FedPop: A Bayesian Approach for Personalised Federated Learning

Nikita Kotelevskii[*]
Skolkovo Institute of Science and Technology
Moscow, Russia
Nikita.Kotelevskii@skoltech.ru
[*] Equal contribution

Maxime Vono[*]
Criteo AI Lab
Paris, France
m.vono@criteo.com
[*] Equal contribution

Alain Durmus
ENS Paris-Saclay
alain.durmus@ens-paris-saclay.fr

Eric Moulines
Ecole Polytechnique
eric.moulines@polytechnique.edu

Abstract

Personalised federated learning (FL) aims at collaboratively learning a machine learning model tailored for each client. Albeit promising advances have been made in this direction, most of the existing approaches do not allow for uncertainty quantification which is crucial in many applications. In addition, personalisation in the cross-silo and cross-device setting still involves important issues, especially for new clients or those having a small number of observations. This paper aims at filling these gaps. To this end, we propose a novel methodology coined `FedPop` by recasting personalised FL into the population modeling paradigm where clients' models involve *fixed* common population parameters and *random* effects, aiming at explaining data heterogeneity. To derive convergence guarantees for our scheme, we introduce a new class of federated stochastic optimisation algorithms that relies on Markov chain Monte Carlo methods. Compared to existing personalised FL methods, the proposed methodology has important benefits: it is robust to client drift, practical for inference on new clients, and above all, enables uncertainty quantification under mild computational and memory overheads. We provide non-asymptotic convergence guarantees for the proposed algorithms and illustrate their performances on various personalised federated learning tasks.

1 Introduction

Federated learning (FL) is a recent machine learning paradigm in which distributed clients holding siloed data collaborate in solving a learning problem, usually under the coordination of a central server [51, 25]. One of the main focuses of FL is on so-called *cross-device* applications where a large number of personal electronic devices such as mobile phones, wearable devices or home assistants collect and store data at the edges of a decentralised network [39].

While standard FL methods [39, 1, 27, 23, 32] have focused on training a global model that can be applied to individual agents, the relevance of such inferences has recently been questioned due to statistical *heterogeneity* between clients. Indeed, the considered common model may not generalise well on a client with a local data distribution that differs significantly from the global data distribution, especially if that client has not participated in the training process. In fact, it might even be better for such clients to derive a local model from their own data set. To circumvent these issues, a number of *personalised federated learning* approaches have recently been proposed, that use local models to fit client-specific data distribution while capturing some shared knowledge via a federated scheme

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[49]. Personalisation has previously been approached using multi-task learning [48], meta-learning [24, 28], client clustering [4], data interpolation [38], model interpolation [19, 20] or partially local models [47, 8]. While these methodologies are promising, they only partially solve the personalisation problem in highly heterogeneous federated settings and have no means of quantifying uncertainty. In addition, cross-device FL also faces other important challenges such as (extreme) partial device participation, small local data sets, limited upload bandwidth, and device capabilities [25]. Addressing these problems for personalised FL requires new paradigms regarding how model knowledge is shared and personalisation is performed locally.

Proposed Approach. In this paper, we adopt a novel perspective to model the problem of personalised FL. This paradigm, called *mixed-effects modeling* (also known as multi-level or population approach) is widely used to analyse data that have a clustered or nested structure, as in medical or biological research where multiple observations per patient are available [17, 37, 30]. Although the hierarchical structure of FL has already been noted [42, 18, 22], the mixed-effects paradigm has interestingly never been considered. Leveraging this framework, we develop a new model for personalised FL called `FedPop` and propose an efficient computational solution to perform inference under this model. More precisely, we introduce a novel class of federated stochastic approximation algorithms based on parallel Markov Chain Monte Carlo (MCMC) methods. In the proposed framework, we also pay special attention to the cross-device setting by taking into account partial client participation, and by addressing the communication bottleneck with both multiple local updates and the use of lossy compression operators.

Benefits. Up to the authors' knowledge, `FedPop` is the first *personalised FL* approach that allows *cheap uncertainty quantification* with a theoretically-grounded methodology. The proposed framework also comes with other interesting properties. First, in contrast to most personalised FL methods that only consider "fixed-effects" models [8, 21, 48], `FedPop` provides credibility information (via credibility regions) and allows more accurate inference for clients with small and heterogeneous local data via *partial pooling* [17]. In addition, inference for new clients which did not participate in the training phase can be easily performed by sampling from the prior over the local random effects. Second, contrary to existing Bayesian FL approaches that aim to provide credibility information by sampling from a target posterior distribution [22, 53, 50, 15], `FedPop` allows to perform personalisation and cheaper on-device uncertainty quantification taking an empirical Bayes prediction approach. Finally, an important benefit of `FedPop` is its ability to allow for multiple local updates without suffering from the client-drift phenomenon [27].

Outline and Contributions. Our contributions are fourfold. First, in Section 2, we propose a novel probabilistic methodology, which we call `FedPop`, to address personalisation under the cross-device FL paradigm. To perform efficient inference under this model, we introduce a general class of stochastic approximation algorithms based on MCMC. Second, we provide in Section 3 non-asymptotic convergence guarantees for the proposed methodology. Then, we perform in Section 4 a comparison between the proposed approach and exisiting works. Finally, we illustrate in Section 5 the benefits of our methodology on several federated learning benchmarks involving both synthetic and real data.

2 Proposed Approach

In this section, we present the statistical estimation problem we are considering and the proposed methodology called `FedPop` to address it.

Problem Formulation. We are interested in the *cross-device* FL setting involving a large number $b \in \mathbb{N}^*$ of clients, potentially unreliable *i.e.* not necessarily available at each communication round. These clients are assumed to own sensitive local data sets $\{\mathrm{D}_i\}_{i\in[b]}$. In this framework, we aim to make both uncertainty quantification and personalised statistical inference by learning a local model tailored to each client. To this end, and inspired by the population approach used in the biological and physical sciences [30], we consider mixed-effects modeling for each client leading to the local marginal likelihood function defined, for any $i \in [b]$, by

$$p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi\,, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\,, \tag{1}$$

where $\phi \in \Phi \subseteq \mathbb{R}^{d_\Phi}$ stands for a *fixed effect* and $\{z^{(i)}\}_{i\in[b]} \in \mathsf{Z}$, $\mathsf{Z} = \prod_{i=1}^{b} \mathbb{R}^d$, represent *random effects* aimed at explaining statistical heterogeneity between local data sets $\{\mathrm{D}_i\}_{i\in[b]}$.

The objective of the fixed (*i.e.* constant across all clients) part is to capture a common representation (*e.g.* same features across different classes of images) while the random part, which is typically low-dimensional, performs personalisation and is assumed to be drawn from a *population* prior whose variance aims at modeling data heterogeneity.

Figure 1 illustrates this statistical framework, referred to as FedPop, by showing its directed acyclic graph (DAG) where grey-filled shapes indicate observed variables, white-filled shapes unknown variables and squared shapes variables to be estimated.

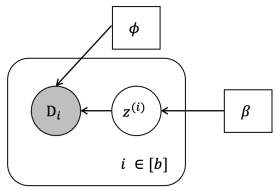

Figure 1: DAG for FedPop.

When the size of the local data set D_i is small, this common prior leverages information from other clients to limit the risk of overfitting and is often called *partial pooling* in the multi-level statistical literature [17, Section 12]. Examples of model architectures involving ϕ and $\{z^{(i)}\}_{i\in[b]}$ include for instance *composition*-based architectures $p(\mathrm{D}_i \mid \phi, z^{(i)}) = p(\mathrm{D}_i \mid h_\phi \circ h_{z^{(i)}})$ where h_ϕ and $h_{z^{(i)}}$ are two neural networks [8, 2]. For the sake of generality, we propose to adopt a flexible energy-based prior distribution of the form for each $i \in [b]$,

$$p(z^{(i)} \mid \beta) = \frac{1}{Z(\beta)} \exp\left\{-E(z^{(i)};\beta)\right\} \text{ , where } Z(\beta) = \int_{\mathbb{R}^d} \exp\left\{-E(z^{(i)};\beta)\right\} \mathrm{d}z^{(i)} \text{ .}$$

Here, $Z(\beta)$ is a normalising constant, and $E(\cdot;\beta)$ represents an *energy* function, typically a neural network, parameterised by a set of parameters $\beta \in \mathsf{B} \subseteq \mathbb{R}^{d_\mathsf{B}}$ [31]. This framework is particularly interesting in the cross-device setting where the number of clients b is large as it allows for efficient enrichment of the model. However, in the case where b is small, the inference of the parameter β is difficult. In this situation, a more pragmatic solution is to consider a common prior for the local random effects $\{z^{(i)}\}_{i\in[b]}$ which is held fixed, *i.e.* $p(z^{(i)} \mid \beta) \propto \exp\{-E(z^{(i)})\}$ for any $\beta \in \mathsf{B}$. Finally, for completeness, we allow the use of a prior model $p(\phi,\beta) = p(\phi)p(\beta)$ for the hyperparameters $\{\phi,\beta\}$. Note, that this assumption has been made for the sake of simplicity and can be generalized easily. Indeed, we can set a general joint prior $p(\phi,\beta)$ (e.g. a joint Gaussian prior with a non-diagonal covariance matrix) and perform stochastic optimization, similar to the one introduced below. Using Bayes' rule [45] and by denoting $\mathrm{D} = \sqcup_{i=1}^b \mathrm{D}_i$ the global data set, the posterior distribution associated with these hyperparameters admits a probability density function which can be written as

$$p\left(\phi,\beta \mid \mathrm{D}\right) \propto p(\phi)p(\beta) \prod_{i=1}^{b} \left[\int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\right] \text{ .}$$

Note, that the independence between local likelihood functions is not an assumption but a core property of the FL paradigm [26]. Set $\theta = \{\phi,\beta\} \in \Theta$ with $\Theta = \Phi \times \mathsf{B}$. In the sequel, we will be interested in solving the maximum a posteriori problem given by

$$\theta^\star \in \underset{\theta\in\Theta}{\arg\max}\ \log p(\phi,\beta \mid \mathrm{D}) \text{ ,} \tag{2}$$

$$\log p(\phi,\beta \mid \mathrm{D}) \propto \log p(\phi) + \log p(\beta) + \sum_{i=1}^{b} \left[\log \int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\right] \text{ .} \tag{3}$$

Once we have estimated $\theta^\star$, using an empirical Bayesian approach, we can perform "for free" on-device uncertainty quantification for each client $i \in [b]$ by sampling from the local posterior distribution $p(z^{(i)} \mid \mathrm{D}_i, \phi^\star, \beta^\star)$, which is typically designed to be low-dimensional.

Algorithm. To solve the optimisation problem (2), we can either use an *alternating maximisation* algorithm or perform global maximisation over Θ. Since the former approach requires more upload bandwidth, in this work we consider the second alternative which is more suitable for FL. The gradient of the objective function (3) being intractable, we propose to resort to the stochastic approximation framework [43] which iteratively defines $(\phi_k,\beta_k)_{k\in\mathbb{N}}$, starting from any $(\phi_0,\beta_0) \in \Theta$, via the recursions for any $k \in \mathbb{N}$,

$$\beta_{k+1} = \Pi_\mathsf{B}\left(\beta_k + \eta_{k+1}^{(1)}\left[\nabla_\beta \log p(\beta) + \sum_{i=1}^{b} \mathrm{g}_k^{(i)}(\phi_k,\beta_k)\right]\right) \text{ ,}$$

$$\phi_{k+1} = \Pi_\Phi \left(\phi_k + \eta_{k+1}^{(2)} \left[\nabla_\phi \log p(\phi) + \sum_{i=1}^{b} h_k^{(i)}(\phi_k, \beta_k) \right] \right) ,$$

where Π_{C} denotes the projection onto $\mathsf{C} \in \{\Phi, \mathsf{B}\}$, $(\eta_k^{(1)}, \eta_k^{(2)})_{k \in \mathbb{N}^*}$ are sequences of step-sizes, and $\{g_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ and $\{h_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ are estimators of the intractable gradients $(\phi, \beta) \mapsto \nabla_\beta \log p(\mathrm{D}_i \mid \phi, \beta)$ and $(\phi, \beta) \mapsto \nabla_\phi \log p(\mathrm{D}_i \mid \phi, \beta)$ at (ϕ_k, β_k), where $p(\mathrm{D}_i \mid \phi, \beta)$ is defined in (1) for any $i \in [b]$.

The choices of the estimators $\{g_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ and $\{h_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ are motivated by the Fisher identity. More precisely, under mild regularity assumptions, and using the Lebesgue dominated convergence theorem, we have for any, $(\phi, \beta) \in \Theta$, $i \in [b]$

$$\nabla_\beta \log p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} \left[\nabla_\beta \log p(\mathrm{D}_i, z^{(i)} \mid \phi, \beta) \right] p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta) \mathrm{d}z^{(i)} ,$$

$$\nabla_\phi \log p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} \left[\nabla_\phi \log p(\mathrm{D}_i, z^{(i)} \mid \phi, \beta) \right] p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta) \mathrm{d}z^{(i)} ,$$

which suggests to consider

$$g_k^{(i)}(\phi, \beta) = \frac{1}{M} \sum_{m=1}^{M} \nabla_\beta \log p(Z_k^{(i,m)} \mid \beta) , \tag{4}$$

$$h_k^{(i)}(\phi, \beta) = \frac{1}{M} \sum_{m=1}^{M} \nabla_\phi \log p(\mathrm{D}_i \mid Z_k^{(i,m)}, \phi) , \tag{5}$$

where $M \in \mathbb{N}^*$ and $Z_k^{(i,1:M)} = (Z_k^{(i,m)})_{m \in [M]}$ are approximate samples from $p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta)$. More precisely, we consider a family $\{Q_{\gamma,\theta}^{(i)} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ where for any step-size γ, $Q_{\gamma,\theta}^{(i)}$ is a Markov kernel which targets a close approximation of $p(z^{(i)} \mid \mathrm{D}_i, \theta)$ with $\theta = \{\phi, \beta\}$. As an example, we can use overdamped Langevin dynamics [46, 52] to generate these samples. In this case, $Q_{\gamma,\theta}^{(i)}$ is associated with a Gaussian probability density function $q_{\gamma,\theta}^{(i)}(z^{(i)}, \cdot)$ with mean $z^{(i)} - \gamma \nabla_z \log p(z^{(i)} \mid \mathrm{D}_i, \theta)$ and variance $2\gamma \mathrm{I}_d$. Note that the number of Monte Carlo draws per iteration k is considered constant here but we can easily generalise our scheme to the non-constant setting. In addition, our scheme can also be generalised by taking into account *stochastic* gradient estimators of (4) and (5). For the sake of simplicity, we present our approach with standard gradients.

In this framework, we present the main steps of the corresponding stochastic approximation algorithm, called `FedSOUK`, in Algorithm 1. Since we consider the *cross-device* federated setting, note that only a random subset A_{k+1} of active (*i.e.* available) clients communicates with the central server at each iteration $k \in \mathbb{N}$. In addition, due to limited upload bandwidth, the potentially high-dimensional gradient estimator (5) is compressed locally via an unbiased stochastic compression operator $\mathscr{C}_{k+1}$ before being sent to the central server [1, 40].

Stateful and Stateless Versions. Depending on local memory constraints and the participation rate, we allow for a possible warm-start strategy across communication rounds to improve the convergence properties of the proposed algorithm so that the proposed algorithm becomes stateful, see steps 4 - 7 in Algorithm 1. In cases the participation rate is very low (e.g. each client might only participate once in the training process), we replace this warm-start strategy with the initialisation $Z_k^{(i,0)} \sim p(\cdot|\beta_k)$ for any $i \in [b]$ and $k \in \{0, ..., K-1\}$. This yields a stateless version of our algorithm more suitable to the cross-device setting. Obviously, compared to the previously proposed warm-start strategy, the performances of Algorithm 1 will be affected negatively if we are using the same number of local iterations M. We end up with an interesting trade-off between local computations and communication: if client-server communication is a bottleneck, the stateless version of the algorithm allows to reduce the communication overhead at the price of longer sampling procedures on each client. Such analyses will be illustrated empirically in Section 5.

Computation complexity. Compared to standard FL methods, our approach has an additional computational cost on the client side associated with Monte Carlo approximations $\{I_k^i\}$ and $\{J_k^i\}$. In practice, this cost is negligible. Indeed, in our experiments, we found that using a small value of $M \in [1, 10]$ was sufficient to obtain state-of-the-art results in terms of accuracy on the test dataset.

We would like to emphasize that this additional computational cost has also two side advantages compared to existing FL approaches: (1) it allows us to communicate less frequently with the central server and (2) it allows us to converge faster when the number of local iterations M increases since Monte Carlo approximation becomes better.

Communication overhead. As pointed out in Table 1, our methodology `FedPop` improves upon existing FL approaches regarding the communication overhead. Indeed, `FedPop` offers the flexibility to use both compression for sending updates to the server, and multiple local steps to reduce the communication frequency. As such, depending on the bandwidth and local computational power, the practitioner can adapt the number of local iterations M and the parameter of the compression operator. Up to our knowledge, this work is the first one combining compression and multiple local steps for personalized FL.

Algorithm 1 FL via Stochastic Optimisation using Unadjusted Kernel (`FedSOUK`)

Input: nb. outer iterations K, nb. local iterations M, Markov kernels $\{Q^{(i)}_{\gamma,\theta}\}_{\gamma,\theta,i}$, step-sizes $\{\eta^{(1)}_k, \eta^{(2)}_k\}_{k\in[K],i\in[b]}$ and initial points $Z^{(0)}_0 \in \mathbb{R}^d$, $\beta_0 \in \mathsf{B}$ and $\phi_0 \in \Phi$.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathsf{A}_{k+1}$ `// On active clients` A_{k+1} **do**
    `// Warm-start of the SA scheme if possible`
    Send $\{\beta_k, \phi_k\}$ to clients belonging to A_{k+1}.
    if $k \geq 1$ **then**
      Set $Z^{(i,0)}_k = Z^{(i,M)}_{k-1}$.
    end if
    `// Computation of key quantities using MCMC`
    for $m = 0$ **to** $M-1$ **do**
      Draw $Z^{(i,m+1)}_k \sim Q^{(i)}_{\gamma,\theta_k}\left(Z^{(i,m)}_k, \cdot\right)$.
      `// For Langevin dynamics`
      `// Draw` $\xi^{(i,m+1)}_k \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
      `// Set` $Z^{(i,m+1)}_k = Z^{(i,m)}_k + \gamma\nabla_z \log p(Z^{(i,m)}_k \mid \mathrm{D}_i, \phi_k, \beta_k) + \sqrt{2\gamma}\xi^{(i,m+1)}_k$.
    end for
    `// Communication with the server`
    Set $I^{(i)}_k = \frac{1}{M}\sum_{m=1}^{M} \nabla_\beta \log p\left(Z^{(i,m)}_k \mid \beta_k\right)$.
    Set $J^{(i)}_k = \frac{1}{M}\sum_{m=1}^{M} \nabla_\phi \log p\left(\mathrm{D}_i \mid Z^{(i,m)}_k, \phi_k\right)$.
    Send $I^{(i)}_k$ and $\mathscr{C}_{k+1}\left(J^{(i)}_k\right)$ to the central server.
  end for
  Set $\beta_{k+1} = \Pi_{\mathsf{B}}\left(\beta_k + \eta^{(1)}_{k+1}\left[\nabla_\beta \log p(\beta_k) + \frac{b}{|\mathsf{A}_{k+1}|}\sum_{i\in\mathsf{A}_{k+1}} I^{(i)}_k\right]\right)$.
  Set $\phi_{k+1} = \Pi_{\Phi}\left(\phi_k + \eta^{(2)}_{k+1}\left[\nabla_\phi \log p(\phi_k) + \frac{b}{|\mathsf{A}_{k+1}|}\sum_{i\in\mathsf{A}_{k+1}} \mathscr{C}_{k+1}\left(J^{(i)}_k\right)\right]\right)$.
end for
Output: $\{\phi_K, \beta_K\}$ and samples $\{Z^{(1:b,m)}_{K-1}\}_{m=1}^{M}$.

Robustness to client drift. For simplicity, we will take the example of `FedSOUL` (see Algorithm 1), which uses the Markov kernel associated with Langevin Monte Carlo to compute gradient estimates of the local marginal likelihood. However, the reasoning holds for general Markov kernels (adjusted or unadjusted). In this scenario, M steps of Langevin Monte Carlo are performed on each device to draw samples $\{Z^{(i,m)}_k\}$ used to compute Monte Carlo estimates $\{I^{(i)}_k\}$ and $\{J^{(i)}_k\}$. Increasing the number of local steps M does not slow down convergence but instead allows for more accurate Monte Carlo integration and hence better convergence properties. In contrast, the client drift phenomenon for classical FL approaches (e.g. `FedAvg` proposed in [39]) slows down convergence as the number of local iterations M increases.

Simple inference on new clients. Typical personalized FL approaches such as `DITTO` or `FedRep` require additional local training for inference on new clients. In contrast, the proposed methodology `FedPop` allows for a cheaper two-step approach once we have estimated $\theta^* = (\phi^*, \beta^*)$, as detailed below for a new client $b+1$ with feature vector x:

1. Sample $\{Z_{b+1}^{(l)}\}_{l\in[L]} \sim p(\cdot|\beta^*)$ in an i.i.d. manner.
2. Estimate the posterior predictive function $p(\cdot|x) \approx L^{-1}\sum_{l=1}^{L} p(\cdot|\phi^*, Z_{b+1}^{(l)}, x)$.

The prior $p(z_{b+1}|\beta^*)$ is typically chosen so that sampling is computationally cheap, e.g. a Gaussian with diagonal covariance matrix as in our experiments (see Section 5).

3 Theoretical Guarantees

In this section, we present non-asymptotic convergence guarantees for Algorithm 1 when the family of Markov kernels $\{Q_{\gamma,\theta}^{(i)} : \gamma \in (0,\bar{\gamma}], \theta \in \Theta, i \in [b]\}$ is associated to unadjusted, *i.e.* without Metropolis acceptance step, overdamped Langevin dynamics [14, 10]. The bounds we derive allow us to showcase explicitly the impact of FL constraints, namely partial participation and compression. Results for general unadjusted Markov kernels are postponed to the supplement.

To show our theoretical results and resort to standard assumptions made in the stochastic approximation literature, we consider a minimisation problem and rewrite the opposite of the objective function (3) for any $\theta \in \Theta$ as

$$f(\theta) = b^{-1}\sum_{i=1}^{b} f_i(\theta)\,, \quad \text{where } f_i(\theta) = -\log p(\phi) - \log p(\beta) - b\log p(\mathrm{D}_i \mid \phi, \beta)\,. \tag{6}$$

Non-Asymptotic Convergence Bounds. For the sake of better readability, we only detail in the main paper assumptions regarding the objective function, compression operators and the partial participation scenario. Technical assumptions related to the Markov kernels $\{Q_{\gamma,\theta}^{(i)}\}$ are postponed to the supplement. In spirit, we require, for any $i \in [b], \theta \in \Theta$ and γ, that $Q_{\gamma,\theta}^{(i)}$ satisfies some ergodic condition and can provide samples sufficiently close to the local posterior distribution $p(z^{(i)} \mid \mathrm{D}_i, \theta)$. For the sake of simplicity, we also assume that for any $k \in \mathbb{N}^*$, $\eta_k^{(1)} = \eta_k^{(2)} = \eta_k$, see Algorithm 1.

We make the following assumptions on Θ and the family of functions $\{f_i : i \in [b]\}$.

H1. *Θ is convex, closed subset of $\mathbb{R}^{d_\Theta}$ and $\Theta \subset \mathrm{B}(0, R_\Theta)$ for $R_\Theta > 0$.*

H2. *For any $i \in [b]$, the following conditions hold.*

(i) The function f_i defined in (S1) is convex.
(ii) There exist an open set $\mathsf{U} \in \mathbb{R}^{d_\Theta}$ and $L_f > 0$ such that $\Theta \subset \mathsf{U}$, $f_i \in \mathrm{C}^1(\mathsf{U}, \mathbb{R})$ and for any $\theta_1, \theta_2 \in \Theta$,

$$\|\nabla f_i(\theta_2) - \nabla f_i(\theta_1)\| \le L_f \|\theta_2 - \theta_1\|\,.$$

The assumption below requires compression operators $\{\mathscr{C}_k\}_{k\in\mathbb{N}^*}$ to be unbiased and to have a bounded variance. Such an assumption is for instance verified by stochastic quantisation operators, see Alistarh et al. [1].

H3. *The compression operators $\{\mathscr{C}_k\}_{k\in\mathbb{N}^*}$ are independent and satisfy the following conditions.*

(i) For any $k \in \mathbb{N}^*$, $v \in \mathbb{R}^d$, $\mathbb{E}[\mathscr{C}_k(v)] = v$.
(ii) There exists $\omega \ge 1$, such that for any $k \in \mathbb{N}^*$, $v \in \mathbb{R}^d$, $\mathbb{E}[\|\mathscr{C}_k(v) - v\|^2] \le \omega \|v\|^2$.

We finally assume that each client has probability $p \in (0, 1]$ to be active at each communication round. We would like to point out that this partial participation assumption can be associated with a specific compression operator satisfying **H3**.

H4. *For any $k \in \mathbb{N}^*$, $\mathsf{A}_k = \{i \in [b] : B_{i,k} = 1\}$ where for any $i \in [b]$, $\{B_{i,k} : k \in \mathbb{N}^*\}$ is a family of i.i.d. Bernouilli random variables with success probability $p \in (0, 1]$.*

Under these assumptions, the next result establishes that $(\bar{\theta}_k)_{k\in\mathbb{N}}$ defined by $\bar{\theta}_k = \sum_{j=1}^{k} \eta_j\theta_j / (\sum_{j=1}^{k} \eta_j)$ converges towards an element of $\arg\min_\Theta f$.

Theorem 1. *Assume **A1**-**H4** along with **A8** detailed in the supplement and let for any $k \in [K]$, $\eta_k \in (0, 1/L_f]$. Then, for any $K \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[f(\bar{\theta}_k) - f(\theta_\star)\right] \le \mathbb{E}\left[\frac{\sum_{k=1}^{K} \eta_k\{f(\theta_k) - f(\theta_\star)\}}{\sum_{k=1}^{K} \eta_k}\right] \le A(\gamma) + \frac{E_K}{\sum_{k=1}^{K} \eta_k}\,,$$

where E_K depends linearly on $(\omega/p)\sum_{k=1}^{K}\eta_k^2$; and $A(\gamma) = C\gamma^\alpha$ with $\alpha > 0$ and C is independent of ω, p and (η_k). Closed-form formulas for these constants are provided in the supplement.

An interesting feature of Algorithm 1 is that convergence towards a minimum of f is possible and the impact of partial participation and compression vanishes when $\lim_{k\to\infty}\eta_k = 0$. More precisely, $\limsup_{k\to\infty} E_K/(\sum_{k=1}^{K}\eta_k) = 0$ and $\lim_{\gamma\to 0^+} A(\gamma) = 0$ which shows that we can tend towards a minimum of f with arbitrary precision $\epsilon > 0$ by setting the step-size γ to a small enough value.

4 Related Works

As pointed out in Section 1, many different approaches have been proposed to address personalisation and uncertainty quantification under the federated learning paradigm. This section reviews the main related existing lines of research and shows that the proposed methodology provides many benefits; see Table 1. Interestingly, we also show that `FedPop` encompasses some of the existing FL models.

Bayesian FL. One of our main motivations is the possibility to perform grounded uncertainty quantification in FL by resorting to the Bayesian paradigm. In recent years, many works have suggested adapting serial workhorse stochastic simulation approaches such as MCMC or variational inference to the FL setting [7, 36, 35, 50, 15, 9, 5, 42, 12]. Although some of these approaches address important FL challenges such as the communication bottleneck, partial participation, or limited computational device resources, they are not suitable for uncertainty quantification in the cross-device FL scenario. Indeed, all these approaches assume that the posterior distribution targeted by each client is parametrised by a single potentially high-dimensional parameter of size $d_\Phi + d$, see (1). This prevents a sufficient number of samples from being stored locally to perform uncertainty quantification and Bayesian model averaging, especially when the model is a large neural network. In contrast, our approach decouples this unique high-dimensional parameter into a fixed part ϕ and a low-dimensional random part $z^{(i)}$, significantly reducing the memory footprint of local sample storage.

In addition, Bayesian FL methods aim at sampling a random parameter θ from a target probability distribution $\pi(\theta) \propto e^{-f(\theta)}$ where $f(\theta) = (1/b)\sum_{i=1}^{b} f_i(\theta)$ with f_i denoting the negative log-likelihood associated to the i-th client. On the other hand, the proposed methodology `FedPop` considers a mixed-effects modeling approach where parameters are divided into two categories: a fixed component ϕ and a random one z_i for each client $i \in [b]$. As such, the mixed-effects approach is in essence an empirical Bayesian/marginal likelihood approach (see [6, 44]). It corresponds to a hierarchical model that aims to combine the modeling flexibility and uncertainty assessment of Bayesian inference with computational pragmatism. More precisely, a part of the parameters (fixed-effects ϕ) is estimated via marginal likelihood maximization and the rest (random effects z_i) using common Bayesian techniques, which are in most cases low dimensional.

As a result, up to our knowledge, the model and approach that we proposed is novel in FL and comes with many benefits as shown in Table 1.

Personalised FL. Besides uncertainty quantification, we also aim at providing each client with a dedicated personalised model. Among the numerous existing personalised FL approaches, those related to `FedPop` can be broadly classified into two groups: *meta-learning* and *partially local methods*. Meta-learning based FL methods aim at training a global model conducive to fast training of personalised models. Such a goal can be achieved, for example, by local fine-tuning [16], regularisation of local models towards their average [19, 21] – or the opposite [33], and model interpolation [34]. On the other hand, FL methods based on partial decoupling take an approach similar to ours by splitting the initial model into a backbone component and a local one aimed at personalisation [8, 2, 41]. This partial decoupling could also enhance privacy as discussed in Singhal et al. [47]. The main difference with `FedPop` is that such approaches based on empirical risk minimisation cannot provide credible information.

`FedPop`**: A Compromise between Standard and Personalised FL.** Interestingly, we show here that the `FedPop` framework allows existing FL approaches to be retrieved in certain regimes. To this end, we assume that the prior $p(z^{(i)} \mid \beta)$ is Gaussian with mean μ and covariance matrix $\sigma^2 \mathrm{I}_d$ so that $\beta = \{\mu, \sigma\}$. If $\sigma \to 0^+$, then this Gaussian prior tends towards the Dirac distribution centered at μ and the local likelihood becomes $p(\mathrm{D}_i \mid \phi, \mu)$, which corresponds to the local objective of standard FL approaches such as `FedAvg` [39]. On the other hand, when $\sigma \to \infty$, no common information β is

Table 1: Overview of the main existing personalised FL (top rows) and Bayesian FL (bottom rows) approaches related to the proposed framework. Column "PP" refers to partial participation, "perso." to personalised approaches, "bounds" to available convergence guarantees, "UQ" to available uncertainty quantification, "com." to the scheme (multiple local steps and/or compression) used to address the communication bottleneck and "memory" to the client memory footprint where M stands for the number of samples.

METHOD	PP	PERSO.	BOUNDS	UQ	COM.	MEMORY	FEDPOP INSTANCE
PER-FEDAVG	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
PFEDME	✗	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
FEDREP	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✓
DITTO	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
LG-FEDAVG	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
QLSD	✓	✗	✓	✓	COMPRESSION	$M(d + d_\Phi)$	✗
FSGLD	✗	✗	✓	✓	LOCAL STEPS	$M(d + d_\Phi)$	✗
FEDBE	✓	✗	✗	✓	LOCAL STEPS	$M(d + d_\Phi)$	✗
DG-LMC	✗	✗	✓	✓	LOCAL STEPS	$M(d + d_\Phi)$	✓
FEDPOP	✓	✓	✓	✓	BOTH	$Md + d_\Phi$	–

used to locally regress $z^{(i)}$ and we end up with the FedRep algorithm [8]. This shows that FedPop stands for a subtle compromise between standard and personalised FL which should benefit clients with small data sets by pooling information via a common prior. Finally, in the extreme scenario where ϕ is the null vector, our approach amounts to the Bayesian FL approach DG-LMC proposed in Plassier et al. [42].

5 Numerical Experiments

In this section, we illustrate the benefits of our methodology on several FL benchmarks associated with both synthetic and real data. Since existing Bayesian FL approaches are not suited for personalisation (see Table 1), we only compare the performances of Algorithm 1 with personalised FL methods. In all our experiments, we use overdamped Langevin dynamics to sample locally and call this specific instance of Algorithm 1, FedSOUL. In addition, we set $p(z^{(i)} \mid \beta) = \mathrm{N}(\mu, \sigma^2 \mathrm{I}_d)$ with $\beta = \{\mu, \sigma\}$ for simplicity. To be comparable with existing personalised FL approaches that only consider periodic communication via multiple local steps, we do not resort to the proposed compression mechanism although the latter could be of interest for real-world applications. Additional experiments and details about experimental design are provided in the supplement.

Synthetic Data. We start by showcasing the benefits of FedSOUL for clients having small and highly heterogeneous data sets as pointed out in Section 1 and Section 2. To this end, we consider a similar experimental setting as in Collins et al. [8] where synthetic observations $\{y_j^{(i)}\}_{j \in [N_i]} \in \mathrm{D}_i$ are generated via the following procedure: $x_j^{(i)} \sim \mathrm{N}(0_k, \mathrm{I}_k)$ and $y_j^{(i)} \sim \mathrm{N}(z_{\text{true}}^{(i)} \phi_{\text{true}}^\top x_j^{(i)}, 0.1)$. The ground-truth parameters $z_{\text{true}}^{(i)} \in \mathbb{R}^d$ and $\phi_{\text{true}} \in \mathbb{R}^{k \times d}$ have been randomly generated beforehand with $(d, k) = (2, 20)$. Compared to Collins et al. [8], we use heterogeneous data partitions across clients so that 90% of the $b = 100$ clients have small data sets of size 5 and the remaining 10% have data sets of size

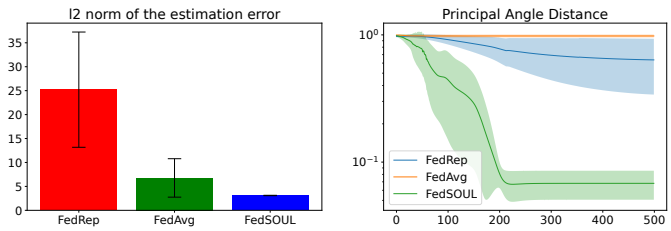

Figure 2: Small data sets - synthetic data.

10. We compare our results with FedRep [8] and FedAvg [39] since they stand for two limiting instances of the proposed methodology, see Section 4 and Gelman and Hill [17, Section 12]. Figure 2 compares the different approaches by computing the principle angle distance[1] (respectively the ℓ_2 norm) between ϕ_{true} (respectively $z^{(i)}_{\text{true}}$) and its estimated value; the lesser the better. In contrast to its main competitors and based on both metrics, FedSOUL provides an impressive improvement. This illustrates the benefits of the introduction of a common prior $p(z^{(i)} \mid \beta)$ which allows for preventing overfitting on clients with small data sets while performing personalisation. Additional results with other choices for (b, d, k) and data partitioning strategies are available in the supplement.

Moreover, to compare our algorithm with a non-FL setting, we perform a non-distributed and non-federated stochastic approximation algorithm to find θ^* using a large number of iterations to get an accurate approximation of the optimal parameter θ^*. Then, we use FedPop to obtain an estimate $\tilde{\theta}^*$ and measure the relative error in l_2- distance between θ^* and $\tilde{\theta}^*$. For some outer iterations $T = 100$, the relative error was less than 10^{-3}, which illustrates the relevance of our theoretical results. We also test the performances of the proposed approach when the warm-start strategy is not used. In this case, we have to set $M = 50$ to achieve the same performances as in the stateful variant of FedSOUL.

Real Data. We consider now real image data sets, namely CIFAR-10 and CIFAR-100 [29]. For our likelihood model defined by $p(\mathrm{D}_i \mid \phi, z^{(i)})$, we use 5-layer convolutional neural networks and perform personalisation for the last layer. We set $b = 100$ for convenience and control data heterogeneity by assigning to each client N_i images belonging to only S different classes.

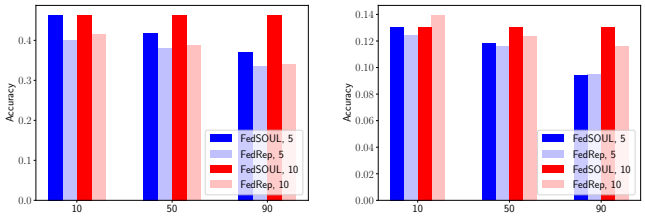

Figure 3: (right) CIFAR-10 with $S = 5$ and (left) CIFAR-100 with $S = 20$. The x-axis refers to the percentage of clients having $N_i \in \{5, 10\}$ images.

Small data sets. Under this setting, we first consider (10%, 50%, 90%) of clients having small data sets of size either $N_i = 5$ or $N_i = 10$; while remaining clients have larger data sets of size $N_i = 25$. We compare our approach with FedRep since it stands for the state-of-the-art personalised FL approach. The algorithms are trained to fulfill the same computational budget. Figure 3 shows the average accuracy across clients for the two approaches on both CIFAR-10 and CIFAR-100. We can see that FedSOUL is consistently better than FedRep over different configurations.

Full data sets. In addition, to show that the proposed approach achieves state-of-the-art performances on small data sets (which is common in the cross-device scenario), we now illustrate that FedSOUL is also competitive on larger data sets. To this end, we use all training images in CIFAR-10 and CIFAR-100 image data sets and consider the same data partitioning as in Collins et al. [8]. More precisely, in this case, the number of observations and the number of classes per client are uniformly shared among the clients. Table 2 shows our results in comparison with state-of-the-art personalised FL approaches. We can see that our model outperforms other methods on both CIFAR-10 and CIFAR-100 by a large margin. Additional results with other personalised FL algorithms are postponed to the supplement.

Uncertainty Quantification on Real Data. As highlighted in Table 1, one advantage of the proposed approach compared to existing personalised FL methods is the ability to perform uncertainty quantification by sampling locally from the posterior $p(z^{(i)} \mid \mathrm{D}_i, \phi_K, \beta_K)$, see Algorithm 1. We illustrate this feature by computing CIFAR-10 calibration curves and scores (*e.g.* expected calibration error aka ECE) on a specific client; and by performing an out-of-distribution analysis on MNIST/FashionMNIST data sets. Figure 4 shows that the proposed approach provides relevant uncertainty diagnosis. Additional results on uncertainty quantification can be found in the supplement.

[1] defined in [8, Definition 1]

Table 2: Real data - Full data sets. Accuracy (in %) on test samples. FedAvg and SCAFFOLD are not personalised FL approaches but stand for well-known FL benchmarks.

	CIFAR-10		CIFAR-100	
(# clients b, # classes per client S)	(100, 2)	(100, 5)	(100, 5)	(100, 20)
Local learning only	89.79	70.68	75.29	41.29
FedAvg [39]	42.65	51.78	23.94	31.97
SCAFFOLD [27]	37.72	47.33	20.32	22.52
LG-FedAvg [34]	84.14	63.02	72.44	38.76
Per-FedAvg [16]	82.27	67.20	72.05	52.49
L2GD [19]	81.04	59.98	72.13	42.84
APFL [13]	83.77	72.29	78.20	55.44
DITTO [33]	85.39	70.34	78.91	56.34
FedRep [8]	87.70	75.68	79.15	56.10
FedSOUL (this paper)	91.12	79.48	79.56	59.73

6 Conclusion

In this paper, we proposed a general Bayesian methodology based on a natural mixed-effects modeling approach to model personalisation in federated learning. Our FL method is the first that allows for both personalisation and cheap uncertainty quantification for (cross-device) federated learning. By introducing a common prior on the local parameters, we tackle the local overfitting problem in the scenario where clients have highly heterogeneous and small data sets. In addition, we have shown that the proposed approach has favorable convergence properties. Some limitations of FedPop pave the way for more advanced personalised FL approaches. As an example, our model does not allow for training heterogeneous architectures across clients because of the introduced common prior and only satisfies first-order privacy guarantees. Regarding the latter, further works include for instance deriving differentially private versions of our framework.

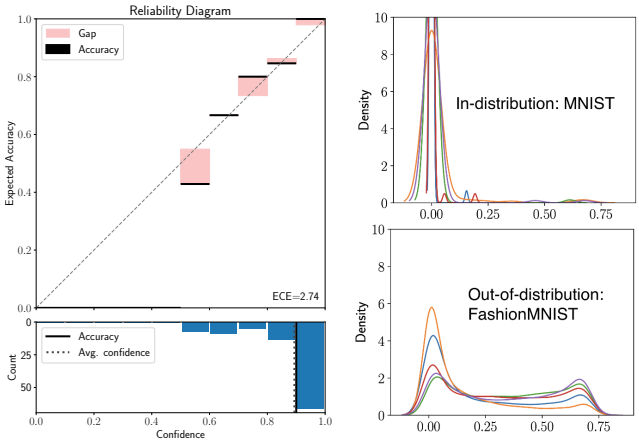

Figure 4: (right) Calibration on CIFAR-10 for a specific client and (left) OOD analysis with MNIST training & FashionMNIST inference – one curve corresponds to one client.

Acknowledgments and Disclosure of Funding

The authors acknowledge the Lagrange Mathematics and Computing Research Center for supporting the project. The development of the algorithm and conducting experiments (Section 5) was supported by Russian Science Foundation grant 20-71-10135.

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 2017.

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated Learning with Personalization Layers. *arXiv preprint arXiv:1912.00818*, 2019.

[3] Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

[4] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[5] Thang D. Bui, Cuong V. Nguyen, Siddharth Swaroop, and Richard E. Turner. Partitioned Variational Inference: A unified framework encompassing federated and continual learning. *arXiv preprint arXiv:1811.11206*, 2018.

[6] George Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[7] Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2020.

[8] 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.

[9] Luca Corinzia, Ami Beuret, and Joachim M. Buhmann. Variational Federated Multi-Task Learning. *arXiv preprint arXiv:1906.06268*, 2019.

[10] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society, Series B*, 79(3):651–676, 2017.

[11] Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F. Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Statistics and Computing*, 31(3), 2021.

[12] Wei Deng, Yi-An Ma, Zhao Song, Qian Zhang, and Guang Lin. On Convergence of Federated Averaging Langevin Dynamics. *arXiv preprint arXiv:2112.05120*, 2021.

[13] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive Personalized Federated Learning, 2021.

[14] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238.

[15] Khaoula El Mekkaoui, Diego Mesquita, Paul Blomstedt, and Samuel Kaski. Distributed stochastic gradient MCMC for federated learning. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

[16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *Advances in Neural Information Processing Systems*, 2020.

[17] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

[18] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. In *International Conference on Learning Representations*, 2018.

[19] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

[20] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint arXiv:2010.02372*, 2020.

[21] Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv: 2102.09743*, February 2021.

[22] Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

[23] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction . *arXiv preprint arXiv:1904.05115*, 2019.

[24] Yihan Jiang, Jakub Konevcnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

[25] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, 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 Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, 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. ISSN 1935-8237. doi: 10.1561/2200000083.

[26] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[27] 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.

[28] 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.

[29] Alex Krizhevsky. Learning multiple layers of features from tiny images. Available at http://www.cs.toronto.edu/~kriz/cifar.html, 2009.

[30] Marc Lavielle. *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, 2014.

[31] Yann LeCun, Sumit Chopra, Raia Hadsell, Fu Jie Huang, and et al. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

[32] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020. URL https://proceedings.mlsys.org/paper/2020/file/38af86134b65d0f10fe33d30dd76442e-Paper.pdf.

[33] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368, 2021. URL http://proceedings.mlr.press/v139/li21h.html.

[34] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. In *NeurIPS 2019 Workshop on Federated Learning*, 2019.

[35] Dongzhu Liu and Osvaldo Simeone. Channel-Driven Monte Carlo Sampling for Bayesian Distributed Learning in Wireless Data Centers. *IEEE Journal on Selected Areas in Communications*, 2021.

[36] Dongzhu Liu and Osvaldo Simeone. Wireless Federated Langevin Monte Carlo: Repurposing Channel Noise for Bayesian Sampling and Privacy. *arXiv preprint arXiv:2108.07644*, 2021.

[37] Jeffrey Long. *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Sage Publications, Inc, 2011.

[38] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. 2020.

[39] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[40] Constantin Philippenko and Aymeric Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees . *arXiv preprint arXiv:2006.14591*, 2020.

[41] Krishna Pillutla, Kshitiz Malik, Abdelrahman Mohamed, Michael Rabbat, Maziar Sanjabi, and Lin Xiao. Federated Learning with Partial Model Personalization, 2022. URL https://openreview.net/forum?id=iFf26yMjRdN.

[42] Vincent Plassier, Maxime Vono, Alain Durmus, and Eric Moulines. DG-LMC: a turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs. In *International Conference on Machine Learning (ICML)*, 2021.

[43] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 09 1951. doi: 10.1214/aoms/1177729586.

[44] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in statistics*, pages 388–394. Springer, 1992.

[45] C. P. Robert. *The Bayesian Choice: from decision-theoretic foundations to computational implementation*. Springer, New York, 2 edition, 2001.

[46] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[47] Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11220–11232. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/5d44a2b0d85aa1a4dd3f218be6422c66-Paper.pdf.

[48] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

[49] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–17, 2022. doi: 10.1109/TNNLS.2022.3160699.

[50] Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, and Eric Moulines. QLSD: Quantised Langevin stochastic dynamics for Bayesian federated learning. In *AISTATS*, 2022.

[51] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A Field Guide to Federated Optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[52] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning*, 2011.

[53] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e1021d43911ca2c1845910d84f40aeae-Paper.pdf.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section ...
- Did you include the license to the code and datasets? [Yes] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Some assumptions are explicitly stated in Section 3. For the sake of readability and due to the space limit, we postponed other technical assumptions in the supplement.
   (b) Did you include complete proofs of all theoretical results? [Yes] All our proofs are postponed to the supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplement.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Some details have been stated in Section 5 in the main paper. Because of the limited number of pages, others have been postponed to the supplement

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We did it for some experiments we conducted, see e.g. Figure 2.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is postponed to the supplement.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] See for example the citation for CIFAR image data sets

   (b) Did you mention the license of the assets? [No] We use only open source code and datasets.

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]