Proceedings of Machine Learning Research vol 211:1–13, 2023 5th Annual Conference on Learning for Dynamics and Control

FedSysID: A Federated Approach to Sample-Efficient System Identification

Han Wang HW2786@COLUMBIA.EDU
Columbia University, New York, NY

Leonardo F. Toso LT2879@COLUMBIA.EDU
Columbia University, New York, NY

James Anderson JAMES.ANDERSON@COLUMBIA.EDU
Columbia University, New York, NY

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

We study the problem of learning a linear system model from the observations of M clients. The catch: Each client is observing data from a different dynamical system. This work addresses the question of how multiple clients collaboratively learn dynamical models in the presence of heterogeneity. We pose this problem as a federated learning problem and characterize the tension between achievable performance and system heterogeneity. Furthermore, our federated sample complexity result provides a constant factor improvement over the single agent setting. Finally, we describe a meta federated learning algorithm, FedSysID, that leverages existing federated algorithms at the client level.

Keywords: Federated learning; System identification; System heterogeneity

1. Introduction

The system identification problem aims to estimate the parameters of a dynamical system from observed data. The data can take many forms, however, we focus on data sets comprised of a given sequence of input-state trajectories. System identification plays a crucial role in a diverse set of applications including time-series analysis, control theory, robotics, and reinforcement learning. Motivated by trends in big data and the explosion in IoT devices, there has been an increasing interest in finite-sample complexity and non-asymptotic analysis. This line of works can broadly be divided into two categories: i) learning fully observed linear time-invariant (LTI) systems (Sarkar and Rakhlin, 2019; Dean et al., 2020; Simchowitz et al., 2018); ii) learning partially observed LTI systems (Oymak and Ozay, 2019; Sun et al., 2020; Zheng and Li, 2020; Simchowitz et al., 2019). Despite the mature body of the literature on the finite-sample properties for system identification, it appears that a non-asymptotic convergence rate of $\frac{1}{\sqrt{N}}$, where N is the number of observed trajectories, is a hard limit. Such dependence on the number of trajectories places a limit on the scalability of standard techniques for applications where sampling different trajectories is a distinctly difficult and costly task (e.g., large-scale and safety-critical systems).

To circumvent this apparent fundamental sample-efficiency limit, we propose a modified problem setting where multiple similar agents, orchestrated by a central server, collaborate

to learn, from a broad data set, a common model that fits all of them well. To achieve this, we leverage the popular federated learning (FL) paradigm (Konečnỳ et al., 2016a; McMahan et al., 2017b). Federated learning (FL) is a machine-learning setting where multiple clients collaborate to train a model under the coordination of a central server. By allowing data to be stored client-side, the FL paradigm has many favorable privacy properties and achieves almost the same performance guarantees as centralized methods. Canonical examples of FL include building treatment outcome models from patients' medical records (without sharing the sensitive data) and creating predictive text models from an ensemble of cell phone users. Compared to traditional distributed optimization, one key challenge and characteristic of FL is heterogeneity; both in an *objective and statistical* sense. In the system identification problem we pose, this heterogeneity manifests through the fact that each client observes a similar (but not identical) dynamical system.

We formulate an offline federated system identification problem. Our problem assumes a central server connected to M clients. The objective is to identify a linear dynamical system model based on observational data held locally by the M clients. The catch - clients observe trajectories from different systems. Specifically, each client observes N_i trajectories generated by a system, with the assumption that all the systems are "similar" to each other. Motivating examples of this scenario include modeling multiple autonomous vehicles that operate in similar dynamic environments, based on observations from a fleet of identical vehicles. Our objective is to answer the following questions: i) *Through collaboration, can each client achieve an improved finite-time convergence guarantee?* ii) *How does the heterogeneity of the observed systems influence the convergence rate?*

Contributions

- **Problem formulation:** To the best of our knowledge, this is the first paper to use the FL framework for system identification. We consider the case where clients observe similar, but not identical, dynamical systems. Although we demonstrate our solution technique by focusing on the problem of learning fully observed linear dynamical systems from multiple trajectories, our framework can be easily extended to analyze system identification problems with single trajectory, from partially observed linear dynamical systems (Oymak and Ozay, 2019; Wang and Anderson, 2022), or sparse dynamical system models (Fattahi et al., 2019). The latter can be solved via `FedADMM` (Wang et al., 2022a).

- **Improved convergence rate:** Our solution provides a convergence rate that is a constant factor better than if a single agent tries to learn a model independently. In particular, the convergence rate is improved by a factor of $\sqrt{M}$, where M is the number of clients, if each client has the same amount of local data. As the number of agents increases, the benefit from collaboration increases. Since the number of clients in FL is usually large, the benefit from collaboration always shows up.

- **Influence of heterogeneity:** We assume that the systems participating in the collaboration are similar but not identical. We provide a theoretical analysis of how the dissimilarity amongst those systems influences the non-asymptotic convergence rate of the proposed technique.

- **Meta Algorithm:** We present `FedSysID`, a federated learning algorithm to solve the proposed system identification problem. `FedSysID` can call any FL optimization method as a subroutine. Furthermore, we analyze the convergence property of this algorithm for the specific case of `FedAvg` (McMahan et al., 2017b) and `FedLin` (Mitra et al., 2021).

Notation: Given a matrix $A \in \mathbb{R}^{m \times n}$, the Frobenius norm of A is denoted by $||A||_F = \sqrt{Tr(AA^\top)}$. $||A||$ corresponds to the spectral norm of A, namely, $||A|| = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the largest singular value of A. Consider a symmetric matrix Σ, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote its minimum and maximum eigenvalues, respectively.

2. Related Work

System identification: Classically, system identification methods provided asymptotic results, see for example (Ljung, 1998). Unfortunately, these results prove consistency using the central limit theorem and law of large numbers type theorems, thus requiring the number of data samples to tend to infinity. Finite-time (i.e., non-asymptotic) analysis of system identification problems is now an active area of research. Subspace methods (Verhaegen and Verdult, 2007; Van Overschee and De Moor, 2012) are often widely used in practice, however, recent work (Jedra and Proutiere, 2020; Sarkar and Rakhlin, 2019; Simchowitz et al., 2018) has shown that the least squares estimator (that we consider) is nearly-optimal in the fully-observed setting using finite-time analysis. System theoretic properties quantifying controllability (Dean et al., 2020; Tsiamis and Pappas, 2021) and stability (Sarkar and Rakhlin, 2019) frequently appear in sample complexity bounds. Broadly, identification techniques can be broken down into two categories, depending on whether they require a single trajectory (most suitable for stable systems) or multiple trajectories (suitable for unstable, requires the ability to reset the system). Upper and lower bounds on the sample complexities for such algorithms are concisely reviewed in (Tsiamis et al., 2022). Specifically, various algorithms are shown to have an $\mathcal{O}(\frac{1}{\sqrt{N}})$ upper bound on the sample complexity. Our work is motivated by the results of (Xin et al., 2022), who introduced the idea of learning system dynamics by leveraging data from a similar system. However, the authors do not formulate the system identification problem in a federated manner. Moreover, they did not establish a connection between their results and the centralized solution where the system has access to the data sets from all the systems.

Federated learning (FL): FL is a *distributed* machine learning framework with the objective of training a model from multiple client's data (McMahan et al., 2017a; Kairouz et al., 2021). Motivated by the proliferation of Internet of Things (IoT) devices, FL aims to: i) *provide privacy* by ensuring data does not leave the client device (Truex et al., 2019); ii) *reduce communication costs* by reducing the volume of communication between the clients and the server (Konečnỳ et al., 2016b); iii) gracefully handle *client dropout* (Tran Dinh et al., 2021; Wang et al., 2022a). Typical FL applications comprise of a massive number of clients. Not all clients may be able to fully participate, for example, a cell phone may lose signal and be forced to drop out; iv) handle data heterogeneity (Sahu et al., 2018) – the distributions characterizing client data sets is likely to vary. Although peer-to-peer model exists in the architecture of FL (Yang et al., 2019), we focus on the classical server-client

model where a single server periodically aggregates data from its clients. Clients maintain their own data and cannot communicate directly with each other. Numerous FL algorithms have been developed and analyzed. While our algorithm is independent of a specific FL implementation, we highlight results using the classical `FedAvg` (McMahan et al., 2017b) and the more recent `FedLin` (Mitra et al., 2021) which can handle data heterogeneity.

3. Centralized System Identification

Given state trajectory data, the centralized system identification problem is to learn a linear, time-invariant invariant (LTI) system model that best fits the observations. The LTI model takes the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 0, 1, 2, \ldots, T-1 \tag{1}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$ and $w_t \in \mathbb{R}^n$ are the state, input, and process noise of the system at time t, respectively. Moreover, it is assumed that $\{u_t\}_{t=1}^{\infty}, \{w_t\}_{t=1}^{\infty}$ are independent and identically distributed random variables, namely, $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_u^2 I_p\right)$ and $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_w^2 I_n\right)$. It is further assumed that the initial state $x_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_x^2 I_n\right)$. The state-input trajectory pair $\{x_t, u_t\}$ from a single experiment is referred to as a *rollout*. Multiple rollouts of length T are allowed, with the resulting data set stored as $\{x_{l,t}, u_{l,t}\}_{t=0}^{T-1}$, for $l = 1, \ldots N$, where l denotes l-th rollout and t denotes t-th time-step in the rollout. We work from the assumption that the data set is generated by (1) and the objective is to provide estimates $(\hat{A}, \hat{B})$ of best fit. We now revisit the standard least-square estimation procedure.

3.1. Least-Square Procedure

Denote $\Theta \triangleq \begin{bmatrix} A & B \end{bmatrix} \in \mathbb{R}^{n \times (n+p)}$ where A and B correspond to the ground truth system (1) that generated the data. The augmented state input pair of rollout l at time t, is $z_{l,t} \triangleq \begin{bmatrix} x_{l,t} \\ u_{l,t} \end{bmatrix} \in \mathbb{R}^{n+p}$. Therefore, the state update $x_{l,t+1}$ can be written as

$$x_{l,t+1} = \Theta z_{l,t} + w_{l,t} \quad \forall\ 1 \le l \le N \text{ and } 0 \le t \le T-1,$$

where $x_{l,t}$ can be expanded recursively as follows

$$x_{l,t} = G_t \begin{bmatrix} u_{l,0} \\ \vdots \\ u_{l,t-1} \end{bmatrix} + F_t \begin{bmatrix} w_{l,0} \\ \vdots \\ w_{l,t-1} \end{bmatrix} + A^t x_{l,0}, \quad t = 1, 2, \ldots, T-1$$

with $G_t \triangleq \begin{bmatrix} A^{t-1}B & A^{t-2}B & \cdots & B \end{bmatrix}$ and $F_t \triangleq \begin{bmatrix} A^{t-1} & A^{t-2} & \cdots & I_n \end{bmatrix}$ for all $t \ge 1$.

Lemma 1 *Let $u_{l,t}, w_{l,t}$, and $x_{l,0}$ be drawn from i.i.d. Gaussian distributions. It holds for $z_{l,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_t)$, and for all $t \ge 1$, that*

$$\Sigma_t \triangleq \begin{bmatrix} \sigma_u^2 G_t (G_t)^\top + \sigma_w^2 F_t (F_t)^\top + \sigma_{i,x}^2 A^t (A^t)^\top & 0 \\ 0 & \sigma_u^2 I_p \end{bmatrix} \succ 0.$$

Proof *All proofs can be found in Wang et al. (2022b) (the extended version of this paper).* ■

For each rollout l, the data is concatenated as $X_l = \begin{bmatrix} x_{l,T} & \cdots & x_{l,1} \end{bmatrix} \in \mathbb{R}^{n\times T}$, $Z_l = \begin{bmatrix} z_{l,T-1} & \cdots & z_{l,0} \end{bmatrix} \in \mathbb{R}^{(n+p)\times T}$, and $W_l = \begin{bmatrix} w_{l,T-1} & \cdots & w_{l,0} \end{bmatrix} \in \mathbb{R}^{n\times T}$. We further stack the data from all the rollouts to obtain the batch matrices $X = \begin{bmatrix} X_1 & \dots & X_N \end{bmatrix} \in \mathbb{R}^{n\times NT}$, $Z = \begin{bmatrix} Z_1 & \cdots & Z_N \end{bmatrix} \in \mathbb{R}^{(n+p)\times NT}$, and $W = \begin{bmatrix} W_1 & \cdots & W_N \end{bmatrix} \in \mathbb{R}^{n\times NT}$. Pulling it all together, the state, input noise, and model parameters are related via the system of linear equations,

$$X = \Theta Z + W.$$

From this system description, an estimate of system dynamics can be obtained by solving the unconstrained least-square problem:

$$\hat{\Theta} \triangleq \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} = \arg\min_{\Theta\in\mathbb{R}^{n\times(n+p)}} \|X - \Theta Z\|_F^2. \tag{2}$$

Proposition 2 [1] *(Dean et al., 2020) Assume we collect data from the linear, time-invariant system initialized at $x_0 = 0$, using inputs $u_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_u^2 I_p\right)$ for $t = 1, \cdots, T$. Suppose $w_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_w^2 I_n\right)$ and $N \geq 8(n+p) + 16\log(4/\delta)$. Thus, with probability at least $1-\delta$, the least squares estimator using only the final sample of each trajectory satisfies*

$$\max\left\{\|\hat{A} - A\|, \|\hat{B} - B\|\right\} \leq \frac{16\sigma_w}{\sqrt{\lambda_{\min}\left(\Sigma_{T-1}\right)}}\sqrt{\frac{(n+2p)\log(36/\delta)}{N}}.$$

The above proposition indicates that the estimation error for the least-squares method scales as $\mathcal{O}(N^{-\frac{1}{2}})$. However, the magnitude of N is typically small due to the difficulty and cost of collecting samples, especially for large-scale and safe-critical systems, leading us to a large estimation error. In order to address this issue of insufficient data volume for system identification, we exploit the federated learning (FL) framework, where $N \triangleq \sum_{i=1}^{M} N_i$ rollouts are generated from similar, but *not idential* LTI systems, that each generates N_i rollouts. Moreover, the number of clients M is usually large, which enables each client to borrow other clients' data indirectly in order to achieve a better sample efficiency, e.g. $\mathcal{O}(\frac{1}{\sqrt{\sum_{i=1}^{M} N_i}})$, by participating in FL and benefiting from the participation of other clients.

4. Federated System Identification

We now pose the federated system identification problem, introduce `FedSysID` - a federated learning algorithm for system identification, and analyze its performance. Let there be M clients collaborating to solve the system identification problem under the coordination of a central server. We consider a client-server model of computation, i.e., there is no direct communication between clients, and communication between the central server and client only happens periodically. Each client observes and collects data from a separate dynamical system that is assumed to follow the LTI system dynamics,

$$x_{t+1}^{(i)} = A^{(i)} x_t^{(i)} + B^{(i)} u_t^{(i)} + w_t^{(i)} \tag{3}$$

1. It is important to remark that (Dean et al., 2020) consider the last time-step of Σ_t with $x_0 = 0$ for the sample complexity characterization, whereas, in our work, we include the information of Σ_t from all time steps with $x_0^{(i)} \sim \mathcal{N}\left(0, \sigma_{i,x}^2 I_n\right)$ to characterize the non-asymptotic convergence rate of our federated system identification approach.

for $i \in [M]$. To reduce notation, we assume that each system is observed for T time steps. All the results presented carry over to the case where each client collects measurements over differing time horizons. Client i performs N_i rollouts of length T where they excite their systems with an i.i.d. Gaussian input $u_t^{(i)} \sim \mathcal{N}\left(0, \sigma_{i,u}^2 I_p\right)$, process noise $w_t^{(i)} \sim \mathcal{N}\left(0, \sigma_{i,w}^2 I_n\right)$, and an initial state $x_0^{(i)} \sim \mathcal{N}\left(0, \sigma_{i,x}^2 I_n\right)$. The resulting data set for each client is defined as $\left\{x_{l,t}^{(i)}, u_{l,t}^{(i)}\right\}_{t=0}^{T-1}$, for $l = 1, \dots N_i$.

The objective of the federated system identification problem is to find a common estimation of the system matrices $\bar{\Theta} = \begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix}$ which provides a small estimation error with respect to each client i's true system $\Theta^{(i)} = \begin{bmatrix} A^{(i)} & B^{(i)} \end{bmatrix}$. However, each client i maintains its own data (without sharing it) and can estimate its unknown system dynamics $(A^{(i)}, B^{(i)})$ via the least-square procedure (2) described previously. We aim to answer the question: *Can clients in the FL setting achieve a lower estimation error and a better sample complexity cost compared to the centralized setting?* To answer this question, we first need to make the following heterogeneity assumption regarding each client's system dynamics.

Assumption 1 *(Bounded System Heterogeneity) There exists a constant ϵ such that*

$$\max_{i,j \in [M]} \|A^{(i)} - A^{(j)}\| \leq \epsilon, \text{ and } \max_{i,j \in [M]} \|B^{(i)} - B^{(j)}\| \leq \epsilon, \quad \text{holds for all } i, j \in [M].$$

At the heart of the federated system identification problem is determining when collaboration helps the identification. Intuitively, when clients observe very similar systems, it makes sense to collaborate. For wildly divergent systems, the collaboration will be detrimental. One of our objectives is to provide a more subtle characterization of the effect of system heterogeneity on identification performance. We use ϵ to characterize this heterogeneity, we will see that it explicitly appears in FedSysID's performance analysis.

Intuitively, each client can cooperate to take advantage of other clients' data successfully through FL if the system heterogeneity parameter ϵ is small. Concretely, the objective of the federated system identification problem can be formulated as

$$\bar{\Theta} = \left[\bar{A}, \bar{B}\right] = \frac{1}{M} \sum_{i=1}^{M} \arg\min_{\Theta \in \mathbb{R}^{n \times (n+p)}} \|X^{(i)} - \Theta Z^{(i)}\|_F^2 \tag{4}$$

where, $X^{(i)} = \begin{bmatrix} X_1^{(i)} & \dots & X_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n \times N_i T}$, and $X_l^{(i)} = \begin{bmatrix} x_{l,T}^{(i)} & \cdots & x_{l,1}^{(i)} \end{bmatrix} \in \mathbb{R}^{n \times T}$, and similarly for $Z^{(i)}$ and $W^{(i)}$. We characterize the distance between the estimated system dynamics $\bar{\Theta}$, given by Eq (4), and each client's true system $\Theta^{(i)}$ in the following theorem.

Theorem 3 *For any fixed $\delta > 0$, let $N_i \geq \max\left\{8(n+p) + 16 \log \frac{2MT}{\delta}, (4n+2p) \log \frac{MT}{\delta}\right\}$. Then, with probability at least $1 - 3\delta$, the estimated dynamics $\bar{\Theta}$ in (4) satisfies*

$$\max\left\{\left\|\bar{A} - A^{(i)}\right\|, \left\|\bar{B} - B^{(i)}\right\|\right\}$$
$$\leq \frac{1}{\sqrt{\sum_{i=1}^{M} N_i}} \times \underbrace{\frac{C_0 \sqrt{\sum_{i=1}^{M} \sigma_{i,w}^2 \left(\sum_{t=0}^{T-1} \left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|\right)^2}}{\min_{i \in [M]} \lambda_{\min}\left(\sum_{t=0}^{T-1} \Sigma_t^{(i)}\right)}}_{C_1: \text{error constant}} + \epsilon \times \underbrace{\frac{9 \sum_{j \neq i} \sqrt{\left(\sum_{t=0}^{T-1} \left\|\Sigma_t^{(i)}\right\|\right)^2}}{\min_{i \in [M]} \lambda_{\min}\left(\sum_{t=0}^{T-1} \Sigma_t^{(i)}\right)}}_{C_2: \text{heterogeneity constant}},$$

for all $i \in [M]$, *where* $C_0 = 16\sqrt{(2n+p)\log\frac{9MT}{\delta}}$ *and* $\Sigma_t^{(i)}$ *is the covariance of* $z_{l,t}^{(i)} \triangleq \begin{bmatrix} x_{l,t}^{(i)\top} & u_{l,t}^{(i)\top} \end{bmatrix}^\top \in \mathbb{R}^{n+p}$.

Proof All proofs can be found in Wang et al. (2022b) (extended version of this paper). ■

Examining Theorem 3, the estimation errors consist of two terms. The first term matches the estimation error of the centralized setting in the setting where each client has access to all the data from all clients, i.e., access to all $N \triangleq \sum_{i=1}^{M} N_i$ rollouts. The error constant C_1 in Eq (5) comes from the noise that each client's system is subject to. Approximately, this is similar to the average signal-noise ratio of all systems across all time steps. The second term is due to the heterogeneity of the clients. Theorem 3 confirms our intuition that each client can improve its estimation performance from $\mathcal{O}({N_i}^{-1/2})$ to $\mathcal{O}((\sum_{i=1}^{M} N_i)^{-1/2})$ by participating in the collaboration, if the system heterogeneity ϵ is small i.e., when the first term dominates the second term in Theorem 3.

The results of Theorem 3 characterize the optimal solution of the least squares problem (4). Once this optimal solution $\bar{\Theta}$ is found, this estimation is a good fit for all the similar clients participating in the collaboration. In order to compute $\bar{\Theta}$, we now propose `FedSysID`, a federated algorithm (see Algorithm 1 below) that explicitly describes the client-server interaction. Although not the focus of this paper, the algorithm can also handle the case where only a subset of clients participate in each round.

At the beginning of the algorithm, each client has an initial guess $\bar{\Theta}^0$ about its dynamics and chooses α as the step size. At each round r, the subset of active users S_r is uniformly sampled from $[M]$.[2] All clients in S_r execute the ClientUpdate K_i-times using their local trajectories before communicating to the server (line 7). By performing multiple local model updates and communicating with the central server for a limited time, the algorithm resolves the core bottleneck of the limited communicating capabilities of clients. Once the local updates are completely performed, each client sends its local model update $\bar{\Theta}_{r+1}^{(i)}$ to the server (line 8). The server computes an improved global model $\bar{\Theta}_{r+1}$ by averaging the clients' local model updates and sends $\bar{\Theta}_{r+1}$ back to each client (line 11). Full details are presented in Algorithm 1.

We describe two different implementations of `ClientUpdate` in Algorithm 1: i) Federated Averaging algorithm (`FedAvg`) (McMahan et al., 2017b); ii) Linear Convergent Federated algorithm (`FedLin`) (Mitra et al., 2021).

- `FedAvg`: For each client in $[M]$, `ClientUpdate` iterates execute:

$$\Theta_{r,k}^{(i)} = \Theta_{r,k-1}^{(i)} + \alpha(X^{(i)} - \Theta_{r,k-1}^{(i)} Z^{(i)}) Z^{(i),\top}, \quad k = 1, 2, \cdots, K_i \tag{5}$$

  with input $\Theta_{r,0}^{(i)} = \bar{\Theta}_r$ and output $\Theta_{r+1}^{(i)} = \Theta_{r,K_i}^{(i)}$. From (Pathak and Wainwright, 2020; Mitra et al., 2021), it is known that both `FedAvg` cannot converge with a constant step size α. Thus, with an appropriately chosen stepsize for `FedAvg`, `FedSysID` will converge sub-linearly, i.e., $\mathbb{E}\|\bar{\Theta}_R - \Theta^{(i)}\| \leq \mathcal{O}(\frac{1}{KR})$, where the expectation is taken with respect to the sampling scheme, to the optimal solution of problem (4). See Theorem 3 in (Woodworth et al., 2020) for details.

2. This random sampling step is necessary when a large number of participants are involved as clients may lose connection to the central server due to connectivity or power issues.

Algorithm 1 `FedSysID`

Initialize the server with $\bar{\Theta}_0$ and step size α ;
Initialize each client $i \in [M]$ with $\Theta_{0,0}^{(i)} = \bar{\Theta}_0$;
For each round $r = 0, 1, \ldots, R-1$ **do**
    uniformly sample $S_r \subseteq \{1, 2, \cdots, M\}$
    $\triangleright$ Client side:
    For each client $i \in S_r$ **in parallel do**
        $\Theta_{r+1}^{(i)} = \texttt{ClientUpdate}(i, \bar{\Theta}_r, K_i)$
        send $\Theta_{r+1}^{(i)}$ back to the server
    end for
    $\triangleright$ Server side:
    update $\bar{\Theta}_{r+1} = \frac{1}{M}\sum_{i=1}^{M} \Theta_{r+1}^{(i)}$ and send $\bar{\Theta}_{r+1}$ to each client
end for
Return $\bar{\Theta}_R$

- `FedLin`: Consider the following updating rule of `ClientUpdate` for all the clients:

$$\begin{aligned} \Theta_{r,k}^{(i)} &= \Theta_{r,k-1}^{(i)} + \alpha \left[(X^{(i)} - \Theta_{r,k-1}^{(i)} Z^{(i)}) Z^{(i),T} - (X^{(i)} - \bar{\Theta}_r Z^{(i)}) Z^{(i),T} + g_r \right], \\ g_r &= \frac{1}{M} \sum_{i=1}^{M} (X^{(i)} - \bar{\Theta}_r Z^{(i)}) Z^{(i),T}, \quad k = 1, 2, \cdots, K \end{aligned} \tag{6}$$

  with input $\Theta_{r,0}^{(i)} = \bar{\Theta}_r$ and output $\Theta_{r+1}^{(i)} = \Theta_{r,K_i}^{(i)}$. Note that g_r is the negative gradient of the global function in problem (4). Each client can obtain it by exploiting memory without communication. See Mitra et al. (2021) for more details. With proper constant step sizes, `FedLin` converges linearly to the minimum solution $\bar{\Theta}$, i.e., $\mathbb{E}\|\bar{\Theta}_R - \Theta^{(i)}\| \leq \mathcal{O}(e^{-\beta R})$, where the expectation is taken with respect to the sampling scheme[3] and β is the condition number of problem (4).

Corollary 4 *Frame the hypotheses of Theorem 3. For all $R \geq 1$, the output $\bar{\Theta}_R$ given by the `FedSysID` algorithm satisfies*

$$\max\left\{\mathbb{E}\|\bar{A}_R - A^{(i)}\|, \mathbb{E}\|\bar{B}_R - B^{(i)}\|\right\} \leq \mathcal{O}\left(\frac{1}{KR} + \frac{1}{\sqrt{\sum_{i=1}^{M} N_i}} \times C_1 + \epsilon \times C_2\right) \tag{7}$$

with the `ClientUpdate` being performed according to the local updating rule of `FedAvg`. And,

$$\max\left\{\mathbb{E}\|\bar{A}_R - A^{(i)}\|, \mathbb{E}\|\bar{B}_R - B^{(i)}\|\right\} \leq \mathcal{O}\left(e^{-\beta R} + \frac{1}{\sqrt{\sum_{i=1}^{M} N_i}} \times C_1 + \epsilon \times C_2\right) \tag{8}$$

with the `ClientUpdate` selected from the local updating rule of `FedLin`.

Proof *All proofs can be found in Wang et al. (2022b) (extended version of this paper).* ■

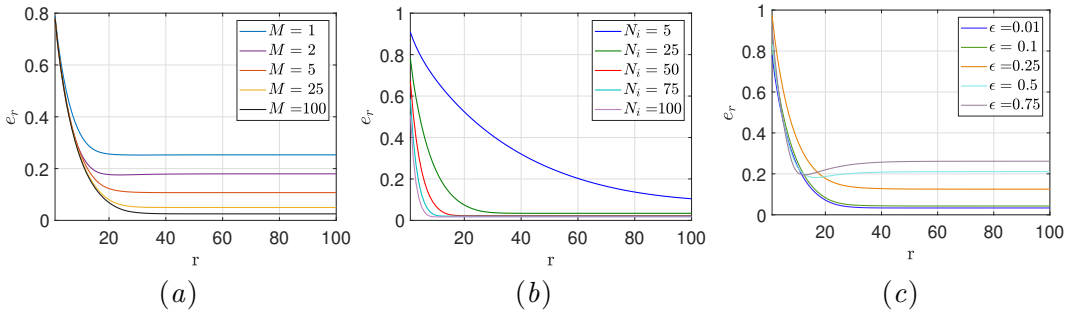
3. Even though, Mitra et al. (2021) does not mention it, `FedLin` can also be extended to handle partial participation of clients.

5. Experimental Results

Numerical results are now shown to illustrate and assess the efficiency `FedSysID`.[4] We consider clients characterized by dynamical systems with $n = 3$ states and $p = 2$ inputs, for which the standard deviations of their initial state, input, and process noise are set to be $\sigma_x = \sigma_u = \sigma_w = 1$. Following (Xin et al., 2022), we consider the nominal system (A_0, B_0), and uniformly distributed variables $(\gamma_1^{(i)}, \gamma_2^{(i)}) \backsim U(0, \epsilon)$, such that $A^{(i)} = A_0 + \gamma_1^{(i)} V$ and $B^{(i)} = B_0 + \gamma_2^{(i)} U$, where $V \in \mathbb{R}^{3\times 3}$ and $U \in \mathbb{R}^{3\times 2}$ represent the modification patterns applied to the nominal system described by,

$$A_0 = \begin{bmatrix} 0.6 & 0.5 & 0.4 \\ 0 & 0.4 & 0.3 \\ 0 & 0 & 0.3 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \\ 0.5 & 0.5 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that we use the same amount of data for each client, i.e., $N_i = N_j$ for all $i, j \in [M]$, and we set $T = 5$ for each client. Once $\bar{\Theta}$ is computed according to (4) via `FedSysID`, the estimation error is defined as the average of the distances between an estimation q at the r-th global iteration, $\bar{\Theta}_r^q$, and the fixed true system, $\Theta^{(1)}$, for different estimations (i.e., estimations made from different data sets). That is, $e_r = \frac{1}{q}\sum_{i=1}^{q} ||\bar{\Theta}_r^i - \Theta^{(1)}||$, where $q = 25$.



(a) (b) (c)

Figure 1: `FedLin` - Estimation error with respect to the number of global iterations with varying M, N_i and ϵ, where each client performs $K_i = 10$ local updates with step size 10^{-4}. (a) $N_i = 25$, $\epsilon = 0.01$ (b) $M = 50$, $\epsilon = 0.01$ and (c) $M = 50$, $N_i = 25$.

Figure 1 illustrates the estimation error with respect to the number of global iterations with varying M, N_i, and ϵ, when the `ClientUpdate` is set to the `FedLin` update. Figure 1 (a) describes the effect of the number of clients M participating in the collaboration on the estimation error, for a fixed number of rollouts $N_i = 25$ and a small dissimilarity parameter $\epsilon = 0.01$. This figure shows that the estimation error is considerably reduced when the number of clients participating increases. Whilst the estimation error is large for the centralized setting, $M = 1$ (i.e., each client learns its dynamics from its own data), it can be reduced by about 10 times when 100 clients collaborate to learn their dynamics by

4. All simulations were developed in MATLAB. Codes can be downloaded from https://github.com/jd-anderson/Federated-ID

leveraging data from similar systems. This highlights how the non-asymptotic convergence rate can be improved and scaled by a factor of $\sqrt{M}$ in the low heterogeneity regime.

Moreover, Figure 1(b–c) illustrate how the estimation error changes when the number of rollouts N_i and the system's dissimilarity ϵ increase. Overall, the figures show that the estimation error reduces with a factor of $\sqrt{MN_i}$ and it increases as the heterogeneity of the systems increase. Figure 2 details a similar analysis in which the same conclusions can be noticed with the `FedAvg ClientUpdate` update rule. As expected from Corollary 4, `FedSysID` converges faster with `FedLin` when compared to the `ClientUpdate` chosen from the local updating rule of `FedAvg`.

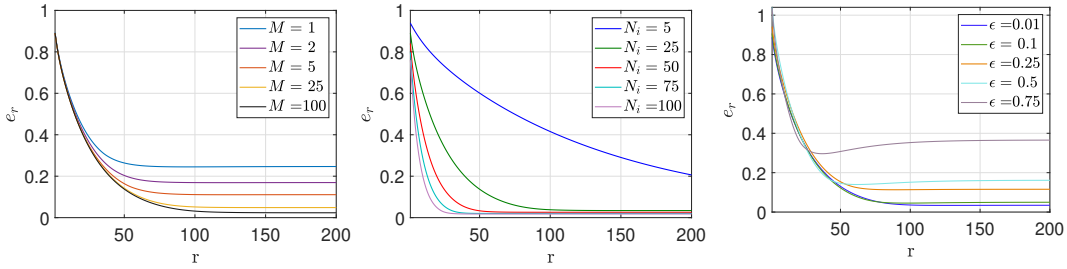


Figure 2: `FedAvg` - Estimation error with respect to the number of global iterations with varying M, N_i and ϵ, where each client performs $K_i = 10$ local updates with linear decreasing step size starting at 10^{-4}. (a) $N_i = 25$, $\epsilon = 0.01$ (b) $M = 50$, $\epsilon = 0.01$ and (c) $M = 50$, $N_i = 25$.

6. Conclusions

In this paper, we considered a federated approach for the system identification problem, where a central server orchestrates the collaboration of M clients, with similar dynamics, to identify a common dynamics that suits well all the clients. We demonstrated that a better non-asymptotic convergence rate scaled by $\sqrt{M}$ is achieved when similar systems learn their dynamics by leveraging data from each other in an FL setting. We also provided a `FedSysID` algorithm that can be used with any FL optimization method as a subroutine. To illustrate the potential of the proposed approach, numerical experiments were demonstrated considering both `FedLin` and `FedAvg` updating rules. Future work will explore an extension of the proposed approach for developing an online sample version of the `FedSysID`. We are currently investigating data-driven and system theoretic heterogeneity metrics.

Acknowledgments

The authors would like to thank Aritra Mitra for valuable and insightful discussions on this work. HW is generously funded by a Wei family fellowship and the Columbia Data Science Institute. The authors also gratefully acknowledge funding from the DoE under grant DE-SC0022234 and NSF awards 2144634 & 2231350.

References

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE, 2019.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001. IEEE, 2020.

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.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016a.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016b.

Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

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

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

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Reese Pathak and Martin J Wainwright. FedSplit: An algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for Dynamics and Control*, pages 16–25. PMLR, 2020.

Quoc Tran Dinh, Nhan H Pham, Dzung Phan, and Lam Nguyen. FedDR–randomized Douglas-Rachford splitting algorithms for nonconvex federated composite optimization. *Advances in Neural Information Processing Systems*, 34:30326–30338, 2021.

Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*, pages 1–11, 2019.

Anastasios Tsiamis and George J Pappas. Linear systems can be hard to learn. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2903–2910. IEEE, 2021.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical Learning Theory for Control: A Finite Sample Perspective. *arXiv preprint arXiv:2209.05423*, 2022.

Peter Van Overschee and BL0888 De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

Michel Verhaegen and Vincent Verdult. *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

Han Wang and James Anderson. Large-scale system identification using a randomized SVD. In *2022 American Control Conference (ACC)*, pages 2178–2185. IEEE, 2022.

Han Wang, Siddartha Marella, and James Anderson. FedADMM: A Federated Primal-Dual Algorithm Allowing Partial Participation. *arXiv preprint arXiv:2203.15104*, 2022a.

Han Wang, Leonardo F. Toso, and James Anderson. FedSysID: A Federated Approach to Sample-Efficient System Identification. *arXiv preprint arXiv:2211.14393*, 2022b.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs Local SGD for Heterogeneous Distributed Learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Lei Xin, Lintao Ye, George Chiu, and Shreyas Sundaram. Identifying the Dynamics of a System by Leveraging Data from Similar Systems. *arXiv preprint arXiv:2204.05446*, 2022.

Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.