Subspace Constraint and Contribution Estimation for Heterogeneous Federated Learning

Supplementary Material

6. Detailed Configuration

In addition to the hyperparameter settings outlined in the main paper, we follow the configurations recommended in the original papers for each baseline method. SOLO requires no additional hyperparameters. For FML [39], we set the knowledge distillation parameters as $\alpha = 0.5$ and $\beta = 0.5$. For FedKD [48], we match the auxiliary model learning rate to the client learning rate, both set at 0.01, with $T_{\text{start}} = 0.95$ and $T_{\text{end}} = 0.95$. For FedDistill [19], we use $\gamma = 0.1$. For FedProto [42], we set $\lambda = 0.1$. For FedTGP [57], we configure $\lambda = 0.1$ (the same as FedProto), $\tau = 100$, and $S = 100$ (server training epochs). The feature vector dimension is set to 512, as suggested in [57] .

7. Label Distribution Skew

Fig. 8 illustrates the data distribution across varying levels of heterogeneity in the *Label skew* scenario, with the number of clients set to 20. We use the size of the point to represent the sample size.

8. Effect of Subspace Update Interval s

In Tab. 7, we present the impact of the subspace update interval on model performance. Our observations indicate that FedSCE outperforms the optimal baseline across all update intervals, demonstrating its robustness to hyperparameters. Additionally, accuracy can be further enhanced by carefully tuning the subspace update interval.

	Office10				***Cifar10***			
s	1	3	5	10	1	3	5	10
Acc	75.55	75.70	76.34	76.66	72.90	72.99	72.86	72.85

Table 7. **Hyperparameter Study:** Test accuracy (%) of FedSCE under **different** s for the *Office10* and *Cifar10* datasets. Please refer to Sec. 8 for a detailed discussion.

9. Further Ablation Experiments

As shown in Tab. 8, we conducted ablation experiments on the adjustment strategies on the client (LSL) and server (CE) to verify their effectiveness. In addition, we further show the mean and standard deviation of different components under three different random seeds in Tab. 9.

10. Discussion and Limitation

The performance of a model fundamentally depends on its generalization ability and convergence behavior. FedSCE enhances generalization performance by constraining model updates to a low-rank subspace and reducing the complexity of the model space. While traditional dimensionality reduction techniques may improve generalization, they usually affect the convergence speed. Exploiting the inherent low-rank property of the update trajectory, FedSCE adopts dynamic singular value decomposition (SVD) to continuously refine the subspace, effectively minimizing the projection error while maintaining convergence. However, if we only use a fixed subspace that is randomly created at the start, it often leads to much larger projection errors, which can slow down or prevent proper convergence.

In addition, considering the computational and storage overheads associated with subspace updates, we provide the per-round computation time and average client memory utilization of various algorithms on the *Office10* dataset in Tab. 10. FedSCE incurs a manageable level of memory and computation overhead due to subspace updates, whereas FedKD [48] requires more storage and computation resources due to local SVD calculations.

LSL	CE	*Amaz*	*Calt*	*DSLR*	*Webc*	***C.Avg***	***D.Avg***
		82.04	63.34	77.21	90.54	78.29	74.45
✓		83.29	63.70	78.48	91.89	79.34	75.31
	✓	84.34	64.59	79.74	92.56	80.31	76.26
✓	✓	84.55	64.59	79.74	92.57	80.36	76.34

Table 8. **Ablation** experiment with adjustments on the client (LSL) and server (CE). Please see details in Sec. 9.

LSL	DML	CE	***C.Avg***	***D.Avg***
			78.49±0.13	73.29±0.10
✓			79.53±0.28	74.08±0.16
✓	✓		79.65±0.25	75.45±0.31
✓	✓	✓	80.40±0.10	76.47±0.14

Table 9. The **mean** and **standard deviation** of the performance of different components. Please see details in Sec. 9.

	SOLO	FML	KD	Distill	Proto	TGP	Our
M.cost	0.18	0.22	0.31	0.18	0.18	0.18	0.24
Time	20.01	27.82	45.53	30.50	29.62	32.17	30.43

Table 10. **Memory cost** (GB) and **running time** (second) of different algorithms. Please see details in Sec. 10.

11. Generalization Analysis for Subspace Constraint

Let $\mathcal{X} \subset \mathbb{R}^D$ represent the input space, $\mathcal{Z} \subset \mathbb{R}^I$ the latent feature space, and $\mathcal{Y} \subset \mathbb{R}$ the output space. A feature mapping function $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Z}$ projects inputs into a feature space. We define a domain $\mathcal{T}$ as a combination of a data distribution $\mathcal{D}$ over $\mathcal{X}$ and a labeling function $c^* : \mathcal{X} \mapsto \mathcal{Y}$. Given a domain $\mathcal{T} := \langle \mathcal{D}, c^* \rangle$ and a representation $\mathcal{G}$, let $\tilde{\mathcal{D}}$ denote the induced distribution of $\mathcal{D}$ under $\mathcal{G}$ [4], such that for any event $\mathcal{B}$,

$$\mathbb{E}_{z\sim\tilde{\mathcal{D}}}[\mathcal{B}(z)] = \mathbb{E}_{x\sim\mathcal{D}}[\mathcal{B}(\mathcal{G}(x))]. \tag{17}$$

Let $h : \mathcal{Z} \mapsto \mathcal{Y}$ denote a hypothesis that maps features to labels, and let $\mathcal{H} \subseteq \{h : \mathcal{Z} \mapsto \mathcal{Y}\}$ be a hypothesis class. For two distributions $\mathcal{D}$ and $\mathcal{D}'$, the $\mathcal{H}$-*divergence* $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ between $\mathcal{D}$ and $\mathcal{D}'$ is defined as:

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') := 2 \sup_{\mathcal{A}\in\mathcal{A}_{\mathcal{H}}} |\Pr_{\mathcal{D}}(\mathcal{A}) - \Pr_{\mathcal{D}'}(\mathcal{A})|, \tag{18}$$

where $\mathcal{A}_{\mathcal{H}}$ is the set of measurable subsets induced by $h \in \mathcal{H}$. Additionally, $\mathcal{H}\Delta\mathcal{H}$ represents the *symmetric difference hypothesis space* [5]:

$$\mathcal{H}\Delta\mathcal{H} := \{h(z) \oplus h'(z) \mid h, h' \in \mathcal{H}\}, \tag{19}$$

where $\oplus$ denotes the XOR operation, indicating disagreement between $h(z)$ and $h'(z)$. Let $\mathcal{A}_{\mathcal{H}\Delta\mathcal{H}}$ be the corresponding set of measurable subsets. The *distribution divergence* $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ induced by the symmetric difference hypothesis space is given by [5]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}, \mathcal{D}') := 2 \sup_{\mathcal{A}\in\mathcal{A}_{\mathcal{H}\Delta\mathcal{H}}} |\Pr_{\mathcal{D}}(\mathcal{A}) - \Pr_{\mathcal{D}'}(\mathcal{A})|. \tag{20}$$

Now, let $\mathcal{D}$ and $\mathcal{D}'$ be two distributions over the input space $\mathcal{X}$, and let $\tilde{\mathcal{D}}$ and $\tilde{\mathcal{D}}'$ denote their respective induced distributions over $\mathcal{G}$. Then, using the definition of $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$, we have:

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}, \tilde{\mathcal{D}}') &= 2 \sup_{\mathcal{A}\in\mathcal{A}_{\mathcal{H}\Delta\mathcal{H}}} |\mathbb{E}_{x\sim\mathcal{D}}[\Pr(\mathcal{A}(\mathcal{G}(x)))] - \mathbb{E}_{x\sim\mathcal{D}'}[\Pr(\mathcal{A}(\mathcal{G}(x)))]| \\ &= 2 \sup_{\mathcal{A}\in\mathcal{A}_{\mathcal{H}\Delta\mathcal{H}}} |\mathbb{E}_{z\sim\tilde{\mathcal{D}}}[\Pr(\mathcal{A}(z))] - \mathbb{E}_{z\sim\tilde{\mathcal{D}}'}[\Pr(\mathcal{A}(z))]| \\ &= 2 \sup_{\mathcal{A}\in\mathcal{A}_{\mathcal{H}\Delta\mathcal{H}}} |\Pr_{\tilde{\mathcal{D}}}(\mathcal{A}) - \Pr_{\tilde{\mathcal{D}}'}(\mathcal{A})|. \end{aligned} \tag{21}$$

Then, we provide the theorem for the generalization performance from [4, 5, 56, 61] under their assumptions.

Theorem 1 *Let $\mathcal{T}_S$ and $\mathcal{T}_T$ be the source and target domains, whose data distributions are $\mathcal{D}_S$ and $\mathcal{D}_T$. Let $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Z}$ be a feature representation function, and $\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T$ be the induced images of $\mathcal{D}_S$ and $\mathcal{D}_T$ over $\mathcal{G}$, respectively. Let $\mathcal{H}$ be a set of hypotheses with VC-dimension d. Then with probability at least $1-\delta$, $\forall h \in \mathcal{H}$:*

$$\mathcal{L}_{\mathcal{T}_T}(h) \le \hat{\mathcal{L}}_{\mathcal{T}_S}(h) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)} + d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T) + \lambda, \tag{22}$$

where e is the base of the natural logarithm, $\hat{\mathcal{L}}_{\mathcal{T}_S}(h)$ is the empirical risk of the source domain given m observable samples, and $\lambda = \min_{h\in\mathcal{H}}(\mathcal{L}_{\mathcal{T}_T}(h) + \mathcal{L}_{\mathcal{T}_S}(h))$ is the optimal risk on the two domains.

In the context of hierarchical federated learning (HFL), where the emphasis is on transferring knowledge from global to local models on the client side [56], we can restate Theorem 1 as follows:

Theorem 2 *Consider a virtual global data domain $\mathcal{T}$ and a local data domain $\mathcal{T}_m$. Define $\mathcal{T} = \langle \mathcal{D}, c^* \rangle$ with $|\hat{\mathcal{D}}| = n$ and $\mathcal{T}_m = \langle \mathcal{D}_m, c^* \rangle$, where $\mathcal{D}, \mathcal{D}_m \subseteq \mathcal{X}$, and $c^* : \mathcal{X} \mapsto \mathcal{Y}$ is a ground-truth labeling function. Let $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Z}$ be a feature extraction function. Suppose $\mathcal{H}$ represents a hypothesis space constrained by subspace with VC dimension d', and $h : \mathcal{Z} \mapsto \mathcal{Y}$ for each $h \in \mathcal{H}$. Then, with probability at least $1-\delta$, the following holds for any $h \in \mathcal{H}$:*

$$\mathcal{L}_{\mathcal{T}_m}(h) \le \hat{\mathcal{L}}_{\mathcal{T}}(h) + \sqrt{\frac{4}{n}\left(d' \log \frac{2en}{d'} + \log \frac{4}{\delta}\right)} + d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}, \tilde{\mathcal{D}}_m) + \lambda_m, \tag{23}$$

where $d' < d < 2n$, d denotes the VC dimension of the standard hypothesis space. $\lambda_m := \min_h(\mathcal{L}_{\mathcal{T}_m}(h) + \mathcal{L}_{\mathcal{T}}(h))$ denotes an oracle performance, e is the base of the natural logarithm, $\tilde{\mathcal{D}}, \tilde{\mathcal{D}}_m$ are the induced distributions of $\mathcal{D}, \mathcal{D}_m$ under $\mathcal{G}$. $d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_m, \tilde{\mathcal{D}})$ denotes the divergence measured over a symmetric-difference hypothesis space [5]. $\hat{\mathcal{L}}_{\mathcal{T}}(h)$ is the empirical risk on $\mathcal{T}$.

12. Convergence Analysis in Non-convex Settings

In $\mathcal{L}_m^t, \mathcal{L}_m^{t,0}$, we use t represents the time step before the subspace update, and $t, 0$ represents the time step between the subspace update and the first iteration of the current round. In addition, we use t, e to denote the e-th local epoch of the t-th communication round. Next, we prove the non-convex convergence of the local cross entropy loss with dynamic subspace constraint.

Assumption 1 *Each local objective function is L_1-Lipschitz smooth; that is, $\|\nabla\mathcal{L}_m^{t_1,e} - \nabla\mathcal{L}_m^{t_2,e}\|_2 \leq L_1\|\theta_m^{t_1,e} - \theta_m^{t_2,e}\|_2$, which implies the following quadratic bound, $\mathcal{L}_m^{t_1,e} - \mathcal{L}_m^{t_2,e} \leq \langle\nabla\mathcal{L}_m^{t_2,e}, (\theta_m^{t_1,e} - \theta_m^{t_2,e})\rangle + \frac{L_1}{2}\|\theta_m^{t_1,e} - \theta_m^{t_2,e}\|_2^2, \forall t_1, t_2 > 0$.*

Assumption 2 *The stochastic gradient $g_m^{t,e} = \nabla\mathcal{L}(\theta_m^{t,e}, \xi_m^t)$ is an unbiased estimator of the local gradient for each client. $\mathbb{E}_{\xi_m\sim D_m}[g_m^{t,e}] = \nabla\mathcal{L}(\theta_m^{t,e}) = \nabla\mathcal{L}_m^{t,e}$, and its variance is bounded by σ^2: $\mathbb{E}[\|g_m^{t,e} - \nabla\mathcal{L}(\theta_m^{t,e})\|_2^2] \leq \sigma^2$.*

Assumption 3 *The expectation of the stochastic gradient is bounded by G: $\mathbb{E}[\|g_m^{t,e}\|_2^2] \leq G^2, \forall t, e$.*

Proposition 1 *For the local update $\Delta\tilde{\theta}_m^{t,e}$, the squared projection error is bounded; that is, $\mathbb{E}[\|\mathtt{Proj}(\Delta\tilde{\theta}_m^{t,e}) - \Delta\tilde{\theta}_m^{t,e}] \leq \rho_R\eta^2 e^2 G^2, \rho_R = \frac{\sum_{r=R+1}^K \sigma_r^2}{\sum_{r=1}^K \sigma_r^2}, \forall t, e$. σ_r denotes singular value.*

In fact, we make this proposition based on the low-rank nature of the model update space. This may be a bit strict, but if we replace $\rho_R = \frac{\sum_{r=R+1}^K \sigma_r^2}{\sum_{r=1}^K \sigma_r^2}$ with $0 < \rho_R < 1$, the following convergence proof also holds.

12.1. Key Lemmas

Lemma 1 *Let Assumption 1 and Assumption 3 hold. The total client loss of an arbitrary client can be bounded:*

$$\mathbb{E}[\mathcal{L}_m^{t+1}] \leq \mathcal{L}_m^{t,0} - (\eta - \frac{L_1\eta^2}{2})\sum_{e=0}^{E-1}\|\nabla\mathcal{L}_m^{t,e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2. \tag{24}$$

Proof. This lemma only focuses on local training at the client level, combined with the initial local goal. By replacing the relevant symbols in Lemma 1 of FedProto [42], it can be obtained.

Lemma 2 *Let Assumption 1 to 3 and Proposition 1 hold. After the local subspace update, the loss function of an arbitrary client can be bounded as:*

$$\mathbb{E}[\mathcal{L}_m^{t+1,0}] \leq \mathcal{L}_m^{t+1} + \mu\rho_R\eta^2 E^2 G^2. \tag{25}$$

Proof.

$$\begin{aligned}
\mathcal{L}_m^{t+1,0} &= \mathcal{L}_m^{t+1} + \mathcal{L}_m^{t+1,0} - \mathcal{L}_m^{t+1} \\
&= \mathcal{L}_m^{t+1} + \mu\left\|P_t P_t^T\Delta\tilde{\theta}_m^{t,E} - \Delta\tilde{\theta}_m^{t,E}\right\|^2 - \mu\left\|P_{t+1}P_{t+1}^T\Delta\tilde{\theta}_m^{t,E} - \Delta\tilde{\theta}_m^{t,E}\right\|^2 \\
&\leq \mathcal{L}_m^{t+1} + \mu\left\|P_t P_t^T\Delta\tilde{\theta}_m^{t,E} - \Delta\tilde{\theta}_m^{t,E}\right\|^2 \\
&\leq \mathcal{L}_m^{t+1} + \mu\rho_R\eta^2 E^2 G^2.
\end{aligned} \tag{26}$$

Take expectations of random variable ξ on both sides, then

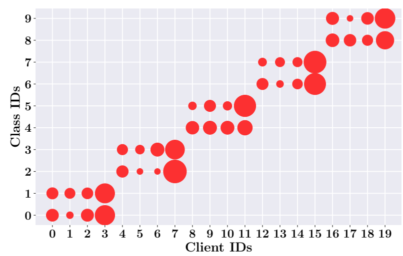
$$\mathbb{E}[\mathcal{L}_m^{t+1,0}] \leq \mathcal{L}_m^{t+1} + \mu\rho_R\eta^2 E^2 G^2. \tag{27}$$

12.2. Theorems

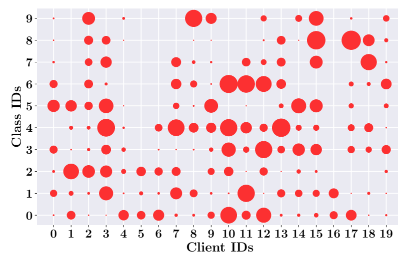
Theorem 3 *(One-round deviation) Let Assumption 1 to 3 and Proposition 1 hold. For an arbitrary client, after every communication round, we have,*

$$\mathbb{E}[\mathcal{L}_m^{t+1,0}] \leq \mathcal{L}_m^{t,0} - (\eta - \frac{L_1\eta^2}{2})\sum_{e=0}^{E-1}\|\nabla\mathcal{L}_m^{t,e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2 + \mu\rho_R\eta^2 E^2 G^2. \tag{28}$$

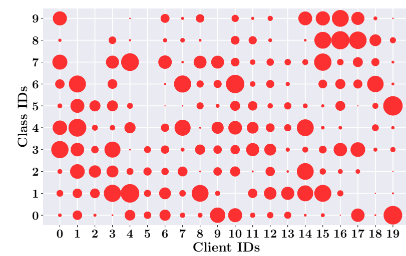
Proof. Taking the expectation of θ_m^t on both sides of Lemma 2, and then adding Lemma 1 and Lemma 2, we get Eq. (28).



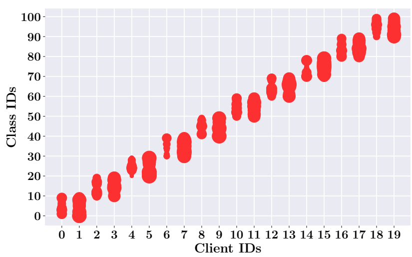
(a) *Cifar10* (Pat)



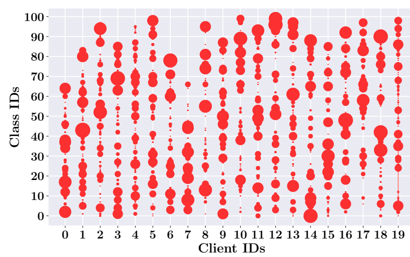
(b) *Cifar10* $(Dir_{0.3})$



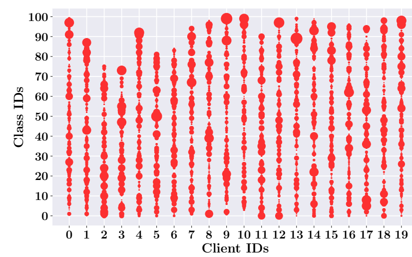
(c) *Cifar10* $(Dir_{0.5})$



(d) *Cifar100* (Pat)



(e) *Cifar100* $(Dir_{0.1})$



(f) *Cifar100* $(Dir_{0.5})$

Figure 8. Data distribution of each client on *Cifar10* and *Cifar100* under different settings. The size of the circle indicates the number of samples.

Theorem 4 *Assuming that Assumption 1 to 3 and Proposition 1 hold, $\mathcal{L}_m^*$ denotes the local optimum, the following inequality is valid for any arbitrary client and any $\epsilon > 0$:*

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_m^{t,e}\|_2^2] \leq \frac{L_1\eta\sigma^2 + 2\mu\rho_R\eta EG^2}{2 - L_1\eta} + \frac{2(\mathcal{L}_m^0 - \mathcal{L}_m^*)}{TE\eta(2 - L_1\eta)} < \epsilon. \tag{29}$$

$Proof$. Considering the communication rounds of Eq. (28) from $t = 0$ to $t = T - 1$ and the local epoch of each communication round of Eq. (28) from $e = 0$ to $e = E$, we have

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_m^{t,e}\|_2^2] \leq \frac{\frac{1}{TE}\sum_{t=0}^{T-1}(\mathcal{L}_m^{t,0} - \mathbb{E}[\mathcal{L}_m^{t+1,0}]) + \frac{L_1\eta^2}{2}\sigma^2 + \mu\rho_R\eta^2EG^2}{\eta - \frac{L_1\eta^2}{2}}. \tag{30}$$

Given any $\epsilon > 0$, let

$$\frac{\frac{1}{TE}\sum_{t=0}^{T-1}(\mathcal{L}_m^{t,0} - \mathbb{E}[\mathcal{L}_m^{t+1,0}]) + \frac{L_1\eta^2}{2}\sigma^2 + \mu\rho_R\eta^2EG^2}{\eta - \frac{L_1\eta^2}{2}} < \epsilon, \tag{31}$$

Let $\Delta = \mathcal{L}_m^0 - \mathcal{L}_m^*$. Since $\sum_{t=0}^{T-1}(\mathcal{L}_m^{t,0} - \mathbb{E}[\mathcal{L}_m^{t+1,0}]) \leq \Delta$, the above equation holds when

$$\frac{\frac{2\Delta}{TE} + L_1\eta^2\sigma^2 + 2\mu\rho_R\eta^2EG^2}{2\eta - L\eta^2} = \frac{2(\mathcal{L}_m^0 - \mathcal{L}_m^*)}{TE\eta(2 - L_1\eta)} + \frac{L_1\eta\sigma^2 + 2\mu\rho_R\eta EG^2}{2 - L_1\eta} < \epsilon, \tag{32}$$

that is:

$$T > \frac{2\Delta}{E\epsilon(2\eta - L_1\eta^2) - E\eta(L_1\eta\sigma^2 + 2\mu\rho_R\eta EG^2)}, \tag{33}$$

then, we have:

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_m^{t,e}\|_2^2] < \epsilon, \tag{34}$$

when

$$\eta < \frac{2\epsilon}{L_1(\epsilon + \sigma^2) + 2\mu\rho_R EG^2}. \tag{35}$$

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021. 2

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 3

[3] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. 5

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 5, 2

[5] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007. 5, 2

[6] Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12077–12086, 2024. 3

[7] 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. PMLR, 2021. 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[9] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10081, 2022. 2

[10] Xiuwen Fang, Mang Ye, and Xiyuan Yang. Robust heterogeneous federated learning under data corruption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5020–5030, 2023. 2

[11] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022. 2

[12] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012. 6

[13] Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low dimensional random bases. *Advances in Neural Information Processing Systems*, 33:12140–12150, 2020. 3

[14] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[17] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10143–10153, 2022. 1, 2, 3

[18] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. 1

[19] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018. 3, 6, 7, 8, 1

[20] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16302–16311, 2023. 2

[21] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous multivariate distributions*. Wiley New York, 1972. 6

[22] 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. PMLR, 2020. 1, 2

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[25] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[26] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. *Advances in Neural Information Processing Systems*, 35:38461–38474, 2022. 2

[27] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018. 3

[28] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. 1, 3

[29] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 2

[30] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021. 1

[31] 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. 1

[32] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 1, 2

[33] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021. 1, 3

[34] Tao Li, Zhehao Huang, Qinghua Tao, Yingwen Wu, and Xiaolin Huang. Trainable weight averaging: Efficient training by optimizing historical solutions. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3

[35] Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional trajectory hypothesis is true: Dnns can be trained in tiny subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2022. 2, 3

[36] Kangyang Luo, Xiang Li, Yunshi Lan, and Ming Gao. Gradma: A gradient-memory-based accelerated federated learning with alleviated catastrophic forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3708–3717, 2023. 2

[37] 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, 2017. 2

[38] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. 2

[39] Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Kun Kuang, Fei Wu, and Chao Wu. Federated mutual learning. *arXiv preprint arXiv:2006.16765*, 2020. 1, 2, 3, 6, 7, 8

[40] Eduardo D Sontag et al. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998. 5

[41] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020. 3

[42] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8432–8440, 2022. 1, 3, 5, 6, 7, 8

[43] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022. 1

[44] Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Quasi-newton's method in the class gradient defined high-curvature subspace. *arXiv preprint arXiv:2012.01938*, 2020. 3

[45] Oriol Vinyals and Daniel Povey. Krylov subspace descent for deep learning. In *Artificial intelligence and statistics*, pages 1261–1268. PMLR, 2012. 3

[46] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020. 2

[47] Dingzhu Wen, Ki-Jun Jeon, and Kaibin Huang. Federated dropout—a simple approach for enabling federated learning on resource constrained devices. *IEEE wireless communications letters*, 11(5):923–927, 2022. 2

[48] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022. 2, 3, 4, 6, 7, 8, 1

[49] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 1

[50] Xiyuan Yang, Wenke Huang, and Mang Ye. Dynamic personalized federated learning with adaptive differential privacy. *Advances in Neural Information Processing Systems*, 36:72181–72192, 2023. 1

[51] Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11986–11995, 2024. 3

[52] Liping Yi, Gang Wang, Xiaoguang Liu, Zhuan Shi, and Han Yu. Fedgh: Heterogeneous federated learning with generalized global header. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8686–8696, 2023. 1

[53] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022. 2

[54] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, Jian Cao, and Haibing Guan. Gpfl: Simultaneously learning global and personalized feature information for personalized federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5041–5051, 2023. 1

[55] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11237–11244, 2023. 1, 6

[56] Jianqing Zhang, Yang Hua, Jian Cao, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Eliminating domain bias for federated learning in representation space. *Ad-*

vances in Neural Information Processing Systems, 36, 2024. 5, 2

[57] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16768–16776, 2024. 1, 3, 6, 7, 8

[58] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. An upload-efficient scheme for transferring knowledge from a server-side pre-trained generator to clients in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12109–12119, 2024. 1

[59] Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, pages 26293–26310. PMLR, 2022. 2

[60] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 561–578. Springer, 2020. 6

[61] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021. 5, 2