APPENDIX

Anonymous authors
Paper under double-blind review

1 DERIVATIONS OF ADAPTIVE TRADE-OFF THEOREM

In this section, we present the derivations for the lower bound of the constrained optimization problem defined in the paper. The problem is shown as follows:

$$\begin{aligned} &\max I(\overline{y}'; \overline{\boldsymbol{z}}|\theta, D) \\ &s.t. I(\overline{y}', D; \theta|\boldsymbol{x}') \leq I_c \end{aligned} \tag{1}$$

where I_c is the constant information constraint.

In the local training, the predictive process can be defined as a Markov chain shown as $Y \leftrightarrow X \leftrightarrow Z$, where Y is the label, X is the input sample and Z is the latent feature. Then, for the problem in Eq 1, we can rewrite the problem by using Lagrange multiplier as follows:

$$\begin{aligned} &I(\overline{y}'; \overline{\boldsymbol{z}}, |\theta, D) - \beta I(\overline{y}', D; \theta|\boldsymbol{x}') \\ &= I(\overline{y}'; \overline{\boldsymbol{z}}, |\theta) + I(\overline{y}'; D|\theta, \overline{\boldsymbol{z}},) - I(\overline{y}'; D|\theta) - \beta I(\overline{y}', D; \theta|\boldsymbol{x}') \end{aligned} \tag{2}$$

There are mainly three terms in Eq . 2, including $I(\overline{y}'; \overline{\boldsymbol{z}}|\theta)$, $I(\overline{y}'; D|\theta, \overline{\boldsymbol{z}})$, and $I(\overline{y}', D; \theta|\boldsymbol{x}')$.

For the first term, we have:

$$\begin{aligned} I(\overline{y}'; \overline{\boldsymbol{z}}|\theta) &= \int p(\overline{\boldsymbol{z}}, y|\theta) log \frac{p(\overline{y}'|\overline{\boldsymbol{z}}, \theta)}{p(\overline{y}'|\theta)} d\overline{y}' d\overline{z} \\ &= \int p(\overline{\boldsymbol{z}}, \overline{y}'|\theta) \log p(\overline{y}'|\overline{\boldsymbol{z}}, \theta) d\overline{y}' d\overline{z} - \int p(\overline{y}'|\theta) \log p(\overline{y}'|\theta) \\ &\geq \int p(\overline{\boldsymbol{z}}, \overline{y}'|\theta) \log q(\overline{y}'|\overline{\boldsymbol{z}}, \theta) d\overline{y}' d\overline{z} + \int p(\overline{y}'|\theta) \log \frac{1}{p(\overline{y}'|\theta)} \\ &\geq \int p(\overline{\boldsymbol{z}}, \overline{y}'|\theta) \log q(\overline{y}'|\overline{\boldsymbol{z}}, \theta) d\overline{y}' d\overline{z} \end{aligned} \tag{3}$$

where $q(\overline{y}'|\overline{\boldsymbol{z}}, \theta)$ is a variational approximation to the target distribution $p(\overline{y}'|\overline{\boldsymbol{z}}, \theta)$. The first inequality is obtained by the theorem $KL(p(\overline{y}'|\overline{\boldsymbol{z}}, \theta)|q(\overline{y}'|\overline{\boldsymbol{z}}, \theta)) \geq 0$ and the second inequality is obtained by ignoring the $\int p(\overline{y}'|\theta) \log \frac{1}{p(\overline{y}'|\theta)}$ which is independent of our optimization procedure.

Then, as $\int p(\overline{\boldsymbol{z}}, \overline{y}'|\theta) d\overline{y}' d\overline{z} = \int p(\boldsymbol{x}', \overline{y}'|\theta) p(\boldsymbol{z}|\boldsymbol{x}', \theta) d\boldsymbol{x}' d\overline{y}' d\overline{z}$, we can approximate the lower bound of $I(\overline{y}'; \overline{\boldsymbol{z}}, |\theta)$ by using the empirical data distribution of $p(\boldsymbol{x}', \overline{y}')$ on D'. We obtain $I(\overline{y}'; \overline{\boldsymbol{z}}, |\theta) \geq E \int p(\overline{\boldsymbol{z}}|\boldsymbol{x}'_n, \theta) \log q(y_n|\overline{\boldsymbol{z}}, \theta) d\overline{z}$. Utilizing the reparameterization trick, this can be estimated as $E_{\boldsymbol{x}'} E_{\epsilon \sim N(0,I)} [\log q(\overline{y}'|\boldsymbol{x}', \theta, \epsilon)]$.

For the second term, we follow the proof in Appendix A.2 of the paper Yin et al. (2020) to obtain the following lower bound:

$$\begin{aligned} I(\overline{y}'; D|\theta, \overline{\boldsymbol{z}}) \geq I(\boldsymbol{x}'; \overline{y}'|\theta, \overline{z}) &\geq I(\boldsymbol{x}'; \overline{y}'|\theta) - I(\boldsymbol{x}'; \overline{z}|\theta) \\ &\geq I(\boldsymbol{x}'; \overline{y}'|\theta) - E\left[KL(p(\overline{z}|\boldsymbol{x}', \theta)||r(\overline{z}))\right] \end{aligned} \tag{4}$$

where the $p(\overline{z}|\boldsymbol{x}', \theta)$ can be computed as a deterministic function by using the reparameterization trick.

For the last term, we can obtain:

$$
\begin{aligned}
&I(\overline{y}', D; \theta|\boldsymbol{x}') = E\left[\log \frac{p(\overline{y}', D, \theta|\boldsymbol{x}')}{p(\overline{y}', D|\boldsymbol{x}')p(\theta|\boldsymbol{x}')}\right] = E\left[\log \frac{p(\theta|D, \boldsymbol{x}', \overline{y}')}{p(\theta|\boldsymbol{x}')}\right] \\
&\leq E\left[KL(p(\theta|D, \boldsymbol{x}', \overline{y}')||r(\theta)\right]
\end{aligned} \tag{5}
$$

where $r(\theta) \sim N(0, I)$ is a variational approximation to the target distribution of θ.

Following the observation in the paper Finn et al. (2018) that adapting the parameters with gradient descent is a good way to update them to a given training set D and the testing set D'. Then, we can get $p(\theta|D, \boldsymbol{x}', \overline{y}') = N(\theta_u - \gamma \triangledown_{\theta_u} L(\theta_u, D) - \gamma \triangledown_{\theta_u} L(\theta_u, D'); \theta_\sigma)$.

Finally, combing these terms, we obtain the lower bound as follows:

$$
\begin{aligned}
&I(\overline{y}'; \overline{\boldsymbol{z}}, |\theta, D) - \beta I(\overline{y}', D; \theta|\boldsymbol{x}') \\
&\geq E_{\boldsymbol{x}'} E_{\epsilon \sim N(0,I)} \left[\log q(\overline{y}'|\boldsymbol{x}', \theta, \epsilon)\right] \\
&+ I(\boldsymbol{x}'; \overline{y}'|\theta) - E\left[KL(p(\overline{z}|\boldsymbol{x}', \theta)||r(\overline{z}))\right] \\
&- I(\overline{y}'; D|\theta) \\
&- \beta E\left[KL(p(\theta|D, \boldsymbol{x}', \overline{y}')||r(\theta)\right]
\end{aligned} \tag{6}
$$

2 Construction Algorithm for Non-IID scenarios used in the Experiment

We propose the non-iid data scenario, namely Semi-Consistent non-IID (sc-non-IID) data for the experiments. Then, non-IID settings are different sc-non-IID data with different construction hyper-parameters. In this section, we first present definitions and discuss the relation between non-IID and sc-non-IID.

2.1 Definitions

The propose of semi-consistent non-IID (sc-non-IID) scenario for devices is motivated by mutually exclusive tasks in meta-learning Yin et al. (2020). As pointed out in work Jiang et al. (2019), the tasks in meta-learning have similar meanings to devices in federate learning. Thus, we transplant the concept of mutual exclusion of tasks in meta-learning to the relationship between devices in federated learning. That is, for arbitrary two mutually-exclusive devices in the system, they perform similar classification tasks sampled from a task distribution $P(\Gamma)$, but the samples, labels, and output of tasks are different. Therefore, unlike the existing non-IID data with only label distribution and sample quantity skewness, sc-non-IID introduces the triple heterogeneity of samples, labels, and tasks to make it highly consistent with the practical heterogeneous scenario required by the personalized FL. Following the definition of non-IID data, the mathematical definition of the sc-non-IID data is:

Definition 2.1 (Semi-consistent non-IID data)**.** For different devices m and n with corresponding datasets $(\mathbf{X}, \boldsymbol{y})^m$ and $(\mathbf{X}, \boldsymbol{y})^n$, respectively, the sc-non-IID statistics can be formulated as:

$$
\begin{aligned}
&(\boldsymbol{x}, y)^m \sim P_m, (\boldsymbol{x}, y)^n \sim P_n \\
&\exists i, j \in R, p(\boldsymbol{x}_i^m, \boldsymbol{x}_j^n) \neq p(\boldsymbol{x}_i^m)p(\boldsymbol{x}_i^n)
\end{aligned} \tag{7}
$$

where $y^m \in \mathcal{Y}^m$, $y^n \in \mathcal{Y}^n$, $\boldsymbol{x}^m \in \mathcal{X}^m$, $\boldsymbol{x}^n \in \mathcal{X}^n$, and $(\mathcal{Y}^m, \mathcal{X}^m) \sim \Gamma^m$, $(\mathcal{Y}^n, \mathcal{X}^n) \sim \Gamma^n$ with the tasks $\Gamma^m, \Gamma^n \sim p(\Gamma)$, .

Definition 2.2 (Non-mutually-exclusive rate)**.** For the decentralized system with C devices, the non-mutually-exclusive rate ζ is defined as $1 - \frac{|S^{me}|}{C}$, where S^{me} is a group of devices, in which any two devices meet definition 2.1.

From the definition 2.2, when $\zeta = 1$, the proposed sc-non-IID scenario equals to the standard non-IID scenario. The basic assumption of sc-non-IID data is the privacy and task heterogeneity enhancement of the standard non-IID data setup in existing work Kairouz et al. (2019). In real-world devices, the datasets among devices are not formed by partitioning an existing dataset based on labels or quantity but are generated by each independent device in an uncontrollable and high cross-device differences manner. Then, heterogeneity exists in both the data structure (i.e., feature distribution skew, label distribution skew, concept shift Li et al. (2020); Kulkarni et al. (2020), and quantity skew) and the tasks (i.e., the difference in objectives, evaluations, and metrics). Therefore, the intuitive understanding of sc-non-IID data is a mixture of these heterogeneities. The sc-non-IID data with a high privacy guarantee gives a better and comprehensive representation of the practical scenario required by the FL personalization. As it is an enhanced and extended non-IID data, we construct the sc-non-IID data based on the existing non-IID setup by further integrating the principles of meta dataset Triantafillou et al. (2019).

Algorithm 1: Construction algorithm of sc-non-IID

```
input  : The dataset $(\mathcal{Y}, \mathcal{X})$;          Device-specific classes upper bound $T$;
         Non-mutually-exclusive rate $\zeta$;                $P\%$ samples for support set in each device
output : The constructed sc-non-IID data for devices
Randomly select $|S^{me}| = \zeta \times C$, and $(1-\zeta) \times C$ devices for $c = 1;\ c < C;\ c = c + 1$ do
    $t \sim Uniform(2, T)$;
    Sample $\boldsymbol{y}^c \in \mathcal{Y}$, $|\ \boldsymbol{y}^c\ | = t$;
    $\boldsymbol{X}^c = \{\boldsymbol{X_i} \sim \mathcal{X}_i; i \in \boldsymbol{y}\}$; where $\mathcal{X}_i$ is a set of samples for class $i$ and $\mathcal{X}_i \in \mathcal{X}$.
    if $c \in S^{me}$ then
        $\overline{\boldsymbol{X}}^c \leftarrow DataAugmentation(\boldsymbol{X}^c)$
                                                   /* Sample heterogeneity */
        $\overline{\boldsymbol{y}}^c \leftarrow RelabelCategories(\boldsymbol{y}^c)$
                                                    /* Label heterogeneity */
        $D = (\overline{\boldsymbol{X}}, \overline{\boldsymbol{y}})$                              /* Task heterogeneity */
        Support set $D^c = (\overline{\boldsymbol{X}}, \overline{\boldsymbol{y}})^c$
        Query set $D'^c = (\overline{\boldsymbol{X}'}, \overline{\boldsymbol{y}'})^c$, where $\frac{|D^c|}{|D|} = P$ and $|D^c| + |D'^c| = |D|$
    else
        Support set, $D^c = (\boldsymbol{X}, \boldsymbol{y})^c$,
        Query set, $D'^c = (\boldsymbol{X}', \boldsymbol{y}')^c$, where $\frac{|D^c|}{|D|} = P$ and $|D^c| + |D'^c| = |D|$
    end
end
```

2.2 Analysis of the sc-non-IID scenario

The constructed sc-non-IID dataset with mixed heterogeneous is utilized for testing algorithms and assessing their resilience to different degrees of heterogeneity among devices. When $\zeta = 1$, we have the existing non-IID scenario. When $\zeta \in (0, 1)$, the sc-non-IID scenario introduces a challenge of achieving performance trade-off between optimality and adaptability of the global model for personalization. When $\zeta = 0$, a global model with high adaptability is perferrable. Thus, each device can gain a personalized model within a few updates with this global model as the initial model. In all three conditions, the main objective is to make the learning process adaptively balance the initial performance and the ability for personalization in local updates and server aggregation.

3 More Experimental details and results

In this section, we present more settings and detailed results of our AFML on CIFAR-100 and Shakespeare datasets under different non-IID settings.

3.1 Experimental Setup

Datasets. We utilize CIFAR100 and Shakespeare datasets. CIFAR100 has 500K training and 100K testing samples. The training and testing examples are partitioned across 500 and 100 clients, respectively. For Shakespeare, there are 16,068 training and 2,356 samples from 715 users. The dataset is partitioned across 528 clients for training and 52 clients for testing. The construction of non-IID settings used in experiments is motivated by the work Triantafillou et al. (2019). Then, the detailed construction procedure of generating the non-IID scenarios under $\zeta \in [0, 1]$ for these datasets is shown in Section 2 of the Appendix.

Table 1: Learning Setting

Datasets		**CIFAR100**		**Shakespeare**	
		server	clients	server	clients
STANDARD	Learning rate	0.001	0.01	0.001	0.001
	Batch	-	100	-	20
	Optimizer	SGD			
META	Learning rate	0.01	0.01	0.1	0.01
	Batch	-	20	-	16
	Optimizer	SGD			
	Support set	P% samples of local training dataset			

Table 2: Performance for personalized models on $\zeta = 0.2, 0.8$.

Rate	**Methods**	**Configurations**	**CIFAR100**			**Shakespeare**		
			Pers.1 Acc.	**Pers.5 Acc.**	**Com. Rou.**	**Pers.1 Acc.**	**Pers.5 Acc.**	**Com. Rou.**
0.2	**FedAvg-FT**	E=3	NONE	NONE	NONE	NONE	NONE	NONE
		E=10	NONE	NONE	NONE	NONE	NONE	NONE
	FedPer-Meta	20% Support	0.3253(0.0071)	0.3872(0.0063)	1561	0.2501(0.0087)	0.3092(0.0055)	285
		80% Support	NONE	NONE	None	NONE	NONE	NONE
	FedMAML	20% Support	0.4133(0.0041)	0.4881(0.0065)	1074	0.2679(0.0051)	0.3455(0.0062)	129
		80% Support	0.4298(0.0035)	0.4922(0.0042)	1230	0.2874(0.0026)	0.3598(0.0032)	156
	AFML	20% Support	0.5114(0.0076)	0.5633(0.0081)	1019	0.2961(0.0054)	**0.4302(0.0065)**	137
		80% Support	**0.5175(0.0071)**	**0.5762(0.0079)**	**963**	**0.3212(0.0049)**	0.4211(0.0054)	**119**
0.8	**FedAvg-FT**	E=3	0.4677(0.0053)	0.4891(0.0076)	1792	0.3602(0.0039)	0.3776(0.0062)	425
		E=10	0.5621(0.0039)	0.5702(0.0065)	1627	0.3688(0.0027)	0.3763(0.0051)	394
	FedPer-Meta	20% Support	0.5872(0.0064)	0.6254(0.0069)	1652	0.3709(0.0039)	0.3953(0.0042)	311
		80% Support	0.6301(0.0058)	0.6592(0.0061)	1408	0.3856(0.0054)	0.4076(0.0059)	271
	FedMAML	20% Support	0.6722(0.0054)	0.6970(0.0066)	1227	0.4218(0.0076)	0.4406(0.0085)	155
		80% Support	0.6893(0.0075)	0.7046(0.0084)	1052	0.4323(0.0048)	0.4458(0.0063)	115
	AFML	20% Support	0.6854(0.0044)	0.7325(0.0049)	884	0.4280(0.0033)	0.4743(0.0042)	126
		80% Support	**0.6912(0.0051)**	**0.7571(0.0058)**	**743**	**0.4376(0.0018)**	**0.4880(0.0029)**	**109**

The ”NONE” represents that algorithm cannot converge.

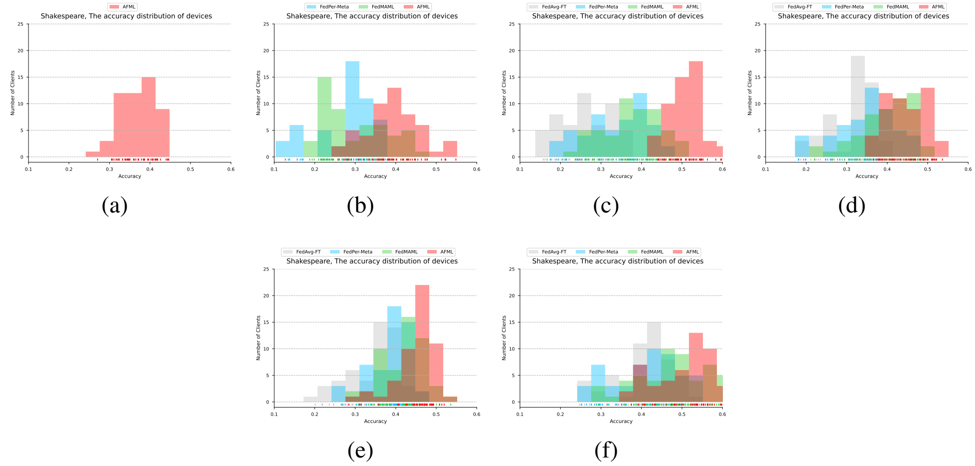
Learning Setting. We cast all evaluation tasks as the classification problem and consider two settings: MODEL and LEARNING. For the MODEL, the neural network (NN) is the only considered architecture for both the encoding and classification networks. They are constructed in simple structures because of the limited computation resources in the clients. Specifically, for CIFAR100, the encoding network contains three VGG blocks and one dense layer, while the classifier contains three dense layers. For Shakespeare, the encoding network is a two-dense layer, and the classification module has two LSTM hidden layers with 100 memory cells following one 100-d dense layer used for the next word prediction. For the LEARNING, there are two settings, including META for FedPer-Meta, FedMAML and STANDARD for FedAvg-FT, summarized in Section 3 Table 1. For coefficient α and β of regularization terms of AFML, we set $\alpha = 0.05$ and $\beta = 0.01$. The learning rates η_1, η_2, η for our AFML are $0.0001, 0.0005, 0.001$, respectively.

Performance metrics: AFML is compared with three state-of-the-art methods, including FedAvg-FT Jiang et al. (2019), FedPer-Meta Fallah et al. (2020), and FedMAML Deng et al. (2020) where FedPer-Meta is an enhancement work of FedPer-Meta Chen et al. (2019). We conduct three metrics for the performance evaluation. Firstly, we study the number of communication rounds (Com. Rou.) to reach convergence. The system budget in terms of the number of bytes downloaded/uploaded in communication and the floating-point operations (Flops) required for a specific accuracy is provided for further analysis. Secondly, we use the Top1-accuracy to evaluate the personalized model after performing 1 and 5 personalization epochs (i.e., Pers1. Acc. and Pers5. Acc). Third, we show the final accuracy distribution of personalized models for all clients and the comparison results with the locally trained models.

REFERENCES

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2019.

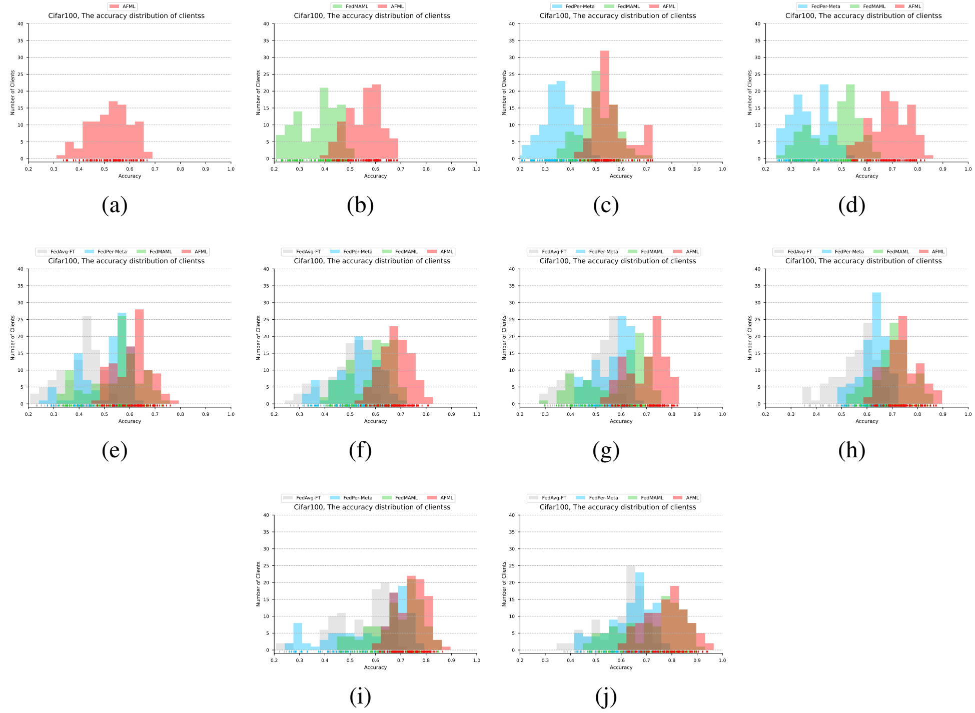
Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.



(a) (b) (c) (d)

(e) (f)

Figure 1: Accuracy distributions of devices on Shakespeare dataset under non-mutually-exclusive rate $\zeta = [0, 1]$. Compare AFML with FedAvg-FT, FedPer-Meta, and FedMAML. (a)$\zeta = 0$, (b)$\zeta = 0.2$, (c)$\zeta = 0.4$, (d)$\zeta = 0.6$, (e)$\zeta = 0.8$, (f)$\zeta = 1$.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Figure 2: Accuracy distributions of devices on CIFAR-100 dataset under non-mutually-exclusive rate $\zeta = [0, 1]$. Compare AFML with FedAvg-FT, FedPer-Meta, and FedMAML. (a)$\zeta = 0$, (b)$\zeta = 0.1$, (c)$\zeta = 0.2$, (d)$\zeta = 0.3$, (e)$\zeta = 0.5$, (f)$\zeta = 0.6$, (g)$\zeta = 0.7$, (h)$\zeta = 0.8$, (i)$\zeta = 0.9$, (j)$\zeta = 1$.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pp. 9516–9527, 2018.

Yihan Jiang, Jakub Konečný, Keith Rush, Ruiyu Li, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. *arXiv preprint arXiv:2003.08673*, 2020.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2020.