with $a_p^{(t)} \in \mathbb{R}^k$ standing for the p -th column of $A^{(t)} \in \mathbb{R}^{k \times d}$; and $a_p^* \in \mathbb{R}^k$ standing for the p -th column of $A^* \in \mathbb{R}^{k \times d}$. Finally, we define for any $i \in A_{t+1}$,

$$\Pi^i = \frac{1}{n} \sum_{j=1}^n \phi_i^*(x_i^{(j)}) [\phi_i^*(x_i^{(j)})]^\top, \quad (\text{S8})$$

$$(G^{(t)})^i = [A^{(t)}]^\top Q \Pi^i Q A^{(t)}, \quad (\text{S9})$$

$$(C^{(t)})^i = [A^{(t)}]^\top Q \Pi^i Q A^*, \quad (\text{S10})$$

$$(D^{(t)})^i = [A^{(t)}]^\top Q A^*. \quad (\text{S11})$$

Using these notations, we also define $\tilde{\beta}^* = [(\beta_1^*)^\top, \dots, (\beta_d^*)^\top]^\top \in \mathbb{R}^{\lfloor rb \rfloor d}$ and

$$F^{(t)} = [([G^{(t)}]^{-1}(G^{(t)}D^{(t)} - C^{(t)})\tilde{\beta}^*)_1, \dots, ([G^{(t)}]^{-1}(G^{(t)}D^{(t)} - C^{(t)})\tilde{\beta}^*)_d]. \quad (\text{S12})$$

Technical results. To prove our main result in Theorem S3, we begin by providing a first upper bound on the quantity of interest namely $\text{dist}(A^{(t+1)}, QA^*)$. This is the purpose of the next lemma.

Lemma S2. For any $t \in \{0, \dots, T-1\}$ and $\eta > 0$, we have

$$\text{dist}(A^{(t+1)}, QA^*) \leq C_1 + C_2, ,$$

where

$$C_1 = \left\| [A_\perp^*]^\top QA^{(t)} \left(I_d - \frac{\eta}{\lfloor rb \rfloor} [B_{\mathcal{A}_{t+1}}^{(t+1)}]^\top B_{\mathcal{A}_{t+1}}^{(t+1)} \right) \right\|_2 \left\| (R^{(t+1)})^{-1} \right\|_2, \quad (\text{S13})$$

$$C_2 = \frac{\eta}{\lfloor rb \rfloor} \left\| \left(\frac{1}{n} [A_\perp^*]^\top (QA^{(t+1)})^\dagger \mathcal{A}^{(t+1)} (Z^{(t+1)}) Q - Z^{(t+1)} \right)^\top B_{\mathcal{A}_{t+1}}^{(t+1)} \right\|_2 \left\| (R^{(t+1)})^{-1} \right\|_2, \quad (\text{S14})$$

where $A^{(t)}$ is defined in (S3), $B_{\mathcal{A}_t}^{(t)}$ is defined in (S2), $Z^{(t)}$ is defined in (S4) and $R^{(t)}$ comes from the QR factorisation of $\bar{A}^{(t)}$, see step 20 in Algorithm S3.

Proof. The proof follows from the same steps as in Collins et al. (2021, Proof of Lemma 6) and by noting that $\text{dist}(A^{(t)}, QA^*) = \text{dist}(QA^{(t)}, A^*)$ for $t \in \{0, \dots, T-1\}$. \square

We now have to control the terms C_1 and C_2 . For the sake of clarity, we split technical results aiming to upper bound of C_1 and C_2 in two different paragraphs.

Control of C_1 .

Lemma S3. Assume H1. Let $\delta_d = cd^{3/2}\sqrt{\log(\lfloor rb \rfloor)}/n^{1/2}$ for some absolute constant $c > 0$. Then, for any $t \in \{0, \dots, T-1\}$, with probability at least $1 - e^{-111k^2 \log(\lfloor rb \rfloor)}$, we have for $\delta_d \leq 1/2$ and $\eta \leq 1/(4\bar{\sigma}_{\max, \star}^2)$

$$C_1 \leq \left[\leq 1 - \eta \left(1 - \text{dist}(A^{(0)}, QA^*) \right) \bar{\sigma}_{\min, \star}^2 + 2\eta \frac{\delta_d}{1 - \delta_d} \bar{\sigma}_{\max}^2 \right] \text{dist}(A^{(t)}, QA^*) \left\| (R^{(t+1)})^{-1} \right\|_2,$$

where $\bar{\sigma}_{\min}^2, \bar{\sigma}_{\max}^2$ are defined in (S15)-(S16), C_1 is defined in (S13), $A^{(t)}$ is defined in (S3) and $R^{(t)}$ comes from the QR factorisation of $\bar{A}^{(t)}$, see step 20 in Algorithm S3.

Proof. Using Cauchy-Schwarz inequality, we have

$$\begin{aligned} C_1 &\leq \left\| (A_\perp^*)^\top QA^{(t)} \right\|_2 \left\| I_d - \frac{\eta}{\lfloor rb \rfloor} [B_{\mathcal{A}_{t+1}}^{(t+1)}]^\top B_{\mathcal{A}_{t+1}}^{(t+1)} \right\|_2 \left\| (R^{(t+1)})^{-1} \right\|_2 \\ &= \text{dist}(A^{(t)}, QA^*) \left\| I_d - \frac{\eta}{\lfloor rb \rfloor} [B_{\mathcal{A}_{t+1}}^{(t+1)}]^\top B_{\mathcal{A}_{t+1}}^{(t+1)} \right\|_2 \left\| (R^{(t+1)})^{-1} \right\|_2. \end{aligned}$$

Define the following minimum and maximum singular values:

$$\bar{\sigma}_{\min, \star}^2 = \min_{\mathcal{A} \subseteq [b], |\mathcal{A}| = \lfloor rb \rfloor} \sigma_{\min} \left(\frac{1}{\sqrt{\lfloor rb \rfloor}} B_{\mathcal{A}}^* \right) \quad (\text{S15})$$

$$\bar{\sigma}_{\max, \star}^2 = \min_{\mathcal{A} \subseteq [b], |\mathcal{A}| = \lfloor rb \rfloor} \sigma_{\max} \left(\frac{1}{\sqrt{\lfloor rb \rfloor}} B_{\mathcal{A}}^* \right). \quad (\text{S16})$$

Using Collins et al. (2021, Proof of Lemma 6, equations (67)-(68)), we have for $\delta_d \leq 1/2$ where δ_d is defined in Lemma S4 and $\eta \leq 1/(4\bar{\sigma}_{\max, \star}^2)$,

$$\left\| I_d - \frac{\eta}{\lfloor rb \rfloor} [B_{\mathcal{A}_{t+1}}^{(t+1)}]^\top B_{\mathcal{A}_{t+1}}^{(t+1)} \right\|_2 \leq 1 - \eta \left(1 - \text{dist}(A^{(0)}, QA^*) \right) \bar{\sigma}_{\min, \star}^2 + 2\eta \frac{\delta_d}{1 - \delta_d} \bar{\sigma}_{\max, \star}^2,$$

with probability at least $1 - e^{-111k^2 \log(\lfloor rb \rfloor)}$. The proof is concluded by combining the two previous bounds. \square

Control of C_2 . We begin by showing four intermediary results gathered in the next four lemmata.

Lemma S4. Assume **H1**. Let $\delta_d = cd^{3/2}\sqrt{\log(\lceil rb \rceil)}/n^{1/2}$ for some absolute constant $c > 0$. Then, for any $t \in \{0, \dots, T-1\}$, with probability at least $1 - e^{-111k^3 \log(\lceil rb \rceil)}$, we have

$$\left\| [G^{(t)}]^{-1} \right\|_2 \leq \frac{1}{1 - \delta_d},$$

where $G^{(t)}$ is defined in (S5).

Proof. The proof stands as a straightforward extension of Collins et al. (2021, Proof of Lemma 2) by noting that the random variable $Q\phi_i^*(x_i^{(j)}) = \hat{\phi}_i(x_i^{(j)})$ is sub-Gaussian under **H1**-(i); and as such is omitted. \square

Lemma S5. Assume **H1**. Let $\delta_d = cd^{3/2}\sqrt{\log(\lceil rb \rceil)}/n^{1/2}$ for some absolute constant $c > 0$. Then, for any $t \in \{0, \dots, T-1\}$, with probability at least $1 - e^{-111k^2 \log(\lceil rb \rceil)}$, we have

$$\left\| (G^{(t)}D^{(t)} - C^{(t)})B_{A_t}^* \right\|_2 \leq \delta_d \|B_{A_t}^*\|_2 \text{dist}(A^{(t)}, QA^*),$$

where $G^{(t)}$ is defined in (S5), $D^{(t)}$ is defined in (S7), $C^{(t)}$ is defined in (S6) and $A^{(t)}$ in (S3).

Proof. Without loss of generality and to ease notation, we remove the superscript (t) in the proof and re-index the indexes of clients in A_{t+1} . Let $H = GD - C$. From (S8), (S9), (S10) and (S11), it follows, for any $i \in \llbracket rb \rrbracket$, that

$$H^i = G^i D^i - C^i = A^\top Q \Pi^i Q (AA^\top - I_k) Q A^*.$$

Hence, by using the definition of H , we have

$$\|(GD - C)\beta^*\|_2^2 = \sum_{i=1}^{\lceil rb \rceil} \|H^i \beta_i^*\|_2^2 \leq \sum_{i=1}^{\lceil rb \rceil} \|H^i\|_2^2 \|\beta_i^*\|_2^2 \leq \frac{d}{\lceil rb \rceil} \|B^*\|_2^2 \sum_{i=1}^{\lceil rb \rceil} \|H^i\|_2^2,$$

where the last inequality follows almost surely from **H1**-(iii). As in Collins et al. (2021, Proof of Lemma 3), we then define for any $j \in \llbracket n \rrbracket$, the vectors

$$\begin{aligned} u_i^{(j)} &= \frac{1}{\sqrt{n}} [A^*]^\top (AA^\top - I_k) Q \phi_i^*(x_i^{(j)}), \\ v_i^{(j)} &= \frac{1}{\sqrt{n}} A^\top Q \phi_i^*(x_i^{(j)}). \end{aligned}$$

Let S^{d-1} denotes the d -dimensional unit spheres. Then, by Vershynin (2018, Corollary 4.2.13), we can define \mathcal{N}_d , the $1/4$ -net over S^{d-1} such that $|\mathcal{N}_d| \leq 9^d$. Therefore, by using Vershynin (2018, Equation (4.13)), we have

$$\|H^i\|_2^2 \leq 2 \max_{z, y \in \mathcal{N}_d} \sum_{j=1}^n \langle z, u_i^{(j)} \rangle \langle v_i^{(j)}, y \rangle.$$

Since $\phi_i^*(x_i^{(j)})$ is a standard Gaussian vector, it is sub-Gaussian and therefore $\langle z, u_i^{(j)} \rangle$ and $\langle v_i^{(j)}, y \rangle$ are sub-Gaussian with norms $\|\frac{1}{\sqrt{n}} [A^*]^\top (AA^\top - I_k) Q\|_2 = (1/\sqrt{n}) \text{dist}(A, QA^*)$ and $(1/\sqrt{n})$, respectively. In addition, we have

$$\begin{aligned} \mathbb{E} \left[\langle z, u_i^{(j)} \rangle \langle v_i^{(j)}, y \rangle \right] &= \frac{1}{n} \mathbb{E} \left[z^\top \frac{1}{\sqrt{n}} [A^*]^\top (AA^\top - I_k) Q \phi_i^*(x_i^{(j)}) [\phi_i^*(x_i^{(j)})]^\top Q A y \right] \\ &= \frac{1}{n} z^\top \frac{1}{\sqrt{n}} [A^*]^\top (AA^\top - I_k) A y \\ &= 0, \end{aligned}$$

where we have used the fact that $\mathbb{E}[\phi_i^*(x_i^{(j)})[\phi_i^*(x_i^{(j)})]^\top] = 1$, $Q^2 = I_k$ and $(AA^\top - I_k)A = 0$. The rest of the proof is concluded by using the Bernstein inequality by following directly the steps detailed in Collins et al. (2021, Proof of Lemma 3, see equations (35) to (39)). \square

Lemma S6. Assume **H1**. Let $\delta_d = cd^{3/2} \sqrt{\log(\lceil rb \rceil)} / n^{1/2}$ for some absolute constant $c > 0$. Then, for any $t \in [T]$, with probability at least $1 - e^{-111k^2 \log(\lceil rb \rceil)}$, we have

$$\|F^{(t)}\|_F \leq \frac{\delta_d}{1 - \delta_d} \|B_{A_t}^*\|_2 \text{dist}(A^{(t)}, QA^*),$$

where $F^{(t)}$ is defined in (S12) and $A^{(t)}$ in (S3).

Proof. By the Cauchy-Schwarz inequality, we have $\|F^{(t)}\|_F = \|[G^{(t)}]^{-1}(G^{(t)}D^{(t)} - C^{(t)})B_{A_t}^*\|_2 \leq \delta_d \|B_{A_t}^*\|_2 \leq \|[G^{(t)}]^{-1}\|_2 \|(G^{(t)}D^{(t)} - C^{(t)})B_{A_t}^*\|_2 \leq \delta_d \|B_{A_t}^*\|_2$. The proof is concluded by combining the upper bounds given in Lemma S4 and Lemma S5. \square

Lemma S7. Assume **H1** and let $\delta'_d = cd\sqrt{k}/\sqrt{\lceil rb \rceil}n$ for some absolute positive constant c . For any $t \in [T]$ and whenever $\delta'_d \leq d$, we have with probability at least $1 - e^{-110k} - e^{-110d^2 \log(\lceil rb \rceil)}$

$$\frac{1}{\lceil rb \rceil} \left\| \left(\frac{1}{n} Q(A^{(t)})^\dagger A^{(t)} (Z^{(t)}) Q - Z^{(t)} \right)^\top B_{A_t}^{(t)} \right\|_2 \leq \delta'_d \text{dist}(A^{(t)}, QA^*),$$

where $B_{A_t}^{(t)}$ is defined in (S2) and $Z^{(t)}$ is defined in (S4).

Proof. Let $t \in [T]$. Note that we have

$$\left(\frac{1}{n} Q(A^{(t)})^\dagger A^{(t)} (Z^{(t)}) Q - Z^{(t)} \right)^\top B_{A_t}^{(t)} = \frac{1}{n} \sum_{i \in A_t} \sum_{j=1}^m \langle Q\phi_i^*(x_i^{(j)}), z_i^{(t)} \rangle Q\phi_i^*(x_i^{(j)}) [\beta_i^{(t)}]^\top - z_i^{(t)} [\beta_i^{(t)}]^\top.$$

Let \mathcal{S}^{k-1} and \mathcal{S}^{d-1} denote the k -dimensional and d -dimensional unit spheres, respectively. Then, by Vershynin (2018, Corollary 4.2.13), we can define \mathcal{N}_k and \mathcal{N}_d , $1/4$ -nets over \mathcal{S}^{k-1} and \mathcal{S}^{d-1} , respectively, such that $|\mathcal{N}_k| \leq 9^k$ and $|\mathcal{N}_d| \leq 9^d$. Therefore, by using Vershynin (2018, Equation (4.13)), we have

$$\begin{aligned} & \left\| \left(\frac{1}{n} Q(A^{(t)})^\dagger A^{(t)} (Z^{(t)}) Q - Z^{(t)} \right)^\top B_{A_t}^{(t)} \right\|_2^2 \\ &= 2 \max_{u \in \mathcal{N}_d, v \in \mathcal{N}_k} u^\top \left[\frac{1}{n} \sum_{i \in A_t} \sum_{j=1}^m \langle Q\phi_i^*(x_i^{(j)}), z_i^{(t)} \rangle Q\phi_i^*(x_i^{(j)}) [\beta_i^{(t)}]^\top - z_i^{(t)} [\beta_i^{(t)}]^\top \right] v \\ &= 2 \max_{u \in \mathcal{N}_d, v \in \mathcal{N}_k} \frac{1}{n} \sum_{i \in A_t} \sum_{j=1}^m \langle Q\phi_i^*(x_i^{(j)}), z_i^{(t)} \rangle \langle u, Q\phi_i^*(x_i^{(j)}) \rangle \langle \beta_i^{(t)}, v \rangle - \langle u, z_i^{(t)} \rangle \langle \beta_i^{(t)}, v \rangle. \end{aligned} \quad (\text{S17})$$

In order to control (S17) using Bernstein inequality as in Lemma S5, we need to characterise, in particular, the sub-Gaussianity of $\langle u, z_i^{(t)} \rangle$ and $\langle \beta_i^{(t)}, v \rangle$ which require a bound on $\|z_i^{(t)}\|$ and $\|\beta_i^{(t)}\|$, respectively. From Lemma S1, we have $[\beta_i^{(t)}]^\top = (\beta_i^*)^\top (A^*)^\top A^{(t)} - (z_i^{(t)})^\top$ which leads to

$$\begin{aligned} \|z_i^{(t)}\|^2 &= \left\| QA^{(t)}(A^{(t)})^\top QA^*\beta_i^* - QA^{(t)}f_i^{(t)} - A^*\beta_i^* \right\|_2^2 \\ &= \left\| (QA^{(t)}(A^{(t)})^\top Q - I_d)A^*\beta_i^* - QA^{(t)}f_i^{(t)} \right\|_2^2 \\ &\leq 2 \left\| (QA^{(t)}(A^{(t)})^\top Q - I_d)A^* \right\|_2^2 \|\beta_i^*\|^2 + 2 \|f_i^{(t)}\|^2 \\ &\leq 2d \text{dist}^2(A^{(t)}, QA^*) + 2 \|f_i^{(t)}\|^2. \end{aligned}$$

Using (S12) and the Cauchy-Schwarz inequality, we have

$$\|f_i^{(t)}\|^2 = \left\| [G^{i,(t)}]^{-1}(G^{i,(t)}D^{i,(t)} - C^{i,(t)})\beta_i^* \right\|_2^2$$

$$\begin{aligned}
&\leq \left\| [G^{i,(t)}]^{-1} \right\|_2^2 \left\| G^{i,(t)} D^{i,(t)} - C^{i,(t)} \right\|_2^2 \|\beta_i^*\|^2 \\
&\leq d \left\| [G^{i,(t)}]^{-1} \right\|_2^2 \left\| G^{i,(t)} D^{i,(t)} - C^{i,(t)} \right\|_2^2,
\end{aligned} \tag{S18}$$

where the last inequality follows from **H1**-(ii).

Using Lemma **S4** and Lemma **S5** and similarly to Collins et al. (2021, Equation (45)), it follows for any $i \in \mathbf{A}_t$ that

$$\left\| z_i^{(t)} \right\|_2^2 \leq 4d \operatorname{dist}(A^{(t)}, QA^*),$$

with probability at least $1 - e^{-110d^2 \log(\lfloor rb \rfloor)}$.

Similarly, using Lemma **S1** and (S18), we have with probability at least $1 - e^{-110d^2 \log(\lfloor rb \rfloor)}$ and for any $i \in \mathbf{A}_t$, that

$$\left\| \beta_i^{(t)} \right\|^2 \leq 2 \left\| [A^{(t)}]^\top QA^* \beta_i^* \right\|^2 + 2 \left\| f_i^{(t)} \right\|^2 \leq 4d.$$

Besides, note we have

$$\mathbb{E} \left[\langle Q\phi_i^*(x_i^{(j)}), z_i^{(t)} \rangle \langle u, Q\phi_i^*(x_i^{(j)}) \rangle \langle \beta_i^{(t)}, v \rangle \right] = \langle u, z_i^{(t)} \rangle \langle \beta_i^{(t)}, v \rangle.$$

The proof is then concluded by applying the Bernstein inequality following the same steps as in the final steps of Collins et al. (2021, Proof of Lemma 5). \square

We are now ready to control C_2 .

Lemma S8. Assume **H1** and let $\delta'_d = cd\sqrt{k}/\sqrt{\lfloor rb \rfloor n}$ for some absolute positive constant c . For any $t \in \{0, \dots, T-1\}$, $\eta > 0$ and whenever $\delta'_d \leq d$, we have with probability at least $1 - e^{-110k} - e^{-110d^2 \log(\lfloor rb \rfloor)}$

$$C_2 \leq \eta \delta'_d \operatorname{dist}(A^{(t)}, QA^*) \left\| \left(R^{(t+1)} \right)^{-1} \right\|_2,$$

where C_2 is defined in (S14), $A^{(t)}$ is defined in (S3) and $R^{(t)}$ comes from the QR factorisation of $\bar{A}^{(t)}$, see step 20 in Algorithm **S3**.

Proof. Let $t \in \{0, \dots, T-1\}$ and $\eta > 0$. Then, whenever $\delta'_d \leq d$, we have with probability at least $1 - e^{-110k} - e^{-110d^2 \log(\lfloor rb \rfloor)}$, we have

$$\begin{aligned}
C_2 &= \frac{\eta}{\lfloor rb \rfloor} \left\| \left(\frac{1}{n} [A_{\perp}^*]^\top (QA^{(t+1)})^\dagger \mathcal{A}^{(t+1)} \left(Z^{(t+1)} \right) Q - Z^{(t+1)} \right)^\top B_{\mathbf{A}_{t+1}}^{(t+1)} \right\|_2 \left\| \left(R^{(t+1)} \right)^{-1} \right\|_2 \\
&\leq \frac{\eta}{\lfloor rb \rfloor} \left\| \left(\frac{1}{n} (QA^{(t+1)})^\dagger \mathcal{A}^{(t+1)} \left(Z^{(t+1)} \right) Q - Z^{(t+1)} \right)^\top B_{\mathbf{A}_{t+1}}^{(t+1)} \right\|_2 \left\| \left(R^{(t+1)} \right)^{-1} \right\|_2 \\
&\leq \eta \delta'_d \operatorname{dist}(A^{(t)}, QA^*) \left\| \left(R^{(t+1)} \right)^{-1} \right\|_2,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality in the second inequality and Lemma **S7** for the last one. \square

Control of $\left\| \left(R^{(t+1)} \right)^{-1} \right\|_2$. To finalise our proof, it remains to bound $\left\| \left(R^{(t+1)} \right)^{-1} \right\|_2$. The associated result is depicted in the next lemma.

Lemma S9. Define $\bar{\delta}_d = \delta_d + \delta'_d$ where δ_d and δ'_d are defined in Lemma **S4** and Lemma **S5**, respectively. Assume **H1**. Then, we have with probability at least $1 - e^{-110k} - e^{-110d^2 \log(\lfloor rb \rfloor)}$,

$$\left\| \left(R^{(t+1)} \right)^{-1} \right\|_2 \leq \left(1 - 4\eta \frac{\bar{\delta}_d}{(1 - \bar{\delta}_d)^2} \bar{\sigma}_{\max, \star}^2 \right)^{-1/2}.$$

Proof. The proof follows from Collins et al. (2021, Proof of Lemma 6). \square

S3. Experimental Details

S3.1. Reference Distribution for Regression

For regression problem, our goal is to map all samples for all clients into a common latent subspace, in which some structural information about regression problem is preserved. As such, in order to reproduce the idea of using a Gaussian mixture model as an anchor distribution, we propose to use an infinite number of Gaussian mixtures in which the distribution of x associated to a response y is going to be mapped on a unit-variance Gaussian distribution whose mean depends uniquely on y . Formally, we define the anchor distribution as

$$\mu_y = \mathcal{N}(\mathbf{m}^{(y)}, \mathbf{I})$$

where $\mathbf{m}^{(y)}$ is a vector of dimension d that is uniquely defined. In practice, we consider as $\mathbf{m}^{(y)} = ya + (1 - y)b$ where a and b are two vectors in \mathbb{R}^d .

When training **FLIC**, this means that for a client i , we can compute $W_2^2(\mu_y, \nu_{\phi_i}^{(y)})$ based on the set of training samples $\{x, y\}$. In practice, if for a given batch of samples we have a single sample x , then the Wasserstein distance boils to $\|\phi_i(x) - \mathbf{m}^{(y)}\|_2^2$.

S3.2. Data Sets

We provide some details about the datasets we used for our numerical experiments

S3.2.1. TOY DATA SETS

The first toy dataset, denoted as *noisy features*, is a 20-class classification problem in which the features for a given class is obtained by sampling a Gaussian distribution of dimension 5, with random mean and Identity covariance matrix. For building the training set, we sample 2000 examples for each class and equally share those examples among clients who hold that class. Then, in order to generate some class imbalances on clients, we randomly subsample examples on all clients. For instance, with 100 clients and 2 classes per clients, this results in a problem with a total of about 16k samples with a minimal number of samples of 38 and a maximal one of 400. In order to get different dimensionality, we randomly append on each client dataset some Gaussian random noisy features with dimensionality varying from 1 to 10.

The second toy dataset, denoted as *linear mapping*, is a 20-class classification problem where each class-conditional distribution is Gaussian distribution of dimension 5, with random mean and random diagonal covariance matrix. As above, we generate 2000 samples per class and distribute and subsample them across clients in the similar way, leading to a total number of samples of about 15k. The dimensionality perturbation is modelled by a random (Gaussian)linear transformation that maps the original samples to a space which dimension goes up to 50.

S3.2.2. MNIST-USPS

We consider a digit classification problem with the original MNIST and USPS data sets which are respectively of dimension 28×28 and 16×16 and we assume that a client hosts either a subset of MNIST or USPS data set. We use the natural train/test split of those datasets and randomly share them across clients.

S3.2.3. TEXTCAPS DATA SET

The TextCaps data set (Sidorov et al., 2020) is an Image captioning dataset for which goal is to develop a model able to produce a text that captions the image. The dataset is composed of about 21k images and 110k captions and each image also comes with an object class. For our purpose, we have extracted pair of 14977 images and captions from the following four classes *Bottle*, *Car*, *Food* and *Book*. At each run, those pairs are separated in 80% train and 20% test sets. Examples from the TextCaps datasets are presented in Figure S5. Images and captions are represented by vectors by feeding them respectively to a pre-trained ResNet18 and a pretrained Bert, leading to vectors of size 512 and 768.

Each client holds either the image or the text representation of subset of examples and the associated vectors are randomly pruned of up to 10% coordinates. As such, all clients hold dataset with different dimensionality.

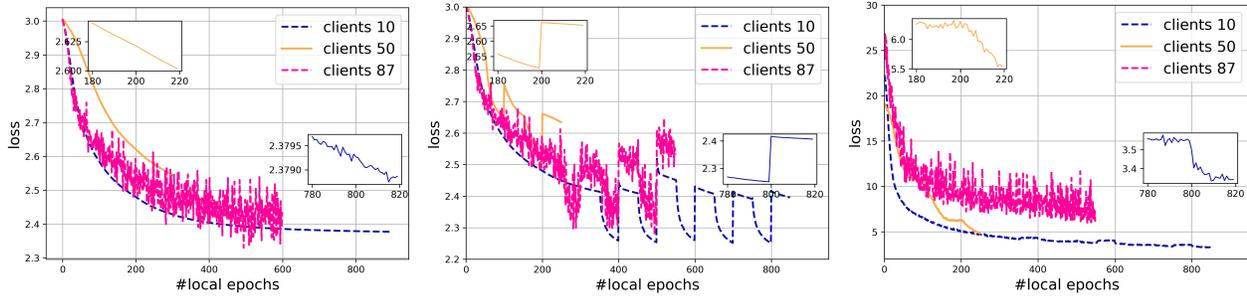


Figure S1. Evolution of the local loss curve of three different clients for three different learning situations. See text for details.

S3.3. Models and Learning Parameters

For the toy problems and the *TextCaps* data set, as a local transformation functions we used a fully connected neural network with one input, one hidden layer and one output layers. The number of units in hidden layer has been fixed to 64 and the dimension of latent space as been fixed to 64. ReLU activation has been applied after the input and hidden layers. For the digits dataset, we used a CNN model with 2 convolutional layers followed by a max-pooling layer and a sigmoid activation function. Once flattened, we have a one fully-connected layer and ReLU activation. The latent dimension is fixed to 64.

For all datasets, as for the local model g_{θ_i} , in order to be consistent with competitors, we first considered a single layer linear model implementing the local classifier as well as a model with one input layer (linear units followed by a LeakyReLU activation function) denoting the shared representation layer and an output linear layer.

For training, all methods use Adam with a default learning rate of 0.001 and a batch size of 100. Other hyperparameters have been set as follows. Unless specified, the regularization strength λ_1 and λ_2 have been fixed to 0.001. Local sample batch size is set to 100 and the participation rate r to 0.1. For all experiments, we have set the number of communication round T to 50 and the number of local epochs to respectively 10 and 100 for the real-world and toy datasets. For **FLIC**, as in FedRep those local epochs is followed by one epoch for representation learning. We have trained the local embedding functions for 100 local epochs and a batch size of 10 for toy datasets and TextCaps and while of 100 for MNIST-USPS. Reported accuracies are computed after local training for all clients.

S3.4. Ablating Loss Curves

In order to gain some understanding on the learning mechanism that involves local and global training respectively due to the local embedding functions, the local classifier and the global representation learning, we propose to look at local loss curves across different clients.

Here, we have considered the *linear mapping* toy dataset as those used in the toy problem analysis. However, the learning parameters we have chosen are different from those we have used to produce the results so as to highlight some specific features. The number of epochs (communication rounds) is set to 100 with a client activation ration of 0.1. Those local epochs are shared for either training the local parameters or the global ones (note that in our reference Algorithm 1, the global parameter is updated only once for each client) Those latter are trained starting after the 20-th communication round and in this case, the local epochs are equally shared between local and global parameter updates. Note that because of the randomness in the client selection at each epoch, the total number of local epochs is different from client to client. We have evaluated three learning situations and plotted the loss curves for each client.

- the local embedding functions and the global models are kept fixed, and only the local classifier is trained. Examples of loss curves for 3 clients are presented in the left plot of Figure S1. For this learning situation, there is no shared global parameters that are trained locally. Hence, the loss curve is typical of those obtained by stochastic gradient descent with a smooth transition, at multiple of 100 local epochs, when a given client is retrained after a communication rounds.
- the local embedding functions are kept fixed, while the classifier and global parameters are updated using half of the

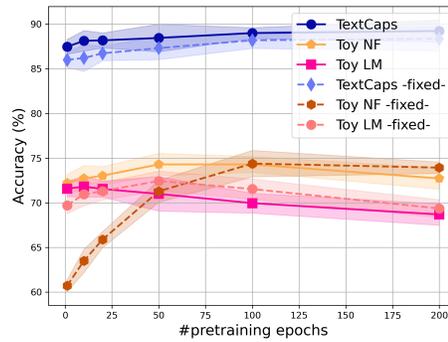


Figure S2. Impact of epochs used for pretraining ϕ_i on the model accuracy as well as updating those functions during the training. Results for three different datasets are reported. Plain and dashed curves are respectively related to local training with and without updates ϕ_i .

local epochs each. This situation is interesting and reported in middle plot in Figure S1. We can see that for some rounds of 100 local epochs, a strong drop in the loss occurs at starting at the 50th local epoch because the global parameters are being updated. Once the local update of a client is finished the global parameter is sent back to the server and all updates of global parameters are averaged by the server. When a client is selected again for local updates, it is served with a new global parameter (hence a new loss value) which causes the discontinuity in the loss curve at the beginning of each local update.

- all the part (local embedding functions, global parameter and the classifier) of the models are trained. Note at first that the loss value for those curves (right plot in Figure S1) is larger than for the two first most left plots as the Wasserstein distance to the anchor distribution is now taken into account and tends to dominate the loss. The loss curves are globally decreasing with larger drops in loss at the beginning of local epochs.

S3.5. On Importance of Alignment Pre-Training and Updates.

We have analyzed the impact of pretraining the local transformation functions and their updates during learning for fixed reference distribution. We have considered two learning situations : one in which they are updated during local training (as usual) and another one in they are kept fixed all along the training. We have chosen the setting with 100 users and have kept the same experimental settings as for the performance figure and made only varied the number of epochs considered for pretraining from 1 to 200. Results, averaged over 5 runs are shown in Figure S2. We remark that for the three datasets, increasing the number of epochs up to a certain number tends to increase performance, but overfitting may occur. The latter is mostly reflected in the *toy linear mapping* dataset for which 10 to 50 epochs is sufficient for good pretraining. Examples of how classes evolves during pretraining are illustrated in Figure 4, through *t-sne* projection. We also illustrate cases of how pretraining impact on the test set and may lead to overfitting as shown in the supplementary Figure S4.

S3.6. On the Impact of the Participation Rate

We have analyzed the effect of the participation rate of each client into our federated learning approach. Figure S3 reports the accuracies, averaged over 3 runs, of our approach for the toy datasets and the *TextCaps* problem with respect to the participation rate at each round. We can note that the proposed approach is rather robust to the participation rate but may rather suffer from overfitting due to overtraining of local models. On the left plot, performances, measured after the last communication round, for *TextCaps* is stable over participation rate while those performances tend to decrease for the *toy* problems. We associate these decrease to overfitting since when we report (see right plot) the best performance over communication rounds (and not the last one), they are stable for all problems. This suggests that number of local epochs may be dependent to the task on each client and the client participation rate.

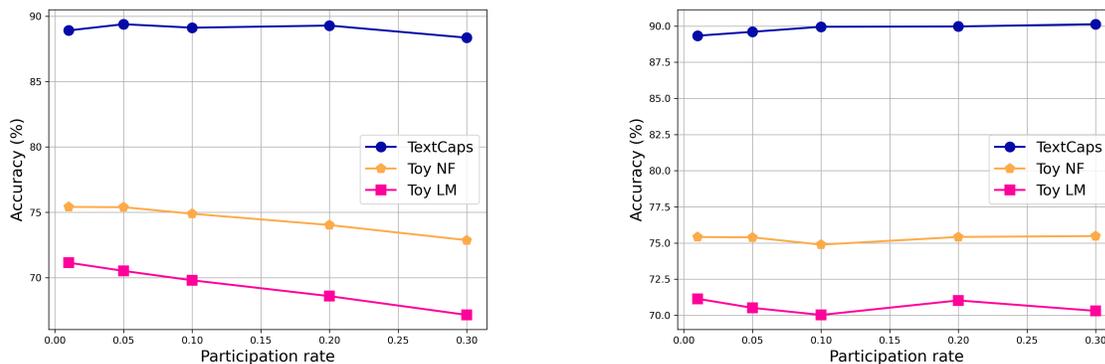


Figure S3. Evolution of the performance of our FLIC-Class algorithm with respects to the participation rate of clients, using the same experimental setting as in Figure 3. (left) evaluating performance after last communication rounds, (right) best performance across communication rounds.

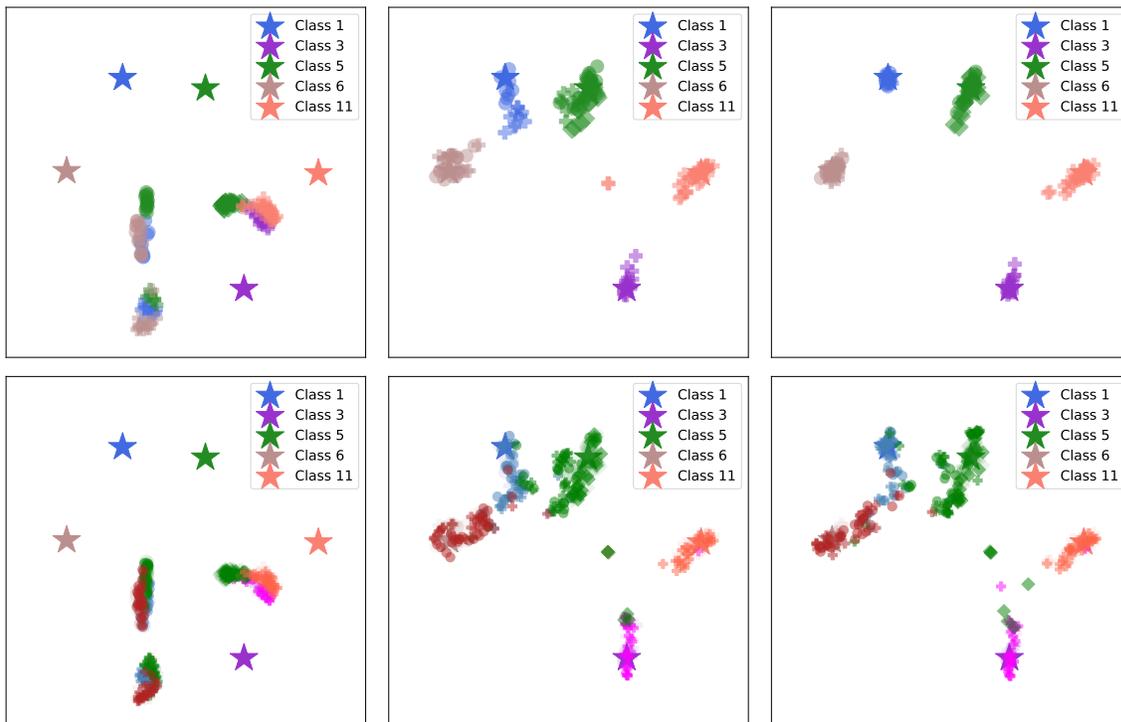


Figure S4. . 2D *t-sne* projection of 5 classes partially shared by 3 clients for the **toy linear mapping** dataset after learning the local embedding functions for (left) 10 epochs, (middle) 50 epochs, (right) 100 epochs. Original dimensions on clients vary from 5 to 50. Top row shows the projection the training set while bottom row plots show both training and test set. Star \star markers represent the projection of the mean of each class-conditional. The three different marker styles represent the different clients. Classes are denoted by colors and similar tones of color distinguish train and test sets. We see that each class from the training set from each client converges towards the mean of its anchor distribution, represented by the star marker. Interestingly, we also remark that unless convergence is reached, empirical class-conditional distributions on each clients are not equal making necessary the learning of a joint representation. From the bottom plots, we can understand that distribution alignment impacts mostly the training set but this alignment does not always generalize properly to the test sets.

A pan sits on a hob with a lid bearing the logo Hamilton Beach Stay n Go



Deciding on whether to drink spring water or a 7UP.



The bookshelf consists of about 20 different types of books.



The yellow convertible is on display somewhere in California.



Figure S5. Examples of some TextCaps pairs of image/caption from the 4 classes we considered of (top-left) Food, (top-right) Bottle, (bottom-left) Book (bottom-right) Car. We can see how difficult some examples can be, especially from the caption point of view since few hint about the class is provided by the text.