A. Appendix

A.1. Experiment Details

We give details omitted from the manuscript related to the experimentation for reproducibility.

Federated Datasets. We use 100 devices in our experiments. We first assign a fixed number of classes to each of the device. More specifically, device i has class list as $\{i \bmod C, (i+1) \bmod C, \ldots, (i+S-1) \bmod C\}$ where C is the total number of classes and S is the size of the class list. For instance, in CIFAR-10, ACID 5 class per device setting, device 10 has class list of $\{0, 1, 2, 3, 4\}$ whereas device 25 has class list of $\{5, 6, 7, 8, 9\}$. With this construction we guarantee that many devices such as device 10 and 25 have non overlapping classes. After fixing the class list, we distribute training and test data instances for each device based on its class list without replacement from the training and test split of the original dataset. This construction gives a training set of size 500 datapoints and a test set of size 100 datapoints for each device. Different from ACID, in ALID setting, we further permute class labels for each device. We use the same label permutation for the training and test dataset of a device.

Models. We use a convolutional network in our experiments similar to the one in McMahan et al. (2017); Acar et al. (2021a). Our architecture has two convolutional layers with 64 filter size and 5×5 kernels. Each convolutional layers are followed by a max pooling layer. After the second max pooling layer, we use two fully connected layers of size 384 and 192 with ReLU activation. Finally, we use a softmax layer to get predictions.

Hyperparameters. We fix the batch size as 50, the number of SGD steps as $K = 50$, the learning rate as $\beta = 0.1$ and the weight decay as 0.001 in our experiments. To avoid divergence, we set a learning rate decay across communication rounds as 0.997.

MAML adaptation has two hyperpameters. First one is the adaptation learning rate which is used to customize to the device model (η). Second hyperparameter is the number of gradient steps. This quantifies the number gradient updates to reach a device model from the meta model. We search the adaptation learning rate in range $\{0.1, 0.01\}$ and the number of gradient steps in range $\{1, 5\}$.

Different from MAML, Proto adaptation is a non parametric adaptation and it does not have extra hyperparameters.

Lastly, PFLDyn has α parameter. We search this parameters in range $\{0.1, 0.01\}$.

We run each method with the aforementioned hyperparameter search list for 100 communication rounds. Then, we pick the best performing configuration for each method and continue to run them for 1000 communication rounds.

Convergence curves. We give the convergence curves for CIFAR-10 and CIFAR-100 in Figure 2 and 3 respectively. We see that PFL based methods using Proto adaptation outperforms the baselines.

No personalization baselines strictly under-perform compared to personalization methods which shows a need to do personalization. We further investigate a case where no personalization methods are given a chance to personalize during inference time. We note that this does not effect training procedure. In no personalization baselines, the server model is used as the device model at each device without personalization. We consider another inference where the server model is personalized at each device using Proto or MAML adaptation. We found out that Proto adaptation gives higher performance than MAML adaptation. However, the performance is still worse than PFLDyn (Proto). We present no personalization baselines with using direct server model and Proto adaptation in inference time as well as PFLDyn (Proto) for CIFAR-10 and CIFAR-100 in Figure 4 and 5 respectively. Methods that perform poorly are omitted from the plots. Even though customization during inference time helps, PFLDyn (Proto) still outperforms the baselines.

The highest and the lowest level of personalization comparison. The Average level personalization metric has been reported in Table 1 and 2. We report the comparison of methods in the highest and the lowest level of personalization metrics for ACID and ALID settings in Table 3 and 4 respectively. Similar to Table 1 and 2, PFLDyn (Proto) method outperforms Fallah et al. (2020).

Implementation best practices. We give some subtle details of the implementation we think as useful practices in the following.

- No personalization baselines draw one batch of data at each round of SGD steps. Different from no personalization baselines, we draw two batches of data for P-Avg (MAML), Fallah et al. (2020) and PFL based methods. For MAML adaptation, first batch is used to customize the meta model into device model and the second batch is used to take gradient

Table 3. The number of model transmissions relative to one round of Fallah et al. (2020) required to reach the target test accuracy for the highest and the lowest level personalization performance in the Active Class Induced Diversity (ACID) scenario. Target accuracies are selected among the highest accuracy of our methods and the highest accuracy of the competing method Fallah et al. (2020). The methods without personalization are omitted due to their poor performance levels. The best method is highlighted and gain with respect to Fallah et al. (2020) method is shown.

Test Performance	Dataset	Accuracy	Fallah et al. (2020)	PFLDyn (Proto)	PFLDyn (MAML)	PFLScaf (Proto)	PFLScaf (MAML)	P-Avg (Proto)	Gain
Highest Level Personalization	**3 Classes per Device**								
	CIFAR-10	100.0	381	**83**	111	118	874	154	**4.6**×
		99.0	106	64	86	106	186	**54**	**2.0**×
	CIFAR-100	100.0	>1000	**144**	388	>1000	>1000	539	>**7.0**×
		99.0	312	**76**	170	990	990	112	**4.1**×
	5 Classes per Device								
	CIFAR-10	99.0	297	**199**	265	680	886	352	**1.5**×
		98.0	221	**114**	199	224	518	170	**1.9**×
	CIFAR-100	99.0	>1000	463	168	**148**	>1000	294	>**6.0**×
		98.0	121	165	166	92	532	**60**	**2.0**×
	7 Classes per Device								
	CIFAR-10	96.0	358	142	**134**	422	656	170	**2.7**×
		95.0	288	**123**	128	336	646	161	**2.3**×
	CIFAR-100	98.0	363	**286**	>1000	532	>1000	397	**1.3**×
		97.0	329	285	320	370	>1000	**281**	**1.2**×
Lowest Level Personalization	**3 Classes per Device**								
	CIFAR-10	80.0	>1000	522	638	780	>1000	**483**	>**2.1**×
		79.0	512	312	**211**	474	>1000	482	**2.4**×
	CIFAR-100	75.0	>1000	**254**	949	750	750	714	>**3.9**×
		66.0	950	**127**	365	660	660	275	**7.5**×
	5 Classes per Device								
	CIFAR-10	76.0	>1000	**240**	698	982	>1000	284	>**4.2**×
		75.0	585	207	**159**	582	892	213	**3.7**×
	CIFAR-100	71.0	857	238	**150**	674	>1000	848	**5.7**×
		70.0	817	235	**148**	510	>1000	284	**5.5**×
	7 Classes per Device								
	CIFAR-10	77.0	782	**180**	306	708	>1000	487	**4.3**×
		76.0	409	**123**	305	492	742	393	**3.3**×
	CIFAR-100	73.0	>1000	**195**	287	616	>1000	825	>**5.1**×
		71.0	307	**160**	252	362	672	538	**1.9**×

with respect to the meta model as in (Finn et al., 2017). For Proto adaptation, first batch is used to construct the class representations $\boldsymbol{c}_{\boldsymbol{w}}^{i,k}$ for all classes k in device i. Then, the second batch is used to calculate the loss of this representation. The first and second batches corresponds to support and query samples respectively according to (Snell et al., 2017).

- During inference time, we adapt meta model using all available training data data for each device. Namely, if MAML adaptation is used, the meta model is updated with the gradient using all training data. In case of proto adaptation, class representations are derived using all available training data. This is for reporting purposes.
- Personalized federated learning is an iterative process where a global meta model is updated over communication rounds. To increase stability of the algorithm, we perform gradient clipping at each device in each round. This stabilizes the cases where device meta models diverge.
- Gradient clipping increases stability. However, even with clipping some device meta models can diverge. This failure in one device causes the server meta model to diverge. To avoid this effect, we check each device before averaging the parameters. If a device meta model has been diverged, we do not include that device in our server model update. We found that this is a rare case but it improves the stability of the algorithms.
- We report performance of the meta model in our tables and figures. There are two options for the meta model which are average of active device meta models and average of all device meta models. We found that having average of all devices meta models give smoother curves than former one. We note that this is just for reporting purposes and we do not change the training dynamics.

A.2. Ablative Analysis of PFL

Analysis of α parameter. We test the sensitivity of α hyperparameter in CIFAR-10, 5 class per device ALID setting using Proto adaptation. By freezing other hyperparameters, we train PFLDyn (Proto) models with α varies in a logarithmic range as $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}\}$. The highest average test accuracies obtained are $\{89.0\%, 89.5\%, 90.0\%, 89.9\%, 89.0\%\}$. The performances are close to each other as such they differ within 1% for the α range.

miniImageNet dataset (Vinyals et al., 2016). We further compare the algorithms in miniImageNet dataset. miniImageNet dataset is a subset of ImageNet ILSVRC-2012 (Deng et al., 2009). There are a total of 100 classes where each class has 600 images. The images in miniImageNet are more realistic and harder than CIFAR-100. To use miniImageNet in personalized

Table 4. The number of model transmissions relative to one round of Fallah et al. (2020) required to reach the target test accuracy for the highest and the lowest level personalization performance in the Anonymous Label Induced Diversity (ALID) scenario. Target accuracies are selected among the highest accuracy of our methods and the highest accuracy of the competing method Fallah et al. (2020). The methods without personalization are omitted due to their poor performance levels. The best method is highlighted and gain with respect to Fallah et al. (2020) method is shown.

Test Performance	Dataset	Accuracy	Fallah et al. (2020)	PFLDyn (Proto)	PFLDyn (MAML)	PFLScaf (Proto)	PFLScaf (MAML)	P-Avg (Proto)	Gain
Highest Level Personalization	**3 Classes per Device**								
	CIFAR-10	100.0	>1000	**84**	92	126	216	173	>**11.9**×
		99.0	153	**59**	73	86	114	83	**2.6**×
	CIFAR-100	100.0	>1000	**133**	685	342	>1000	184	>**7.5**×
		97.0	134	**30**	49	62	126	44	**4.5**×
	5 Classes per Device								
	CIFAR-10	99.0	>1000	**110**	641	478	>1000	547	>**9.1**×
		96.0	123	79	99	172	360	**78**	**1.6**×
	CIFAR-100	100.0	>1000	683	**445**	>1000	>1000	628	>**2.2**×
		97.0	552	122	143	124	446	**41**	**13.5**×
	7 Classes per Device								
	CIFAR-10	98.0	>1000	**245**	475	432	>1000	350	>**4.1**×
		91.0	343	**70**	83	138	300	74	**4.9**×
	CIFAR-100	94.0	>1000	**185**	450	272	968	225	>**5.4**×
		88.0	948	**63**	144	120	478	82	**15.0**×
Lowest Level Personalization	**3 Classes per Device**								
	CIFAR-10	73.0	>1000	**114**	813	350	>1000	278	>**8.8**×
		69.0	710	**100**	250	280	586	166	**7.1**×
	CIFAR-100	73.0	>1000	**192**	721	662	>1000	692	>**5.2**×
		61.0	391	**78**	256	188	846	109	**5.0**×
	5 Classes per Device								
	CIFAR-10	72.0	>1000	**100**	245	306	988	263	>**10.0**×
		61.0	349	60	68	142	292	**55**	**6.3**×
	CIFAR-100	69.0	>1000	330	634	552	>1000	**209**	>**4.8**×
		64.0	896	258	243	300	898	**153**	**5.9**×
	7 Classes per Device								
	CIFAR-10	75.0	>1000	**177**	423	546	>1000	241	>**5.7**×
		63.0	402	**54**	74	130	276	60	**7.4**×
	CIFAR-100	65.0	>1000	165	400	206	>1000	**155**	>**6.5**×
		56.0	934	**58**	231	104	434	65	**16.1**×

Table 5. The number of model transmissions relative to one round of Fallah et al. (2020) required to reach the target test accuracy for the highest, the average and the lowest level personalization performance in miniImageNet, 5 class per device Anonymous Label Induced Diversity (ALID) scenario. Target accuracies are selected among the highest accuracy of our methods and the highest accuracy of the competing method Fallah et al. (2020). The methods without personalization are omitted due to their poor performance levels. The best method is highlighted and gain with respect to Fallah et al. (2020) method is shown.

Test Performance	Accuracy	Fallah et al. (2020)	PFLDyn (Proto)	PFLDyn (MAML)	PFLScaf (Proto)	PFLScaf (MAML)	P-Avg (Proto)	Gain
Highest Level Personalization	94.0	> 1000	**171**	499	370	> 1000	960	>**5.8**×
	90.0	977	**73**	123	152	332	131	**13.4**×
Average Personalization	74.5	> 1000	**141**	353	338	> 1000	349	>**7.1**×
	66.2	943	**63**	137	152	386	83	**14.9**×
Lowest Level Personalization	53.0	> 1000	**140**	224	264	> 1000	487	>**7.1**×
	43.0	778	67	151	124	504	**66**	**11.7**×

setting, we first split dataset into training and test data points as such it becomes a dataset consists of 50000 training and 10000 test points. Then, we repeat the federated dataset generation procedure as explained in Appendix A.1.

We consider ALID, 5 classes per device setting with 100 deivices and 10% participation ratio. Table 5 shows the performances of methods. As seen in the table, PFLDyn (Proto) leads to high communication savings.

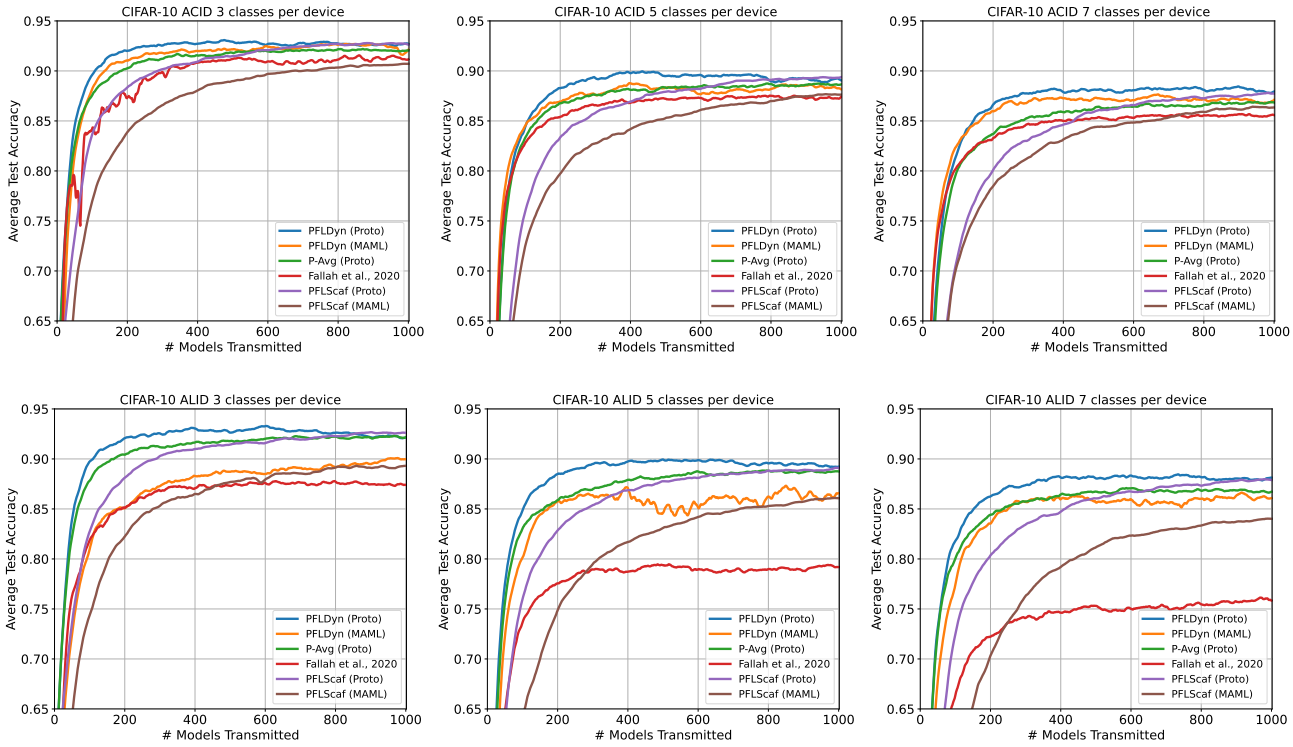

Figure 2. Smoothed convergence curves of methods in CIFAR-10 for average test accuracy among devices.

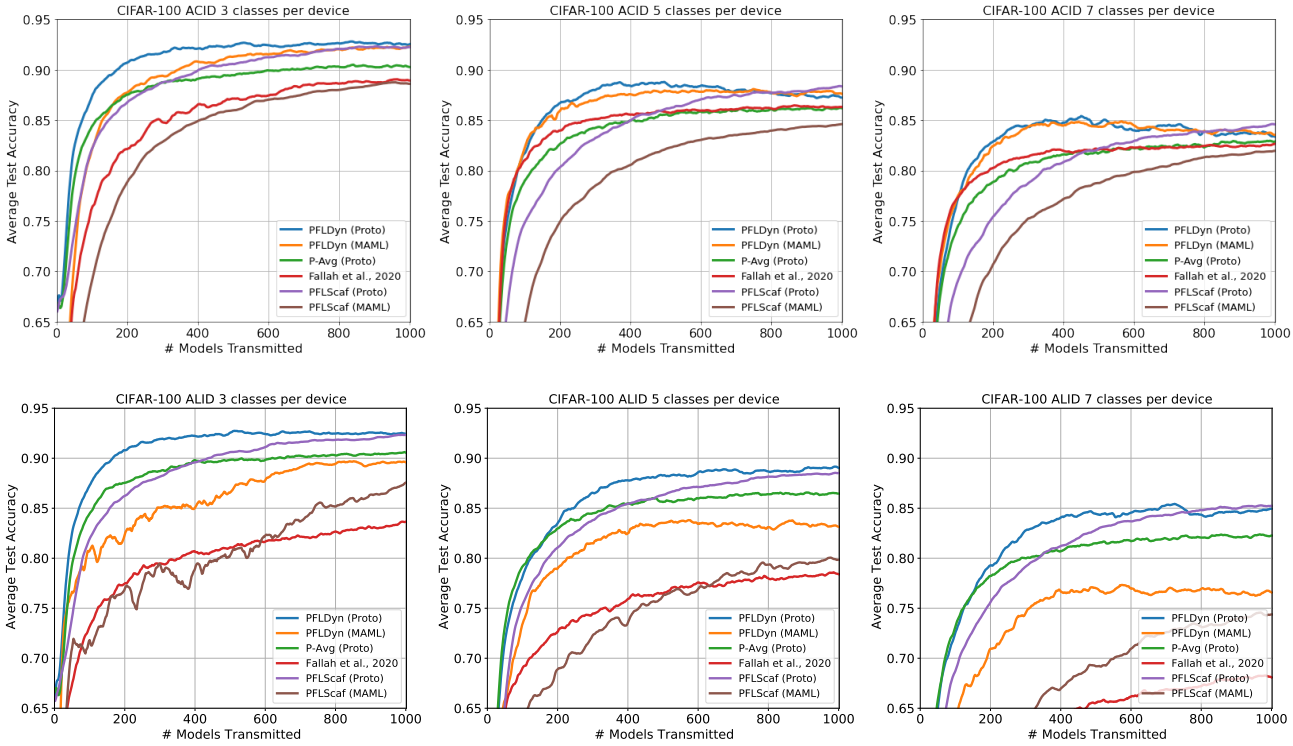

Figure 3. Smoothed convergence curves of methods in CIFAR-100 for average test accuracy among devices.

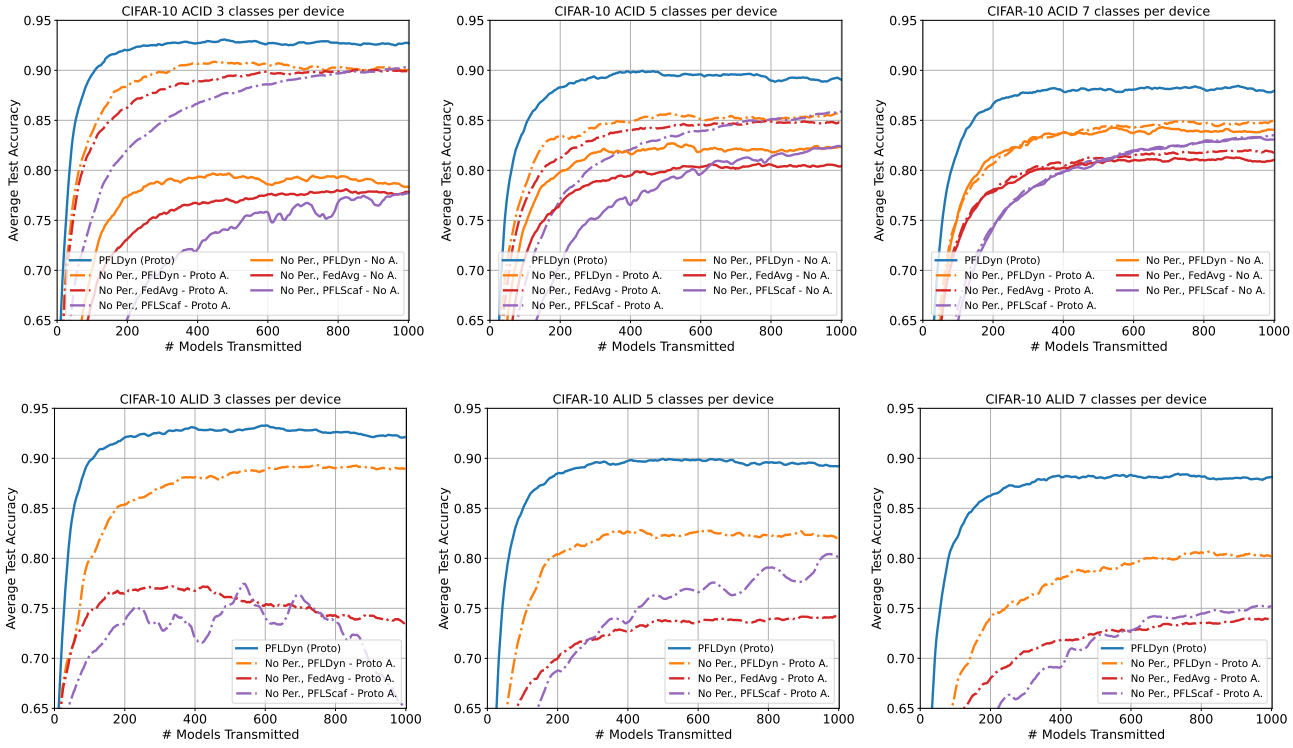

Figure 4. Smoothed convergence curves in CIFAR-10 of PFLDyn (Proto) and no customization baselines without adaptation and with Proto adaptation in inference time.

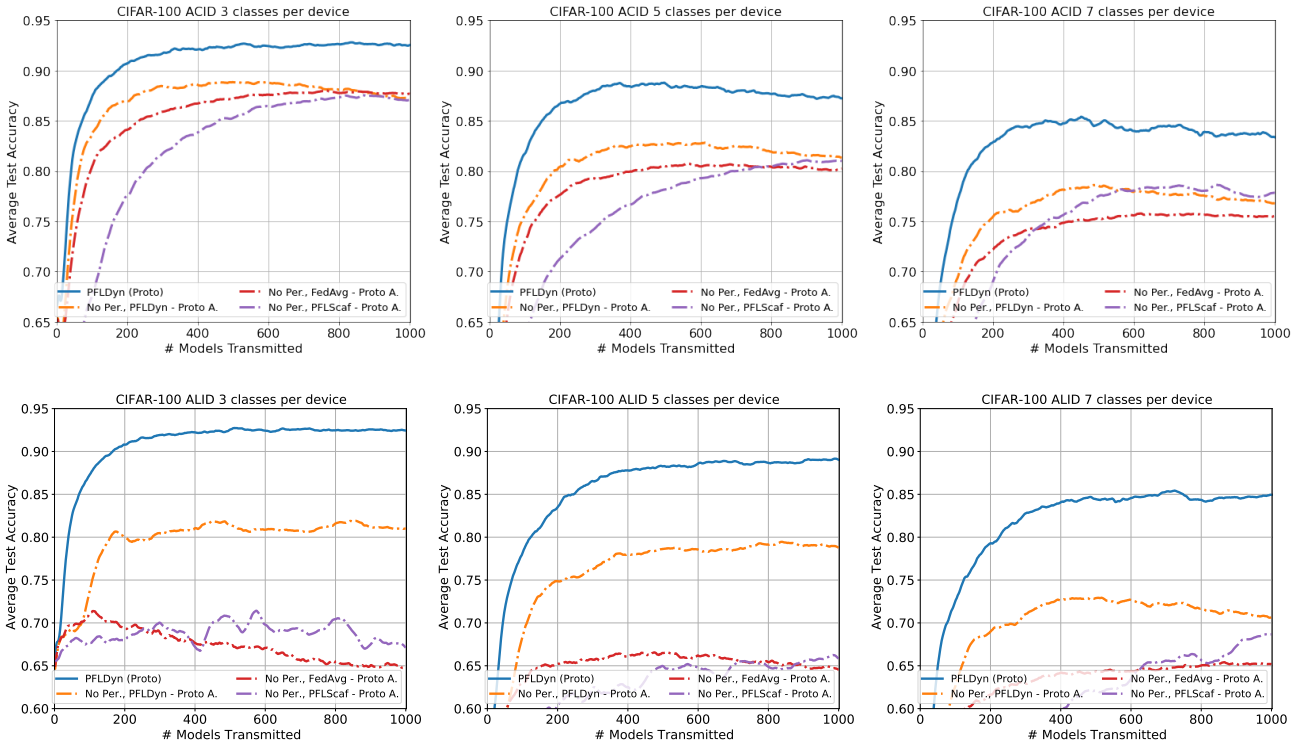

Figure 5. Smoothed convergence curves in CIFAR-100 of PFLDyn (Proto) and no customization baselines without adaptation and with Proto adaptation in inference time.

A.3. Proof

In this section, we mainly follow the analysis in Karimireddy et al. (2019) and Acar et al. (2021a) by modifying device functions so that we now need to consider $f_i \circ T_i$. Additionally, we set variance to be 0 ($\sigma = 0$) and allow for arbitrary SGD updates to ensure reaching a stationary point at each round. While the proof is straightforward, and follows Karimireddy et al. (2019) and Acar et al. (2021a), these modifications make it somewhat necessary to write them down in detail. For the sake of completeness, and for clarity we give the detailed proof here.

We first state our assumptions and the necessary notations, then we give proof of Theorem 1 separately in the following subsections. A similar analysis can be done for PFLScaf by extending the proof in Karimireddy et al. (2019).

A.3.1. Assumptions & Notations

Assumption 1. *(Stationary point) We assume that PFLDyn finds a stationary point of the customized loss it minimizes. Formally, PFLDyn satisfies*

$$\nabla f_i(\overline{\boldsymbol{w}}_i^{t+1}) + \nabla \mathcal{R}_i^t(\boldsymbol{w}_i^{t+1}) = \mathbf{0} \implies \nabla f_i(\overline{\boldsymbol{w}}_i^{t+1}) - \boldsymbol{g}_i^t + \alpha(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t) = \mathbf{0} \tag{6}$$

where $\overline{\boldsymbol{w}}_i^{t+1} = T_i(\boldsymbol{w}_i^{t+1})$.

Assumption 2. *(Smoothness)* $\{f_i \circ T_i\}_{i\in[m]}$ *functions are* L *smooth .i.e*

$$\|\nabla f_i \circ T_i(\boldsymbol{w}_1) - \nabla f_i \circ T_i(\boldsymbol{w}_2)\| \leq L\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \ \ \forall \boldsymbol{w}_1, \boldsymbol{w}_2, i \tag{7}$$

Smoothness imply the following inequality,

$$f_i \circ T_i(\boldsymbol{w}_2) - f_i \circ T_i(\boldsymbol{w}_1) \leq \langle \nabla f_i \circ T_i(\boldsymbol{w}_1), \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{L}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \ \ \forall \boldsymbol{w}_1, \boldsymbol{w}_2, i \tag{8}$$

If $\{f_i \circ T_i\}_{i\in[m]}$ *functions are* μ *strongly convex and* L *smooth, they satisfy,*

$$\frac{1}{2Lm}\sum_{i\in[m]} \|\nabla f_i \circ T_i(\boldsymbol{w}_1) - \nabla f_i \circ T_i(\boldsymbol{w}_*)\|^2 \leq F(\boldsymbol{w}_1) - F(\boldsymbol{w}_*) \ \ \forall \boldsymbol{w}_1, \tag{9}$$

$$-\langle \nabla f_i \circ T_i(\boldsymbol{w}_1), \boldsymbol{w}_3 - \boldsymbol{w}_2 \rangle \leq -f_i \circ T_i(\boldsymbol{w}_3) + f_i \circ T_i(\boldsymbol{w}_2) + \frac{L}{2}\|\boldsymbol{w}_3 - \boldsymbol{w}_1\|^2 - \frac{\mu}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 \ \ \forall \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, i \tag{10}$$

where $\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}} F(\boldsymbol{w})$.

Assumption 2 controls $\{f_i \circ T_i\}_{i\in[m]}$ functions. In MAML transformation, Lemma 4.2 (Fallah et al., 2020) states that this can be achieved for twice continuously differentiable, smooth and Lipschitz continuous f_is. Proto transformation based on Sigmoid function is smooth. Hence, if $\{f_i\}_{i\in[m]}$ are smooth, $\{f_i \circ T_i\}_{i\in[m]}$ functions are smooth.

To simplify the analysis, we set some notation prior to proofs. At each round, a set of devices $\mathcal{P}_t$ are chosen to be active and Algorithm 1 does not update stale devices in each round. We define $\boldsymbol{y}_i^{t+1}$ models as the models that satisfy

$$\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1}) - \boldsymbol{g}_i^t + \alpha(\boldsymbol{y}_i^{t+1} - \boldsymbol{w}^t) = \mathbf{0}. \tag{11}$$

If device i is an active device at time t, $i \in \mathcal{P}_t$, we have that $\boldsymbol{y}_i^{t+1} = \boldsymbol{w}_i^{t+1}$, otherwise $\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t$.

Combining Assumption 1 and $\boldsymbol{g}$ state update .i.e $\boldsymbol{g}_i^{t+1} = \boldsymbol{g}_i^t - \alpha\left(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right)$, we get an important relation as,

$$\boldsymbol{g}_i^{t+1} = \nabla f_i \circ T_i(\boldsymbol{w}_i^{t+1}) \tag{12}$$

where we see that $\boldsymbol{g}_i$ states store gradient information.

$\boldsymbol{g}_i$ state update in local devices and $\boldsymbol{g}$ state update in the server reveals that,

$$\boldsymbol{g}^{t+1} = \frac{1}{m}\sum_{i\in[m]} \boldsymbol{g}_i^{t+1} = \frac{1}{m}\sum_{i\in[m]} \nabla f_i \circ T_i(\boldsymbol{w}_i^{t+1}) \tag{13}$$

where we see that $\boldsymbol{g}$ state store the average gradient information.

We give convergence analysis in terms of the average device meta models. We define average meta model at each round as $\boldsymbol{m}^t = \frac{1}{P}\sum_{i\in\mathcal{P}_t} \boldsymbol{w}_i^t$. Using $\boldsymbol{g}$ state update, we can relate $\boldsymbol{m}^t$ model to PFLDyn Algorithm as $\boldsymbol{w}^{t+1} = \boldsymbol{m}^{t+1} - \frac{1}{\alpha}\boldsymbol{g}^{t+1}$. We continue to define some control variables for the analysis as,

$$E^t = \frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^t - \boldsymbol{m}^{t-1}\|^2,\;\; B^t = \frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t) - \nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2,\;\; C^t = \frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{w}_i^t - \boldsymbol{m}^t\|^2.$$

We use E^t and B^t variables in convex analysis and E^t and C^t in nonconvex analysis. Intuitively, if PFLDyn Algorithm converges as in Proposition 1, all these variables will converge to 0.

A.3.2. Proof of Proposition 1

As stated in Proposition 1, we assume that the device meta models converge. Eq. 12 implies $\boldsymbol{g}_i$s converge as, $\lim_{t\to\infty}\boldsymbol{w}_i^t = \boldsymbol{w}_i^\infty \implies \lim_{t\to\infty}\boldsymbol{g}_i^t = \nabla f_i\left(\overline{\boldsymbol{w}}_{i,i}^\infty\right)$ where $\overline{\boldsymbol{w}}_{i,i}^\infty = T_i(\boldsymbol{w}_i^\infty)$.

Convergence of $\boldsymbol{g}_i$s and the update rule, $\boldsymbol{g}_i^{t+1} = \boldsymbol{g}_i^t - \alpha\left(\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^t\right)$, imply $\boldsymbol{w}_i^\infty = \boldsymbol{w}^\infty$ and $\overline{\boldsymbol{w}}_{i,i}^\infty = T_i(\boldsymbol{w}^\infty)$ i.e. each device meta model converges to the same meta model. Rearranging the server update gives $\boldsymbol{g}^t = \alpha\left(-\boldsymbol{w}^t + \frac{1}{|\mathcal{P}_t|}\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^t\right)$. Since we have $\boldsymbol{w}_i^\infty = \boldsymbol{w}^\infty$ for all is, we get $\lim_{t\to\infty}\boldsymbol{g}^t = \boldsymbol{0}$. Using Eq. 13 we conclude that $\lim_{t\to\infty}\boldsymbol{g}^t = \frac{1}{m}\sum_{i\in[m]}\nabla f_i\left(\overline{\boldsymbol{w}}_i^\infty\right) = \boldsymbol{0}$ where $\overline{\boldsymbol{w}}_i^\infty = T_i(\boldsymbol{w}^\infty)$. Hence, PFL eliminates the bias coming from heterogeneity of devices and it converges to a stationary point of the personalized federated learning objective OPT. □

A.3.3. Strongly Convex Analysis

Theorem 2. *If $\{f_i\circ T_i\}_{i\in[m]}$ functions are μ strongly convex & L smooth, PFLDyn Algorithm satisfies*

$$E\left[F(\boldsymbol{M}^T) - F(\boldsymbol{w}_*)\right] \leq \frac{1}{z^{T-1}} O\left(\alpha D + \frac{m}{P}G\right)$$

where $\boldsymbol{M}^T = \frac{1}{Z}\sum_{t=1}^T z^{t-1}\boldsymbol{m}^t$ is the weighted average of $\boldsymbol{m}^t$ meta models, $\boldsymbol{m}^t = \frac{1}{P}\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^t$ is the average of active device meta models at time t, $z^t = \left(1+\frac{\mu}{\alpha}\right)^t$ weights of the models, $Z = \sum_{t=1}^T z^{t-1}$ is the normalization coefficient, $\alpha = 50\left(\frac{m}{P}\mu + L\right)$ is the hyperparameter, $\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}} F(\boldsymbol{w})$ is the best meta model, $D = \left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2$ is the distance between the initial model and the best meta model, , $G = \frac{1}{m}\sum_{i\in[m]}\left\|\nabla f_i\left(\overline{\boldsymbol{w}}_i^*\right)\right\|^2$, $\overline{\boldsymbol{w}}_i^* = T_i(\boldsymbol{w}^*)$ is a problem dependent constant and the expectation is with respect to randomness due to active device set at each round ($\mathcal{P}_t$).

Due to $\frac{1}{z^T}$ coefficient on RHS of Theorem 2, we conclude that ϵ error can be obtained in $T = O\left(\ln\frac{1}{\epsilon}\right)$ communication rounds. To prove Theorem 2, we start with the following Lemma,

Lemma 1. *If $\{f_i\circ T_i\}_{i\in[m]}$ functions are μ strongly convex & L smooth and $\alpha = 50\left(\frac{m}{P}\mu + L\right)$, Algorithm 1 satisfies*

$$\left(1+\frac{\mu}{\alpha}\right)\left(E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 + rB^{t+1}\right) \leq E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 + rB^t - \frac{1}{5\alpha}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right]$$

where $r = 16\frac{m}{\alpha}\frac{1}{P\alpha + P\mu - m\mu}$.

Multiplying Lemma 1 with $\left(1+\frac{\mu}{\alpha}\right)^{t-1}$ and summing over t give telescoping terms. Rearranging the resulting sum gives,

$$\frac{1}{5\alpha}\sum_{t=1}^T\left(1+\frac{\mu}{\alpha}\right)^{t-1} E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] \leq \left(E\|\boldsymbol{m}^1 - \boldsymbol{w}_*\|^2 + rB^1\right) - \left(1+\frac{\mu}{\alpha}\right)^T\left(E\|\boldsymbol{m}^{T+1} - \boldsymbol{w}_*\|^2 + rB^{T+1}\right)$$

Eliminating the non-negative term and dividing both sides with $\frac{1}{5\alpha}Z$ where $Z = \sum_{t=1}^T\left(1+\frac{\mu}{\alpha}\right)^{t-1}$ give,

$$\sum_{t=1}^T E\left[\frac{\left(1+\frac{\mu}{\alpha}\right)^{t-1}}{Z}F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] \leq \frac{1}{Z}5\alpha\left(E\|\boldsymbol{m}^1 - \boldsymbol{w}_*\|^2 + rB^1\right)$$

Finally, applying Jensen Inq. on LHS gives

$$E\left[F(\boldsymbol{M}^T) - F(\boldsymbol{w}_*)\right] \leq \frac{1}{Z} 5\alpha \left(E\|\boldsymbol{m}^1 - \boldsymbol{w}_*\|^2 + rB^1\right)$$

where $\boldsymbol{M}^T = \frac{1}{Z}\sum_{t=1}^{T}\left(1+\frac{\mu}{\alpha}\right)^{t-1}\boldsymbol{m}^t$. This proves bound in Theorem 2.

To prove Lemma 1, we relate $\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2$ to $\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2$ by expressing the difference as $\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 = \|(\boldsymbol{m}^t - \boldsymbol{w}_*) + (\boldsymbol{m}^{t+1} - \boldsymbol{m}^t)\|^2$ similar to gradient descent analysis. We first give a set of Lemmas that are used for the terms arising in the analysis and prove them at the end.

Lemma 2. *($\boldsymbol{m}$ difference relation)*

$$E\left[\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\right] = -\frac{1}{\alpha m}\sum_{i\in[m]} E\left[\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1})\right].$$

Lemma 3. *($\boldsymbol{m}$ difference bound)*

$$E\left\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\right\|^2 \leq E^{t+1}.$$

We start expanding the term as,

$$\begin{aligned}
E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 =& E\|\boldsymbol{m}^t - \boldsymbol{m}_* + \boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
=& E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 + 2E\left[\left\langle \boldsymbol{m}^t - \boldsymbol{w}_*, \boldsymbol{m}^{t+1} - \boldsymbol{m}^t\right\rangle\right] + E\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
=& E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 + \frac{2}{\alpha m}\sum_{i\in[m]} E\left[\left\langle \boldsymbol{m}^t - \boldsymbol{w}_*, -\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1})\right\rangle\right] + E\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
\leq& \frac{2}{\alpha m}\sum_{i\in[m]} E\left[f_i \circ T_i(\boldsymbol{w}_*) - f_i \circ T_i(\boldsymbol{m}^t) + \frac{L}{2}\|\boldsymbol{y}_i^{t+1} - \boldsymbol{m}^t\|^2 - \frac{\mu}{2}\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}_*\|^2\right] \\
&+ E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 + E\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
=& E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 - \frac{2}{\alpha}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] + \frac{L}{\alpha}E^{t+1} - \frac{\mu}{\alpha}\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}_*\|^2 + E\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
\leq& E\|\boldsymbol{m}^t - \boldsymbol{w}_*\|^2 - \frac{2}{\alpha}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] + \left(\frac{L}{\alpha} + 1\right)E^{t+1} - \frac{\mu}{\alpha}\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}_*\|^2 \qquad (14)
\end{aligned}$$

where we use Lemma 2, Inq. 10 and Lemma 3.

We introduce more Lemmas to handle E^{t+1} and $E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2$ terms.

Lemma 4. *(E^{t+1} bound)*

$$\left(1 - \frac{4L^2}{\alpha^2}\right)E^{t+1} \leq \frac{8}{\alpha^2}B^t + \frac{8L}{\alpha^2}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right].$$

Lemma 5. *(B^{t+1} bound)*

$$B^{t+1} \leq \left(1 - \frac{P}{m}\right)B^t + 2L^2\frac{P}{m}E^{t+1} + 4L\frac{P}{m}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right].$$

Lemma 6. *($\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}_*\|^2$ bound)*

$$E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 \leq \frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}_*\|^2.$$

Let us sum Inq. 14, Lemma 4, 5 and 6 where Lemma 4, 5 and 6 are scaled with 2, $16\frac{m}{\alpha^2}\frac{\mu+\alpha}{P\alpha+P\mu-m\mu}$ and $\frac{\mu}{\alpha}$ respectively. Let $\alpha = 50\left(\frac{m}{P}\mu + L\right)$ and ignore non-positive terms, then we get Lemma 1. □

We give proof of the Lemmas used above.

Proof of Lemma 2.

$$
\begin{aligned}
E\left[\boldsymbol{m}^{t+1}-\boldsymbol{m}^{t}\right] &= E\left[\left(\frac{1}{P}\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^{t+1}\right)-\boldsymbol{w}^t-\frac{1}{\alpha}\boldsymbol{g}^t\right] = E\left[\frac{1}{P}\sum_{i\in\mathcal{P}_t}\left(\boldsymbol{w}_i^{t+1}-\boldsymbol{w}^t-\frac{1}{\alpha}\boldsymbol{g}^t\right)\right] \\
&= E\left[\frac{1}{\alpha P}\left(\sum_{i\in\mathcal{P}_t}\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_i^{t+1})-\boldsymbol{g}^t\right)\right] \\
&= E\left[\frac{1}{\alpha P}\left(\sum_{i\in\mathcal{P}_t}\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\boldsymbol{g}^t\right)\right] \\
&= E\left[\frac{1}{\alpha m}\left(\sum_{i\in[m]}\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\boldsymbol{g}^t\right)\right] = -\frac{1}{\alpha m}\sum_{i\in[m]}E\left[\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})\right]
\end{aligned}
$$

where we use definition of $\boldsymbol{m}^t$ and $\boldsymbol{y}^t$. We then use tower property where we take expectation conditioned on randomness before time t. In this case, only $\mathcal{P}_t$ remains as a random variable. We take expectation noting that the probability of each device becoming active is $\frac{P}{m}$. Finally, we use definition of $\boldsymbol{g}^t$ in Eq. 13. □

Proof of Lemma 3.

$$
E\|\boldsymbol{m}^{t+1}-\boldsymbol{m}^t\|^2 = E\left\|\frac{1}{P}\left(\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^{t+1}-\boldsymbol{m}^t\right)\right\|^2 \leq \frac{1}{P}E\left[\sum_{i\in\mathcal{P}_t}\left\|\boldsymbol{w}_i^{t+1}-\boldsymbol{m}^t\right\|^2\right] = \frac{1}{P}E\left[\sum_{i\in\mathcal{P}_t}\left\|\boldsymbol{y}_i^{t+1}-\boldsymbol{m}^t\right\|^2\right] = E^{t+1}
$$

where we use definition of $\boldsymbol{m}^t$ in the first equality. We then apply Jensen Inq on $\|\|^2$ function. Finally, we use tower property and take expectation with respect to randomness prior to time t similar by noting that the probability of each device becoming active is $\frac{P}{m}$. We arrive the definition of E^{t+1}. □

Lemma 7.

$$
E\left\|\boldsymbol{g}^t\right\|^2 \leq B^t.
$$

Proof.

$$
\begin{aligned}
E\left\|\boldsymbol{g}^t\right\|^2 =& E\left\|\frac{1}{m}\sum_{i\in[m]}\nabla f_i\circ T_i(\boldsymbol{w}_i^t)\right\|^2 = E\left\|\frac{1}{m}\left(\sum_{i\in[m]}\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\right)\right\|^2 \\
\leq& \frac{1}{m}\sum_{i\in[m]}E\left\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\right\|^2 = B^t
\end{aligned}
$$

where we use Eq. 13, optimality condition of $\boldsymbol{w}_*$ as $\sum_{i\in[m]}\nabla f_i\circ T_i(\boldsymbol{w}_*)=\mathbf{0}$ and Jensen Inq. on $\|\|$ function. □

Proof of Lemma 4.

$$
\begin{aligned}
E^{t+1} =& \frac{1}{m}\sum_{i\in[m]}E\|\boldsymbol{y}_i^{t+1}-\boldsymbol{m}^t\|^2 = \frac{1}{m}\sum_{i\in[m]}E\left\|\boldsymbol{y}_i^{t+1}-\boldsymbol{w}^t-\frac{1}{\alpha}\boldsymbol{g}^t\right\|^2 \\
=& \frac{1}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\boldsymbol{g}^t\|^2 \\
=& \frac{1}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)+\nabla f_i\circ T_i(\boldsymbol{w}_*)-\nabla f_i\circ T_i(\boldsymbol{m}^t)+\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\boldsymbol{g}^t\|^2 \\
\leq& \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2+\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{m}^t)\|^2+\frac{4}{\alpha^2}E\|\boldsymbol{g}^t\|^2\\
\leq&\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2+\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
&+\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{m}^t)\|^2+\frac{4}{\alpha^2}B^t\\
\leq&\frac{8}{\alpha^2}B^t+\frac{4L^2}{\alpha^2}E^{t+1}+\frac{8L}{\alpha^2}E\left[F(\boldsymbol{m}^t)-F(\boldsymbol{w}_*)\right]
\end{aligned}
$$

where we use definition of $\boldsymbol{m}^t$, Eq. 11, Jensen Inq. on $\|\|^2$ function, Lemma 7 and smoothness Inq. 7 and Inq. 9. Rearranging the terms give the statement in the Lemma. □

Proof of Lemma 5.

$$
\begin{aligned}
B^{t+1}=&\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
=&\left(1-\frac{P}{m}\right)\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2+\frac{P}{m}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
=&\left(1-\frac{P}{m}\right)B^t+\frac{P}{m}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{m}^t)+\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
\leq&\left(1-\frac{P}{m}\right)B^t+\frac{2P}{m}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1})-\nabla f_i\circ T_i(\boldsymbol{m}^t)\|^2+\frac{2P}{m}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
\leq&\left(1-\frac{P}{m}\right)B^t+\frac{2L^2P}{m}E^{t+1}+\frac{2P}{m}\frac{1}{m}\sum_{i\in[m]}E\|\nabla f_i\circ T_i(\boldsymbol{m}^t)-\nabla f_i\circ T_i(\boldsymbol{w}_*)\|^2\\
\leq&\left(1-\frac{P}{m}\right)B^t+\frac{2L^2P}{m}E^{t+1}+\frac{4LP}{m}E\left[F(\boldsymbol{m}^t)-F(\boldsymbol{w}_*)\right]
\end{aligned}
$$

where we use definition of B^{t+1}, tower property by taking expectation with respect to randomness prior to time t, Jensen Inq., smoothness Inq. 7 and Inq. 9. □

Proof of Lemma 6.

$$
E\|\boldsymbol{m}^t-\boldsymbol{w}_*\|^2=E\left\|\frac{1}{P}\left(\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^t-\boldsymbol{w}_*\right)\right\|^2\leq\frac{1}{P}E\left[\sum_{i\in\mathcal{P}_t}\left\|\boldsymbol{w}_i^t-\boldsymbol{w}_*\right\|^2\right]=\frac{1}{P}E\left[\sum_{i\in\mathcal{P}_t}\left\|\boldsymbol{y}_i^t-\boldsymbol{w}_*\right\|^2\right]=\frac{1}{m}\sum_{i\in[m]}E\left\|\boldsymbol{y}_i^t-\boldsymbol{w}_*\right\|^2
$$

where we use definition of $\boldsymbol{m}^t$, Jensen Inq., definition of $\boldsymbol{y}_i^t$ and tower property by conditioning on randomness prior to time t. Rearranging the terms gives the statement. □

A.3.4. Convex Analysis

Theorem 3. *If $\{f_i\circ T_i\}_{i\in[m]}$ functions are convex & L smooth, PFLDyn Algorithm satisfies*

$$
E\left[F(\boldsymbol{M}^T)-F(\boldsymbol{w}_*)\right]\leq\frac{1}{T}O\left(\sqrt{\frac{m}{P}}\left(LD+\frac{1}{L}G\right)\right)
$$

where $\boldsymbol{M}^T=\frac{1}{T}\sum_{t=1}^T\boldsymbol{m}^t$ is the weighted average of $\boldsymbol{m}^t$ meta models, $\boldsymbol{m}^t=\frac{1}{P}\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^t$ is the average of active device meta models at time t, $\boldsymbol{w}_*=\arg\min_{\boldsymbol{w}}F(\boldsymbol{w})$ is the best meta model, $D=\left\|\boldsymbol{w}^1-\boldsymbol{w}^*\right\|^2$ is the distance between the initial model and the best meta model, , $G=\frac{1}{m}\sum_{i\in[m]}\|\nabla f_i\left(\overline{\boldsymbol{w}}_i^*\right)\|^2$, $\overline{\boldsymbol{w}}_i^*=T_i(\boldsymbol{w}^*)$ is a problem dependent constant and the expectation is with respect to randomness due to active device set at each round ($\mathcal{P}_t$).

We need $T = O\left(\frac{1}{\epsilon}\sqrt{\frac{m}{P}}\left(LD + \frac{1}{L}G\right)\right)$ communication rounds to reach ϵ expected precision as stated in Theorem 1.

The proof Theorem 3 is similar to strongly convex case. First we state a Lemma in which sum over time gives the statement in the Theorem.

Lemma 8. *If $\{f_i \circ T_i\}_{i\in[m]}$ functions are convex & L smooth $\alpha = 50L\sqrt{\frac{m}{P}}$, Algorithm 1 satisfies*

$$E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 + \frac{16}{\alpha^2}\frac{m}{P}B^{t+1} \leq E\|\boldsymbol{m}^{t} - \boldsymbol{w}_*\|^2 + \frac{16}{\alpha^2}\frac{m}{P}B^{t} - \frac{1}{4\alpha}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right].$$

Summing Lemma 8 over time gives,

$$\frac{1}{4\alpha}\sum_{t=1}^{T} E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] \leq \left(E\|\boldsymbol{m}^{1} - \boldsymbol{w}_*\|^2 + \frac{16}{\alpha^2}\frac{m}{P}B^{1}\right) - \left(E\|\boldsymbol{m}^{T+1} - \boldsymbol{w}_*\|^2 + \frac{16}{\alpha^2}\frac{m}{P}B^{T+1}\right)$$

Eliminating the non-negative term, dividing both sides with $\frac{T}{4\alpha}$ and applying Jensen give,

$$E\left[F(\boldsymbol{M}^T) - F(\boldsymbol{w}_*)\right] \leq \sum_{t=1}^{T} E\left[\frac{1}{T}F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] \leq \frac{1}{T}\alpha\left(\|\boldsymbol{m}^{1} - \boldsymbol{w}_*\|^2 + \frac{16}{\alpha^2}\frac{m}{P}B^{1}\right)$$

where $\boldsymbol{M}^T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{m}^t$. We reach to the statement in Theorem 3.

The proof of Lemma 8 follows the same process as in the proof for strongly convex case. Namely, we start with expanding $\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 = \|(\boldsymbol{m}^t - \boldsymbol{w}_*) + (\boldsymbol{m}^{t+1} - \boldsymbol{m}^t)\|^2$ term and relate $\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2$ to $\|\boldsymbol{m}^{t} - \boldsymbol{w}_*\|^2$. We can use the Lemmas defined for the strongly convex proof by noting that $\mu = 0$. Then, Inq. 14 gives,

$$E\|\boldsymbol{m}^{t+1} - \boldsymbol{w}_*\|^2 \leq E\|\boldsymbol{m}^{t} - \boldsymbol{w}_*\|^2 - \frac{2}{\alpha}E\left[F(\boldsymbol{m}^t) - F(\boldsymbol{w}_*)\right] + \left(\frac{L}{\alpha} + 1\right)E^{t+1} \tag{15}$$

Let us sum Inq. 15, Lemma 4, and 5 where Lemma 4 and 5 are scaled with 2 and $16\frac{m}{p}\frac{1}{\alpha^2}$ respectively. Let $\alpha = 50L\sqrt{\frac{m}{P}}$ and ignore non-positive terms, then we get Lemma 8. □

A.3.5. Nonconvex Analysis

Theorem 4. *If $\{f_i \circ T_i\}_{i\in[m]}$ functions are nonconvex & L smooth, PFLDyn Algorithm satisfies*

$$E\left[\left\|\nabla F\left(\boldsymbol{M}^T\right)\right\|^2\right] \leq O\left(\frac{1}{T}\left(L\frac{m}{P}\Delta_1 + L^2\Delta_2\right)\right)$$

where $\boldsymbol{M}^T = \boldsymbol{m}^\tau$ is a random model, τ is uniformly drawn from $\{1, 2, \ldots, T\}$, $\boldsymbol{m}^t = \frac{1}{P}\sum_{i\in\mathcal{P}_t}\boldsymbol{w}_i^t$ is the average of active device meta models at time t, $\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}} F(\boldsymbol{w})$ is the best meta model, Δ_1=$F(\boldsymbol{w}^1) - F(\boldsymbol{w}_*)$ is the distance between the initial model and the best meta model in the function values, Δ_2=$\frac{1}{m}\sum_{i\in[m]}\|\boldsymbol{w}_i^1 - \boldsymbol{w}^1\|^2$ is another problem dependent constant and the expectation is with respect to randomness due to active device set at each round ($\mathcal{P}_t$) and τ.

We need $T = O\left(\frac{1}{\epsilon}\left(L\frac{m}{P}\Delta_1 + L^2\Delta_2\right)\right)$ communication rounds to reach ϵ expected precision as stated in Theorem 1.

Similar to the previous proofs, we start with a Lemma as,

Lemma 9. *If $\{f_i \circ T_i\}_{i\in[m]}$ functions are nonconvex & L smooth $\alpha = 50L\frac{m}{P}$, Algorithm 1 satisfies*

$$E\left[F(\boldsymbol{m}^{t+1}) - F(\boldsymbol{w}_*)\right] + \frac{8L^3}{\alpha^2}\frac{2m-P}{P}C^{t+1} \leq E\left[F(\boldsymbol{m}^{t}) - F(\boldsymbol{w}_*)\right] + \frac{8L^3}{\alpha^2}\frac{2m-P}{P}C^{t} - \frac{1}{4\alpha}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2.$$

Summing Lemma 9 over time gives,

$$\frac{1}{4\alpha}\sum_{t=1}^{T} E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2 \leq \left(E\left[F(\boldsymbol{m}^{1}) - F(\boldsymbol{w}_*)\right] + \frac{8L^3}{\alpha^2}\frac{2m-P}{P}C^{1}\right) - \left(E\left[F(\boldsymbol{m}^{T+1}) - F(\boldsymbol{w}_*)\right] + \frac{8L^3}{\alpha^2}\frac{2m-P}{P}C^{T+1}\right)$$

$E\left[F(\boldsymbol{m}^{T+1}) - F(\boldsymbol{w}_*)\right] + \frac{8L^3}{\alpha^2}\frac{2m-P}{P}C^{T+1}$ term is non-negative. Dividing both sides with $\frac{1}{4\alpha}T$ and eliminating the non-negative term give,

$$\sum_{t=1}^{T}\frac{1}{T}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2 \leq \frac{1}{T}4\alpha\left(F(\boldsymbol{m}^1) - F(\boldsymbol{w}_*) + \frac{8L^3}{\alpha^2}\frac{2m}{P}C^1\right)$$

Consider an $\boldsymbol{m}^t$ model from a random time t. Let τ is uniformly drawn from $\{1, 2, \ldots, T\}$ and $\boldsymbol{M}^T = \boldsymbol{m}^\tau$ is the corresponding random model. Then we can express LHS as,

$$E\left\|\nabla F\left(\boldsymbol{m}^T\right)\right\|^2 = \sum_{t=1}^{T}\frac{1}{T}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2 \leq \frac{1}{T}4\alpha\left(F(\boldsymbol{m}^1) - F(\boldsymbol{w}_*) + \frac{8L^3}{\alpha^2}\frac{2m}{P}C^1\right)$$

where the expectation is with respect to randomness due to active device set at each round ($\mathcal{P}_t$) and τ. This inequality is the statement in Theorem 4.

The proof of Lemma 9 follows a similar idea in gradient descent analysis for nonconvex functions. We use the quadratic smoothness bound (Inq. 8) on $\boldsymbol{m}^t$ models. We start with Inq. 8 as,

$$\begin{aligned}
E\left[F(\boldsymbol{m}^{t+1})\right] - E\left[F(\boldsymbol{m}^t)\right] \leq& E\left[\left\langle\nabla F(\boldsymbol{m}^t), \boldsymbol{m}^{t+1} - \boldsymbol{m}^t\right\rangle\right] + \frac{L}{2}E\|\boldsymbol{m}^{t+1} - \boldsymbol{m}^t\|^2 \\
=& \frac{1}{\alpha}E\left[\left\langle\nabla F(\boldsymbol{m}^t), \frac{1}{m}\sum_{i\in[m]} -\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1})\right\rangle\right] + \frac{L}{2}E^{t+1} \\
\leq& \frac{1}{2\alpha}E\left\|\frac{1}{m}\left(\sum_{i\in[m]}\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1}) - \nabla f_i \circ T_i(\boldsymbol{m}^t)\right)\right\|^2 - \frac{1}{2\alpha}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2 + \frac{L}{2}E^{t+1} \\
\leq& \frac{1}{2\alpha}\frac{1}{m}\sum_{i\in[m]}E\left\|\nabla f_i \circ T_i(\boldsymbol{y}_i^{t+1}) - \nabla f_i \circ T_i(\boldsymbol{m}^t)\right\|^2 - \frac{1}{2\alpha}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2 + \frac{L}{2}E^{t+1} \\
\leq& \left(\frac{L^2}{2\alpha} + \frac{L}{2}\right)E^{t+1} - \frac{1}{2\alpha}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2
\end{aligned} \tag{16}$$

where we use Lemma 2, 3, inequality $\langle\boldsymbol{w}_1, \boldsymbol{w}_2\rangle \leq \frac{1}{2}\|\boldsymbol{w}_1 + \boldsymbol{w}_2\|^2 - \frac{1}{2}\|\boldsymbol{w}_1\|^2\ \forall \boldsymbol{w}_1, \boldsymbol{w}_2$, Jensen Inq. on $\|\|^2$ and smoothness 7.

We note that we need another set of Lemmas to bound E^{t+1} terms because Lemma 4 uses convexity. Let us introduce nonconvex equivalence of the lemmas as,

Lemma 10. *(Nonconvex E^{t+1} bound)*

$$\left(1 - \frac{4L^2}{\alpha^2}\right)E^{t+1} \leq \frac{8L^2}{\alpha^2}C^t + \frac{4}{\alpha^2}E\left\|\nabla F(\boldsymbol{m}^t)\right\|^2.$$

Lemma 11. *(Nonconvex C^{t+1} bound)*

$$C^{t+1} \leq 2\frac{m-P}{2m-P}C^t + 2\left(\frac{P}{2m-P} + \frac{m}{P}\right)E^{t+1}.$$

Let us sum Inq. 16, Lemma 10, and 11 where Lemma 10 and 11 are scaled with L and $\frac{8L^3}{\alpha^2}\frac{2m-P}{P}$ respectively. Let $\alpha = 50L\frac{m}{P}$ and ignore non-positive terms, then we get Lemma 9. □

Proof of Lemma 10.

$$E^{t+1} = \frac{1}{m}\sum_{i\in[m]}E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{m}^t\|^2 = \frac{1}{m}\sum_{i\in[m]}E\left\|\boldsymbol{y}_i^{t+1} - \boldsymbol{w}^t - \frac{1}{\alpha}\boldsymbol{g}^t\right\|^2$$

$$
\begin{aligned}
=&\frac{1}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t) - \nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1}) - \boldsymbol{g}^t\|^2 \\
=&\frac{1}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t) - \nabla f_i\circ T_i(\boldsymbol{w}^t) + \nabla f_i\circ T_i(\boldsymbol{w}^t) - \nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1}) - \nabla F(\boldsymbol{m}^t) + \nabla F(\boldsymbol{m}^t) - \boldsymbol{g}^t\|^2 \\
\leq&\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t) - \nabla f_i\circ T_i(\boldsymbol{w}^t)\|^2 + \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1}) - \nabla f_i\circ T_i(\boldsymbol{w}^t)\|^2 \\
&+ \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla F(\boldsymbol{m}^t)\|^2 + \frac{4}{\alpha^2}E\|\nabla F(\boldsymbol{m}^t) - \boldsymbol{g}^t\|^2 \\
\leq&\frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{w}_i^t) - \nabla f_i\circ T_i(\boldsymbol{w}^t)\|^2 + \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{y}_i^{t+1}) - \nabla f_i\circ T_i(\boldsymbol{w}^t)\|^2 \\
&+ \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla F(\boldsymbol{m}^t)\|^2 + \frac{4}{\alpha^2}\frac{1}{m}\sum_{i\in[m]} E\|\nabla f_i\circ T_i(\boldsymbol{m}^t) - \nabla f_i\circ T_i(\boldsymbol{w}_i^t)\|^2 \\
\leq&\frac{8}{\alpha^2}L^2C^t + \frac{4L^2}{\alpha^2}E^{t+1} + \frac{4}{\alpha^2}E\|\nabla F(\boldsymbol{m}^t)\|^2
\end{aligned}
$$

where we use definition of $\boldsymbol{m}^t$, Eq. 11, Jensen Inq., definition of $\boldsymbol{g}^t$, Jensen Inq. and smoothness Inq. 7 respectively. Rearranging the terms give the statement in the Lemma. □

Proof of Lemma 11.

$$
\begin{aligned}
C^{t+1} =&\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{w}_i^{t+1} - \boldsymbol{m}^{t+1}\|^2 = \frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{w}_i^{t+1} - \boldsymbol{m}^t + \boldsymbol{m}^t - \boldsymbol{m}^{t+1}\|^2 \\
\leq&\left(1 + \frac{P}{2m-P}\right)\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{w}_i^{t+1} - \boldsymbol{m}^t\|^2 + \left(1 + \frac{2m-P}{P}\right)\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{m}^t - \boldsymbol{m}^{t+1}\|^2 \\
=&\left(1 + \frac{P}{2m-P}\right)\left(\frac{P}{m}\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{y}_i^{t+1} - \boldsymbol{m}^t\|^2 + \left(1 - \frac{P}{m}\right)\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{w}_i^t - \boldsymbol{m}^t\|^2\right) \\
&+ \left(1 + \frac{2m-P}{P}\right)\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{m}^t - \boldsymbol{m}^{t+1}\|^2 \\
=&\frac{2P}{2m-P}E^{t+1} + \frac{2m-2P}{2m-P}C^t + \frac{2m}{P}\frac{1}{m}\sum_{i\in[m]} E\|\boldsymbol{m}^t - \boldsymbol{m}^{t+1}\|^2 \\
\leq&\left(\frac{2P}{2m-P} + \frac{2m}{P}\right)E^{t+1} + \frac{2m-2P}{2m-P}C^t
\end{aligned}
$$

where we use definition of C^{t+1}, inequality $\|\boldsymbol{w}_1 + \boldsymbol{w}_2\|^2 \leq (1+k)\|\boldsymbol{w}_1\|^2 + \left(1 + \frac{1}{k}\right)\|\boldsymbol{w}_2\|^2$, $\forall z > 0$, tower property by taking expectation with respect to randomness prior to time t, definition of E^t and Lemma 3. □