Leveraging Invariance in non-i.i.d Federated Learning

Abstract

Popular gradient descent learning strategies rely on sampling independent and identically distributed (i.i.d) data. However, in federated learning (FL), we know that different environments comprise of different populations and generate non-i.i.d datasets. We propose a new FL strategy *Invariant Federated Averaging* based on minimising empirical risk and penalizing for invariance across different environments. The invariance penalty is based on fine-tuning the global model using an invariant risk minimisation loss function and we demonstrate this leverages the non-i.i.d assumption to improve out-of-distribution generalisation and accuracy on unseen test environments. We provide empirical results on three federated learning tasks, ColouredMNIST, binary classification and linear regression, where each contain challenging spurious correlations and we show upto a 5X improvement in test set accuracy in OOD settings when using Invariant Federated Averaging.

1 Introduction

Federated learning (FL) is a method increasingly used to train machine learning models across a federation of workers without requiring to physically transfer the underlying data to a central location nor having access to the entire dataset. It has numerous promising applications in training shared predictive models, making more data available to use by removing barriers to data sharing, for example across hospital sites (Dayan et al., 2021), or on mobile edge devices (Hard et al., 2018). In this work we are interested in horizontal federated learning (Zhu et al., 2021), whereby a relatively small number of collaborating institutions each collect a dataset of samples from their unique population. These "local" datasets are then used to train a shared predictive model without transferring any sensitive data to a central location.

The contribution of this work is to identify a FL training strategy we call Invariant Federated Averaging (Inv. FedAvg), based on minimising empirical risk and penalizing for invariance between multiple different environments, to leverage non-i.i.d data inherent in FL. This facilitates better global shared models and better generalisation across a federation of different populations. Invariant Federated Averaging is based on a fine-tuning strategy using invariant risk minimisation (IRM) (Arjovsky et al., 2019) which uses a penalised loss function to learn a consistent and invariant data representation across multiple different environments. The intuition is that if a particular feature is consistently predictive of a target variable across multiple environments, each with different population and data distributions, then we can consider that feature as having a causal relationship with the target variable. Hence, in situations where we can learn the underlying causal structure of a task, we can generalise to new out-of-distribution environments (Peters et al., 2016; Shen et al., 2021).

The invariance penalty requires multiple different environments with which to elucidate invariant features. The different environments lend themselves particularly well to FL settings, where for each available environment we can consider a separate worker or collaborator node in that location having access to its own local population and dataset. It is therefore not a surprise that CausalFed, an algorithm for federating IRM, was recently proposed in (Francis et al., 2021). However, whilst data sampled from within an environment can be assumed to be independent and identically distributed (i.i.d),

Preprint. Under review.

the datasets and data populations between different environments cannot (Quiñonero-Candela et al., 2009). In fact, the IRM loss function *requires* the environments to represent different population distributions in order to elucidate those invariant features as candidate causal ones.

It is well known that most modern machine learning strategies rely on the assumption that training and test data are both sampled i.i.d. from the same distribution. Therefore, in this case, empirical gradients are unbiased approximations of the entire population gradient (Mohamed et al., 2020). Consequently, the optimal solution found using empirical gradients on the training data are also expected to be optimal on the test data. In the FL setup however, in all but trivial cases the data are non-i.i.d. across environments which can cause issues generalising across environments.

Population differences and data distribution shifts between different physical environments leads us to require learning strategies that leverage non-i.i.d data assumptions. In this work we highlight and attempt to quantify this interplay between taking empirical gradients that assume i.i.d data, and minimising invariant risk across a federation that requires non-i.i.d populations and data. We demonstrate that fine-tuning models based on an IRM strategy attains better test-set accuracy on held out environments under population and data distributional shift on a range of tasks. We empirically validate our findings on three synthetic tasks, two classification and one regression task, based on models popular in the causality literature (Pearl, 1995; Arjovsky et al., 2019; Aubin et al., 2021).

2 Related Work

Federated learning (FL) aims to train a global predictive model on data distributed across different devices, possibly in different physical locations, without moving the data. It prevents centralised data collection and aggregation in favour of sharing the model and the local model updates. This approach has many proposed benefits, from minimising data-transfer latency to removing barriers to sharing of sensitive data. The federated averaging (FedAvg) algorithm (McMahan et al., 2017) was proposed in order to do this based on transferring the model between workers instead of transferring the data. Each worker updates their current model based on their local dataset and pass the model back to an aggregator server to produce a weighed-average of these models. This new model is then shared back to the workers to begin a new training round. An alternate approach is to perform distributed selective SGD (Shokri & Shmatikov, 2015) to selectively share parameter gradients during the SGD process.

In this work, we consider the horizontal FL setting (Zhu et al., 2021), where different workers each monitor the same set of features. But each worker has its own set of samples, known as its local dataset. Following this nomenclature, we might encounter non-i.i.d. data categories of attribute skew and/or label skew in a federated setup. This raises the problem of averaged local model parameters being far away from the global model parameters. Federated averaging has been shown to achieve acceptable performance in this setting with shallow neural networks, but may fail to converge with deep neural networks (McMahan et al., 2017).

The most similar literature that highlights the risks involved with training a shared model assuming i.i.d data is Zhao et al. (2018). The authors focus on the statistical challenge of federated learning, i.e. i.i.d. sampling of training data is necessary to ensure the stochastic gradient is an unbiased estimate of the full gradient, but practically data from different workers is seldom i.i.d. They show how models trained using FedAvg break down with non-i.i.d. data and that the accuracy reduction is related to weight divergence. We go beyond these previous efforts, and demonstrate how the concept of invariance specifically relies upon the data distributions between environments being non-i.i.d. and that an environment-dependent invariance penalty term applied to the loss leverages the non-i.i.d nature of the task. This allows us to learn powerful generalisable models and progress beyond the FedAvg algorithm for OOD settings.

A federated causal inference method has previously been proposed in Vo et al. (2021) based on the potential outcomes framework popularised by (Rubin, 1978). In Francis et al. (2021), federated causal inference was also explored using IRM (modified such that each client has a set of local NN layers). They consider the link between improved generalisation (through causal instead of associative models) to better privacy of trained models (e.g. harder membership attacks due to less overfitting). However, whilst there is no consensus on general causal machine learning methods, or in particular invariance based methods (Rosenfeld et al., 2020; Kamath et al., 2021), our findings in the FL setting show promise and our approach could be extended to other invariance-based penalties.

3 Data-Distribution in Training

Standard training of a neural network (NN) involves optimising the parameters, traditionally achieved by a continuous optimisation process (LeCun et al., 2012; Mohamed et al., 2020). As we further explain in Subsection 3.1, for the optimisation to find an optimal solution on both the training and the test datasets, it is necessary for the data **x** to be sampled i.i.d., e.g. $p_{\mathsf{train}}(\mathbf{x}) = p_{\mathsf{test}}(\mathbf{x})$.

In federated learning however, local models have access to their own data-generating populations p_{env}. In all but trivial cases, this will result in non-i.i.d. data between different environments, since $p_{\mathsf{env}_1}(\mathbf{x}) \neq p_{\mathsf{env}_2}(\mathbf{x})$. Therefore, as we show in Subsection 3.2, the optimal parameters on the training data will often not generalise to OOD test datasets. To mitigate this behaviour, in Subsection 3.3 we introduce the idea of using an invariance-based penalty function developed to identify invariant features among environments. This can also be viewed as a regulariser for this behaviour.

3.1 Non-Federated Learning

The training of model Φ with parameters θ is achieved via an optimisation process. The parameter space is searched to find the combinations that minimises the expected loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\ell(\Phi(\theta, \mathbf{x}))\right] \tag{1}$$

with $\ell(\Phi(\theta, \mathbf{x}))$ being the contribution to the overall loss from a single input **x**. However, the distribution of **x**, $p(\mathbf{x})$, is not known, thus the above integral is intractable. As a consequence, it is practice to minimise the empirical loss over a set of N observed data points $\{\mathbf{x}_n\}_{n=1}^{N}$ sampled i.i.d from p,

$$\mathcal{L}_N(\theta) = \frac{1}{N}\sum_{n=1}^{N} \ell(\Phi(\theta, \mathbf{x}_n)). \tag{2}$$

In gradient descent[1], the parameters of the model are iteratively updated by following the gradient of the loss function with respect to the parameters, i.e.,

$$\theta_{t+1} = \theta_t - \eta \bar{g}_N(\theta_t), \tag{3}$$

where,

$$\bar{g}_N(\theta_t) \equiv \nabla_\theta \mathcal{L}_N(\theta_t), \tag{4}$$

and η is a small positive learning rate. The convergence of these methods is conditional on the observations $\mathbf{x}_n$ being i.i.d. (Bottou, 2010; Shen et al., 2021). The sampled gradients $\bar{g}_N$ are an unbiased estimator (Mohamed et al., 2020) and therefore, in expectation, are equal to the gradient of the expected loss function:

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\bar{g}(\theta)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}_n\sim p(\mathbf{x})}\left[\nabla_\theta \ell(\Phi(\theta, \mathbf{x}_n))\right] = \nabla_\theta \mathcal{L}(\theta). \tag{5}$$

Therefore, as $N \to \infty$, if each observation is independent and identically sampled from $p(\mathbf{x})$, by the law of large numbers, $\bar{g}(\theta) \to \nabla_\theta \mathcal{L}(\theta)$. We can therefore expect, for convex problems, that given enough data we will converge to the optimal parameters θ for the expected loss objective in Equation (1).

3.2 Federated Learning

In federated learning, each local model does not have direct access to the underlying data generating distribution $p(\mathbf{x})$. In this setting, a common practice is to follow the *federated averaging* (FedAvg) algorithm proposed in McMahan et al. (2017), where gradient descent is implemented separately on each local model, with an aggregation server performing parameter averaging. However, each collaborator k has access to only their own environment population $p_k(\mathbf{x})$. The FedAvg algorithm is applicable to any finite-sum objective of the form:

$$\mathcal{L}_N^{fed} = \sum_{k=1}^{K} \frac{n_k}{N} F_k(\theta) \quad \text{where} \quad F_k(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell(\Phi(\theta, \mathbf{x}_i^{(k)})) \tag{6}$$

[1] In this section we focus on applying an invariance penalty to the loss and applying standard gradient descent algorithm, but our approach can be extended to other optimisation methods.

and with n_k the number of data samples available on each of the K collaborator sites, i.e. $\sum_{i=1}^{K} n_i = N$. Since we consider each collaborator having access to its own environment, we have K distinct datasets $D_k = \{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$. Each collaborator optimises their own loss function, $F_k(\theta)$, on their available data $\mathbf{x}_i^{(k)} \sim p_k$, for a number of epochs. Then the local models are averaged together by an aggregation central server.

Without loss of generalisation, with only one step of local training, the gradient can be identified as a linear sum of the local gradients:

$$\hat{g}(\theta) = \sum_{k=1}^{K} \hat{g}_k(\theta) = \sum_{k=1}^{K} \frac{n_k}{N} \nabla_\theta F_k(\theta) = \sum_{k=1}^{K} \frac{n_k}{N} \sum_{n=1}^{n_k} \frac{1}{n_k} \nabla_\theta \ell(\Phi(\theta, \mathbf{x}_n^{(k)})) \tag{7}$$

$$= \frac{1}{N} \sum_{n,k=1}^{K,n_k} \nabla_\theta \ell(\Phi(\theta, \mathbf{x}_n^{(k)})). \tag{8}$$

Since the gradients across collaborators are sampled from different distributions p_k, the expected gradient in the federated setting does not coincide with the gradient when we have access to samples from p, i.e. in the non-federated setting[2]:

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}_n \sim p(\mathbf{x})} \left[\nabla_\theta \ell(\Phi(\theta, \mathbf{x}_n))\right] \neq \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{n_k} \mathbb{E}_{\mathbf{x}_n^{(k)} \sim p_k(\mathbf{x})} \left[\nabla_\theta \ell(\Phi(\theta, \mathbf{x}_n^{(k)}))\right] \tag{9}$$

Since the gradient is no longer an unbiased estimator, it will not converge to the gradient of the expected loss as the number of samples increases. Thus it is not going to generalise well to out-of-distribution test cases.

3.3 Leveraging Invariance in Federated Learning

Invariance has become a popular topic in causal machine learning for out-of-distribution generalisation. The main idea assumes that if the value of a feature directly causes changes in a target, then that relationship will be invariant to interventions, or different environments, thus allowing for generalisation to out-of-distribution cases. Invariance relies on the data being non-i.i.d. across environments, and we leverage this insight to formulate the Invariant Federated Averaging method to improve training in FL settings.

Invariance Following Peters et al. (2017), a set of features $\{X_i\}_{i \in S}$ with $S \subseteq \{1, ..., m\}$, is invariant to a target Y if there is a function f defined such that in any environment $e \in \mathcal{E}$ the following is satisfied:

$$Y^e = f(\{X_i^e\}_{i \in S}) + \varepsilon, \tag{10}$$

where ε is a random noise with mean zero, finite variance and the same distribution across all environments $e \in \mathcal{E}$. Among the different methods that have been proposed to identify invariant features, IRM (Arjovsky et al., 2019) identifies invariant variables by considering an invariance-based penalty applied to the loss function which we implement in a federated settings.

Specifically, IRM seeks classifiers w that are optimal across multiple environment datasets $D_e = \{(\mathbf{x}_i^{(e)}, y_i^{(e)})\}_{i=1}^{n_e}$. Considering a model Φ of the observations, e.g. a neural network, this is expressed with a constrained optimization problem of the following form:

$$\min_{\theta, w} \sum_{e \in \mathcal{E}_{\text{train}}} \mathcal{L}_e^{inv}(w, \theta) \tag{11}$$

$$\text{subject to} \quad w \in \operatorname*{argmin}_{\tilde{w}} \mathcal{L}_e^{inv}(\tilde{w}, \theta), \ \forall\, e \in \mathcal{E}_{\text{train}}, \tag{12}$$

where $\mathcal{L}_e^{inv}(w, \theta) = \frac{1}{n_e} \sum_{n=1}^{n_e} \ell\left(w \circ \Phi(\theta, \mathbf{x}_n^{(e)})\right)$, and $\mathcal{E}_{\text{train}}$ is the set of training environments. Notice that removing the constraint in optimization problem (12) recovers the classical empirical risk minimization (ERM) objective. Incorporating this constraint results in a bi-leveled optimization

[2] In the trivial case, where $p_k = p$, then the inequality in Equation (9) does not hold.

problem, which is computationally challenging. Arjovsky et al. (2019) proposes a variant of IRM that is more practical to compute:

$$\min_{\theta} \sum_{e \in \mathcal{E}_{\mathsf{train}}} \mathcal{L}_e^{inv}(w_0, \theta) + \lambda \left\| \nabla_{w|w=w_0} \mathcal{L}_e^{inv}(w, \theta) \right\|^2, \tag{13}$$

where w_0 is a user specified vector with a user specified dimension (e.g., a scalar).

Invariant Federated Averaging To mitigate the learning bias caused by the non-iid datasets available to the collaborators in a federation, we propose leveraging invariance in the training scheme. Under this perspective, we assume that each collaborator has access to data from its own environment, and we regularise the training of a shared model by disincentivising the consideration of models that are overly environment specific, i.e. models that over-perform in one environment relative to the others (or on one collaborator relative to others).

We thus extend the federated averaging algorithm introduced in McMahan et al. (2017) by introducing the invariance penalty to the loss function that appears in Equation (13). We call this new FL training scheme Invariant Federated Averaging, where the objective is to learn a global model Φ, defined by parameters θ, that minimises the following loss function:

$$\mathcal{L}_N^{IFA} = \sum_{k=1}^{K} \frac{n_k}{N} F_k^{IFA}(\theta) \tag{14}$$

where

$$F_k^{IFA}(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(w_0 \circ \Phi(\theta, \mathbf{x}_i^{(k)})) + \lambda \left\| \nabla_{w|w=w_0} \ell\left(w \circ \Phi(\theta, \mathbf{x}_i^{(k)})\right) \right\|^2, \tag{15}$$

and the classifier w is defined as a scalar $w_0 = 1$.

One particular insight, and contributions of our work, is implementing a schedule over the parameter λ. That is, from experimental results, and in accordance with Arjovsky et al. (2019), the parameter λ is chosen to be zero at the beginning of training; once the federated averaging algorithm has converged, minimising the empirical risk, λ is updated such that the invariance penalty is several order of magnitude larger than the empirical loss.

This particular training strategy can be interpreted as first training the parameters of a shared model using the standard empirical loss function, and then fine-tuning using the invariance penalty. We find this training strategy to be rewarding for out-of-distribution generalization, e.g. on novel test sets.

4 Experiments

We introduce three tasks based on popular synthetic datasets from the causality literature: two classification and one regression. The two classification tasks simulate the case where a separate OOD test set is available, and the third task assumes that each collaborator has their own test set available, albeit with shuffled spuriously correlated features for maximal difficulty, i.e. not drawn from the training set distribution.

We begin by describing the datasets, including the details of invariant and spuriously correlated features, and then present our results in Subsection 4.2. In detail the three tasks include: 1) an image classification task on the Coloured MNIST dataset (Arjovsky et al., 2019); 2) a binary classification task based on a hospital scenario; 3) a linear regression task based on the popular Example 1 presented in Arjovsky et al. (2019) and Aubin et al. (2021).

4.1 Datasets

Task 1: Coloured MNIST The coloured MNIST dataset modifies the classical MNIST grayscale image dataset (LeCun et al., 1998) into a binary classification task where the label correlates in a strong but spurious way with the digit's color (Arjovsky et al., 2019). The dataset contains images of digits 0-9 and binary labels are assigned to each image such that digits 0-4 have label $\tilde{y} = 0$ and digits 5-9 have label $\tilde{y} = 1$. The final label y for each image is obtained by flipping $\tilde{y}$ with probability of 25%.

Each digit is coloured either green or red based on the value of y. In two distinct training environments, the colour is assigned after flipping y with 10% and 20% chance, respectively. This results in a setup where classical empirical risk minimization will learn the spurious correlation between the colour and y, rather than the digit's value. However, such an approach will fail on the out-of-distribution test set where the spurious correlation is reversed: the colour of the digit is based on y after flipping it with 90% probability[3].

Task 2: Binary Classification A classification task modeled on a common federated learning scenario across hospital collaborators. A binary target value $y_i^e \in \mathbb{Z}_2$ indicates if a patient carries a particular disease and directly depends on the values of a set of d_{inv} invariant variables x^e_{inv} which correspond to the direct causes of the disease. The target y_i^e takes a value of 1 if $\sum_{d_{inv}} x^e_{inv,i} > 0$. Alongside the invariant variables, a set of d_{rand} uncorrelated random variables $x^e_{rand,i}$ that are not of relevance for the disease in question are observed. Finally, the dataset also contains d_{spu} spuriously correlated variables $x^e_{spu,i}$ (e.g. fever if the considered disease is a bacterial infection) that are environment dependent governed by probability p^e.

Formally, the feature vector is given by $x_i^e = (x^e_{(inv,i)}, x^e_{(rand,i)}, x^e_{spu,i})$ and the variables are drawn as follows:

$$\begin{aligned} x^e_{inv,i} &\sim \mathcal{Z}_{d_{inv}} && (16)\\ x^e_{rand,i} &\sim \mathcal{Z}_{d_{rand}} && (17)\\ y_i^e &= \mathbb{1}\left[\sum_{d_{inv}} x^e_{inv,i} > 0\right] && (18)\\ x^e_{spu,i} &= \text{flip}(y_i^e, p^e) && (19) \end{aligned}$$

where $\text{flip}(y_i^e, p^e)$ reverses the value of y_i^e with probability p^e.

For our experiments, the binary classification task is instantiated with $d_{inv} = d_{rand} = 5$ and $d_{spu} = 1$ across two training environments ($p_{\mathsf{train0}} = 0.1$, $p_{\mathsf{train1}} = 0.2$) and one separate OOD test environment ($p_{\mathsf{test}} = 0.9$). We generate 50,000 samples per environment.

Task 3: Linear Regression A particularly challenging linear least-squares regression task based on Example 1 in Arjovsky et al. (2019) and Aubin et al. (2021), where the features contain both causes and effects of a target variable. This is defined by a collection of environments $e \in \mathcal{E}$ whose data $D_e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ contain n_e samples. The input feature vector is similar to previous tasks: $x_i^e = (x^e_{(inv,i)}, x^e_{(spu,i)}) \in \mathbb{R}^d$ is the combination of the invariant variables, x^e_{inv}, and the spuriously correlated variables, x^e_{spu}, with the target y^e.

Additionally from the work in Aubin et al. (2021), here the noise on the target is fixed among environments to be consistent with invariance in Peters et al. (2017), and the spurious features have an environment noise term to insure that there is a unique solution to the problem.

The samples in D_e, for every $e \in \mathcal{E}$ and $i = 1, \ldots, n_e$, are generated in the following way:

$$\begin{aligned} x^e_{inv,i} &\sim \mathcal{N}_{d_{inv}}(0, (\sigma^e)^2), && (20)\\ \tilde{y}_i^e &\sim \mathcal{N}_{d_{inv}}(W_{yx} x^e_{inv,i}, \sigma_n^2), && (21)\\ x^e_{spu,i} &\sim \mathcal{N}_{d_{spu}}(W_{xy} \tilde{y}_i^e, (\tilde{\sigma}^e)^2), && (22)\\ y_i^e &= 1^\top \tilde{y}_i^e && (23) \end{aligned}$$

where the matrices $W_{yx} \in \mathbb{R}^{d_{inv} \times d_{inv}}$ and $W_{xy} \in \mathbb{R}^{d_{spu} \times d_{inv}}$ have elements drawn i.i.d from a Gaussian normal distribution $\mathcal{N}(0, 1/d_{inv})$ and are fixed for each environment e. The variances $(\sigma^e)^2$ and $(\tilde{\sigma}^e)^2$ are environment dependent.

In our experiments we consider three environments with $n_e = 300$ samples. The variances are $\{(\sigma^e, \tilde{\sigma}^e)\}_{e=1}^3 = \{(0.1, 5), (1.5, 0.1), (5, 1.5)\}$.

[3] To facilitate computations we downsample the MNIST images from 24 × 24 to 14 × 14 pixel.

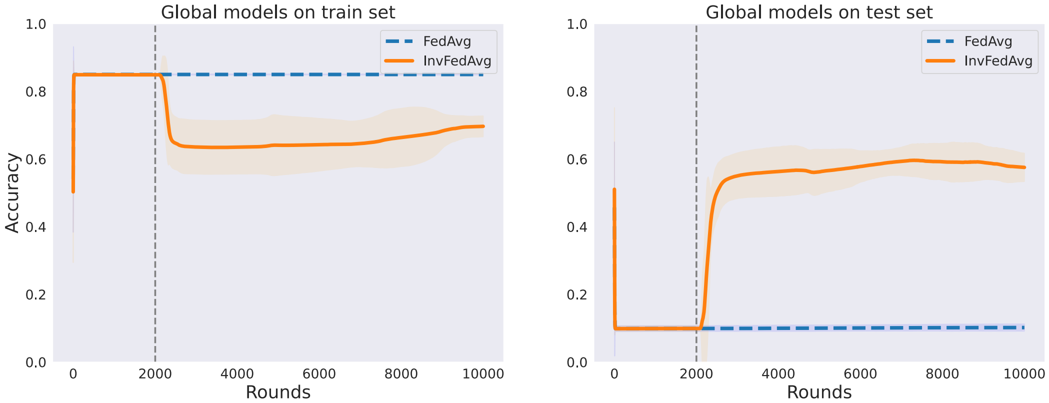

Figure 1: Development of model accuracy on train (**left**) and test set (**right**) for classic federated averaging (FedAvg) and our Invariant Federated Averaging approach (Inv. FedAvg) on the coloured MNIST dataset. The invariance penalty is added to the loss starting with training round 2000 (grey dashed line). Shaded area indicates two standard deviations from the mean.

4.2 Experiment Results

This section explores the empirical benefits of invariance-based penalties in federated learning to improve performance on novel test data on our three tasks defined above. Comparisons with federated averaging (FedAvg) algorithm (McMahan et al., 2017) shows that our method, Invariant Federated Averaging, consistently outperforms this baseline approach. In particular, minimisation of empirical risk can fail catastrophically in the presence of spuriously correlated features in the training data (accuracy of 0.1 on the binary coloured MNIST task). Whereas leveraging environment specific information allows us to successfully distinguish spurious from invariant features, resulting in test accuracy of 0.58 (maximum test accuracy is limited to 0.75 due to random label flipping in this dataset).

Furthermore, we show that the invariant approach is robust with respect to the number of collaborators involved in training, and the number of local model updates per communication round.

Task 1: Coloured MNIST We consider the coloured MNIST dataset in a federated learning setup. The environments are distributed as different workers and the federated averaging (FedAvg) algorithm is used as a baseline. We use a simple Multi-Layer-Perceptron (MLP) with one 40-unit hidden layer and Rectified Linear Unit (ReLU) activation functions. For the FedAvg model, the MLP is trained with the Adam (Kingma & Ba, 2014) optimizer to minimize binary cross entropy on the training data of each environment. This efficiently reduces the training error and achieves an accuracy of 0.85±0.00 on the training data. Evaluation on the test set, however, reveals that the MLP relies on the spurious colour of the digits for its predictions and fails to generalize beyond the training data. The accuracy on the test is 0.10±0.00.

We see that Invariant Federated Averaging (Inv. FedAvg) mitigates these shortcomings by incorporating the invariance penalty term into the binary cross entropy loss function. In particular, we fine-tune the model after the ERM model converges with the adjusted loss function (see Equation (14)), including the environment-dependent penalty term. This allows the model to adjust its dependence on the spurious image colour to the correct features which indicate the digit value. The accuracy on the test dataset subsequently improves to 0.58±0.03 after adding the invariance penalty to the loss.

Task 2: Binary Classification We use a 3-layer MLP, with one hidden layer of 10 units. After training for 10,000 round with the Adam optimizer, the standard FedAvg model reaches an average accuracy of 0.85±0.00 on the training environments. On the test data, however, this approach fails with an accuracy of 0.11±0.00. It is worth noting that a classifier entirely based on the spurious features, i.e. color of digits, results in training accuracy of 0.85 and test accuracy of 0.10.

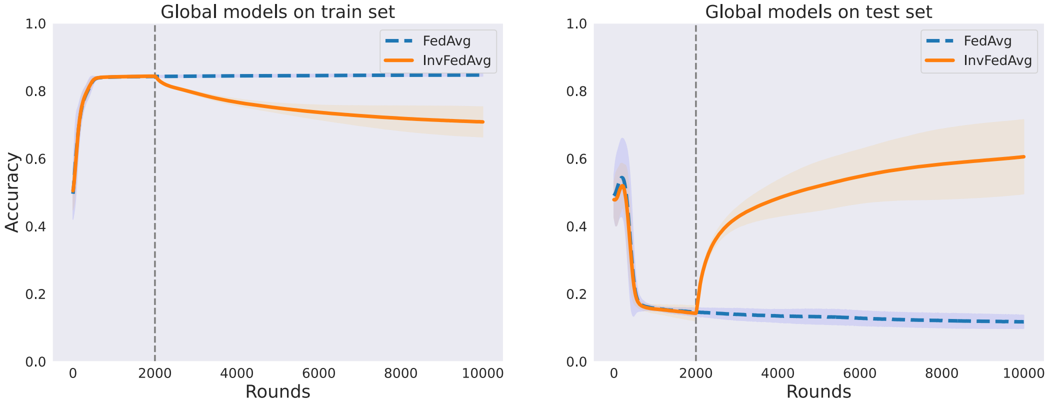

Figure 2: Development of model accuracy on train (**left**) and test set (**right**) for classic federated averaging (FedAvg) and our Invariant Federated Averaging (Inv. FedAvg) approach on the linear regression task. The invariance penalty is added to the loss starting with training round 2000 (grey dashed line). Shaded area indicates 2 standard deviations from the mean.

Table 1: Experiment results.

METHOD	TASK	N	EPOCHS	TRAIN ACC/MSE	TEST ACC/MSE
FEDAVG	COLOUREDMNIST	25,000	10,000	0.85±0.00	0.10±0.00
INV. FEDAVG	COLOUREDMNIST	25,000	10,000	0.69±0.02	**0.58**±0.03
FEDAVG	CLASSIFICATION	50,000	10,000	0.85±0.00	0.11±0.00
INV. FEDAVG	CLASSIFICATION	50,000	10,000	0.70±0.02	**0.62**±0.05
FEDAVG	REGRESSION	900	6,000	2.73±1.15	7.10±4.01
INV. FEDAVG	REGRESSION	900	6,000	4.21±1.81	**6.53**±2.99

Using the environment-dependent loss formulation presented in Equation (13), the model learns invariance to the spurious feature x^e_{spu} and achieves a test set accuracy of 0.62±0.05. Due to random label flipping in the dataset construction, the maximum achievable accuracy would be 0.75.

Robustness in FL Settings: For the binary classification task 2, we also investigate the robustness of our proposed Invariant Federated Averaging across two settings: Varying the number of collaborators and varying the number of local training epochs per communication round. We vary the number of local training epochs E (full batch gradient descent) on each collaborator in the range $E \in \{1, 2, 5, 10, 20\}$ and find consistent test set accuracy after 10,000 communication rounds for Invariant Federated Averaging with all values of E (see Figure 3).

To assess the effect of an increased number of different environments on the efficacy of our invariance-based approach, we increase the number of collaborators K, each of which has access to its own environment, in the range $K \in \{2, 5, 10, 20, 50\}$. The environment variable p_{env}, which governs the strength of the correlation between spurious feature and target, is drawn from the uniform distribution as $p_{env} \sim \mathcal{U}(0, 0.2)$ for each training environment. We find that the approach is robust with respect to the number of collaborators, with a small degradation of test set accuracy as the number of collaborators reaches 50 (see Figure 3).

Table 2: Task 3: Linear Regression experiment MSE results on shuffled environment test sets.

ENVIRONMENT	FEDAVG	INV. FEDAVG
ENV 1	3.91 ± 0.78	**3.25** ± 0.29
ENV 2	4.31 ± 0.86	**4.02** ± 0.67
ENV 3	13.08 ± 11.32	**12.32** ± 8.40

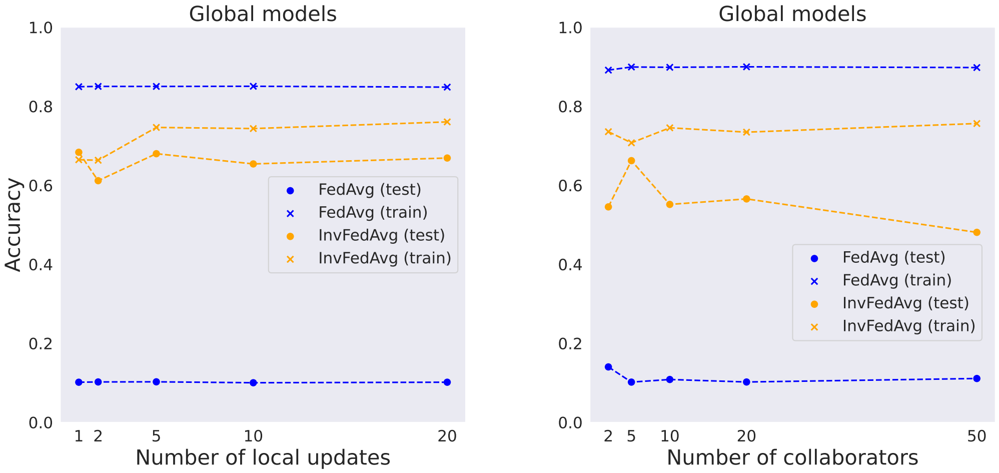

Figure 3: Test and train accuracy on the synthetic classification dataset after 10,000 communication rounds. **Left:** Effect of varying number of local training epochs (full-batch gradient descent) per communication round. **Right:** Effect of varying number of collaborators.

Task 3: Linear Regression In line with Arjovsky et al. (2019), each of three collaborators has access to their own test set, $p_{\mathsf{test}_\mathsf{i}}$ which are a copy of the train distributions $p_{\mathsf{train}_\mathsf{i}}$ but with the spuriously correlated features shuffled between samples. The intuition is that the invariance-based approach will correctly focus on the invariant features in the dataset and ignore the spuriously correlated features, whereas the federated averaging algorithm will rely the spurious correlations and will not generalise as well to OOD dataset.

We use a 3-layer MLP, with one hidden layer of 30 units. After training for 6,000 training rounds, and applying the invariance penalty after round 3000, we show an improvement in average test set accuracy (mean squared error MSE) on out-of-distribution test-sets in Table 1. The standard FedAvg model reaches an impressive average MSE of 2.73±1.15 on the training environments, however, on the shuffled test dataset where spurious correlations are altered it does not generalise as well as the Invariant Federated Averaging. Per environment test set accuracies are presented in Table 2 showing the invariance-penalty provided an improvement on all environment test sets.

5 Conclusion

In the FL setting each collaborator often has access to its own environment and data-generating populations. Under this assumption, we demonstrate that it is beneficial to use a learning strategy that leverages the assumption of non-i.i.d datasets between environments. We draw connection to the invariance literature and apply an invariance-based penalty to the standard empirical loss function which disincentivises the consideration of models that are overly environment specific. To this end, we propose a novel FL strategy *Invariant Federated Averaging* based on minimising empirical risk and fine-tuning penalizing for invariance across different environments. We show across three different experiment settings improvements in out-of-distribution generalisation and test accuracy on unseen test environments.

References

Arjovsky, Martin, Bottou, Léon, Gulrajani, Ishaan, and Lopez-Paz, David. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aubin, Benjamin, Słowik, Agnieszka, Arjovsky, Martin, Bottou, Leon, and Lopez-Paz, David. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Bottou, Léon. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMP-STAT'2010*, pp. 177–186. Springer, 2010.

Dayan, Ittai, Roth, Holger R, Zhong, Aoxiao, Harouni, Ahmed, Gentili, Amilcare, Abidin, Anas Z, Liu, Andrew, Costa, Anthony Beardsworth, Wood, Bradford J, Tsai, Chien-Sung, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature Medicine*, pp. 1–9, 2021.

Francis, Sreya, Tenison, Irene, and Rish, Irina. Towards causal federated learning for enhanced robustness and privacy. *arXiv preprint arXiv:2104.06557*, 2021.

Hard, Andrew, Rao, Kanishka, Mathews, Rajiv, Ramaswamy, Swaroop, Beaufays, Françoise, Augenstein, Sean, Eichner, Hubert, Kiddon, Chloé, and Ramage, Daniel. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Kamath, Pritish, Tangella, Akilesh, Sutherland, Danica, and Srebro, Nathan. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2021.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Yann A, Bottou, Léon, Orr, Genevieve B, and Müller, Klaus-Robert. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

McMahan, Brendan, Moore, Eider, Ramage, Daniel, Hampson, Seth, and y Arcas, Blaise Aguera. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mohamed, Shakir, Rosca, Mihaela, Figurnov, Michael, and Mnih, Andriy. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

Pearl, Judea. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Peters, Jonas, Bühlmann, Peter, and Meinshausen, Nicolai. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Quiñonero-Candela, Joaquin, Sugiyama, Masashi, Lawrence, Neil D, and Schwaighofer, Anton. *Dataset shift in machine learning*. Mit Press, 2009.

Rosenfeld, Elan, Ravikumar, Pradeep, and Risteski, Andrej. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Rubin, Donald B. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp. 34–58, 1978.

Shen, Zheyan, Liu, Jiashuo, He, Yue, Zhang, Xingxuan, Xu, Renzhe, Yu, Han, and Cui, Peng. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Shokri, Reza and Shmatikov, Vitaly. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Vo, Thanh Vinh, Hoang, Trong Nghia, Lee, Young, and Leong, Tze-Yun. Federated estimation of causal effects from observational data. *arXiv preprint arXiv:2106.00456*, 2021.

Zhao, Yue, Li, Meng, Lai, Liangzhen, Suda, Naveen, Civin, Damon, and Chandra, Vikas. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhu, Hangyu, Xu, Jinjin, Liu, Shiqing, and Jin, Yaochu. Federated learning on non-iid data: A survey. *arXiv preprint arXiv:2106.06843*, 2021.