A Limitations, Future Work, and Broader Impact

Learning on naturally heterogeneous datasets can be challenging, as the true data distributions of individual clients are unknown, making it difficult to correlate the divergence between client data distribution and the global data distribution with routing policy decisions. In our approach, we estimate the distribution divergence by measuring the difference between inference losses on global and local models, which helps us reason about routing probabilities for global and local routes. To further improve our understanding of the model performance, we plan to propose a metric that quantifies the difference in performance when a particular dataset is included versus excluded.

Flow has shown the promise of per-instance personalization in improving clients' accuracy. This approach also holds the potential of preserving privacy by protecting against gradient leakage [38–40] and membership inference [41, 42] attacks that are easier to carry out in vanilla FL. Studying the relationship between personalization and privacy, and comparing our approach to traditional methods like Differential Privacy (DP) [43, 44] can reveal properties of personalization that go beyond improved accuracy.

B Datasets and Hyperparameters

Stackoverflow The Stackoverflow dataset [30] is comprised of separate clients designated for training, validation, and testing. The dataset contains a total of 342,477 train clients, whose combined sample count equals 135,818,730. Similarly, the dataset contains 38,758 validation and 204,088 test clients, whose combined sample counts equal 16,491,230 and 16,586,035, respectively. This dataset is naturally heterogeneous [45] since each user of Stackoverflow represents a client, with their posts forming the dataset for that client. The heterogeneity of the dataset arises from the fact that users have different writing styles, meaning the clients' datasets are not i.i.d., and the total number of posts from each user varies, leading to different dataset sizes per client.

We have trained *Flow* and its baselines on the Stackoverflow dataset for 2000 rounds. The one layer LSTM we have used has 96 as embedding dimension, 670 as hidden state size, and 20 as the maximum sequence length [19]. The batch size used for each client on each baseline is 16. The vocabulary of this language task is limited to 10k most frequent tokens. The default learning rate used is 0.1. The number of clients per round is set to 10, as is the common practice in [14, 46, 13, 10, 47]. For client-side training, the default epoch count is 3 for all the algorithms.

For KNNPER, we used 5 nearby neighbors, and the mixture parameter is $\lambda = 0.5$. For APFL, mixture hyperparameter α is set to 0.25. DITTO has regularization hyperparameter $\lambda = 0.1$. There are 2 clusters by default for HYPCLUSTER. *Flow* and its variants were tested on the following choices of regularizing hyperparameters $\gamma \in \{$1e-1, 1e-2, 1e-3, 1e-4$\}$, where 1e-3 gave the best personalized accuracy.

Shakespeare The Shakespeare dataset [31] consists of 715 distinct clients, each of which has its own training, validation, and test datasets. The combined training datasets of all clients contain a total of 12,854 instances, while the combined validation and test datasets contain 3,214 and 2,356 instances, respectively. The Shakespeare dataset is considered heterogeneous due to the fact that each client is a play written by William Shakespeare, and these plays have varying settings and characters.

All the baselines and *Flow* variants have been run for 1500 rounds, with 10 clients per round. The 2 layer LSTM used [19] has 8 as embedding size, vocabulary size of 90 most frequently used characters, and 256 as hidden size. The default epoch count is 5 for each client, for each algorithm. The batch size is 4 since bigger batch sizes resulted in the divergence of the global model across all the different runs. The default learning rate is 0.1.

Since each client has a sample count under 20, we have used 3 as the nearest neighbor sample count for KNNPER. λ and α, the mixture parameters, for KNNPER and APFL respectively, are set to 0.45 and 0.3. The regularization parameter λ for DITTO is set to 0.1. For *Flow*, the learning rate is set to 0.11 and the regularization parameter is picked from $\gamma \in \{$1e-1, 1e-2, **1e-3**, 1e-4$\}$ similar to Stackoverflow.

EMNIST The EMNIST dataset [32] comprises 3400 distinct clients, each of which has its own training, validation, and test datasets. The combined total number of instances in the train datasets of all clients is 671,585, whereas the validation and test datasets of all clients combined contain 77,483 instances each. The heterogeneity of EMNIST clients is due to the individual writing styles of each client, with each client representing a single person. This is discussed in Appendix C.2 of [19].

The default round count for all the baselines and *Flow* variants is 1500, with 10 clients participating per round. Similar to AFO [19], we have used a shallow convolution neural network with 2 convolution layers. Each client uses 3 local epochs for on-device training. The default batch size is 20, and the default learning rate is 0.01.

For LOCAL only training, we have used 10 epochs per client with a learning rate of 0.05. The nearest sample count for KNNPER is 10 and the mixture parameter is $\lambda = 0.4$. For APFL, we have the default mixture parameter as $\alpha = 0.25$. DITTO has regularization hyperparameter as $\lambda = 0.1$. There are 2 clusters for the clustering algorithm HYPCLUSTER. And for *Flow*, along with its variants, we have picked $\gamma \in \{$1e-1, **1e-2**, 1e-3, 1e-4$\}$ as the regularizing hyperparameter.

CIFAR10 The CIFAR10 dataset is derived from the centralized version of the CIFAR10 dataset [33], which comprises 50,000 images. The federated CIFAR10 dataset consists of 500 unique clients, each of which has 100 training samples and 20 testing samples. The training and testing samples for each client are determined according to the Dirichlet distribution [19]. The heterogeneity of a client is determined by the Dirichlet distribution parameter $\alpha \in [0, 1]$, where a client is more heterogeneous than $\alpha \rightarrow 0$. In this context, heterogeneity refers to the dissimilarity of the dataset instances sampled from a distribution. We conducted experiments on clients with α values of 0.1 and 0.6.

We ran all the experiments for 4000 rounds for the CIFAR10 dataset. ResNet18 [34] is used for all the algorithms. The default batch size is 20 and the default learning rate is 0.05. Each client individually trains their local versions of the global model for 3 epochs.

For LOCAL only training, 20 epochs per client were used. The learning rate was 0.1 for the same. The nearest sample count and the mixture hyperparameter for KNNPER are set to 5 and 0.5. PARTIALFED learning rate is set to 0.11, with the local epoch count is 5. APFL has mixture hyperparameter set as $\alpha = 0.2$. And DITTO has a regularization hyperparameter set as $\lambda = 0.01$. *Flow* and its variants have their regularization hyperparameter as $\gamma \in \{$1e-1, 1e-2, **1e-3**, 1e-4$\}$.

CIFAR100 Like CIFAR10, the CIFAR100 dataset [48] is derived from the CIFAR100 dataset [33] consisting of 50,000 images. The number of clients and the count of training and testing images are identical to those of CIFAR10. Similarly, we also conducted experiments with the Dirichlet parameter set to $\alpha = 0.1$ and $\alpha = 0.6$.

Similar to CIFAR10, we have a 4000 round count for all the algorithms ran on the CIFAR100 dataset. We have again used ResNet18 [34]. The default local epoch count is 3, and the default learning rate is 0.05. We have used 20 batch size for all the algorithms. For each round, 10 clients participate as is the norm stated in the Stackoverflow dataset description.

LOCAL only training has 20 epochs per client, and 0.1 learning rate. 5 nearest samples are used for KNNPER, while the mixture parameter λ is set to 0.4. PARTIALFED, just like in CIFAR10, has 0.11 learning rate and 5 local epochs per client. APFL has 0.25 as mixture parameter α. DITTO has 1e-2 as regularization parameter λ. For both CIFAR10 and CIFAR100, we have 2 as the default cluster count for HYPCLUSTER. *Flow* and its variants get {1e-1, 1e-2, **1e-3**, 1e-4} as the regularization hyperparameter γ.

C Additional Results

C.1 Generalized and Personalized Accuracy

Generalized (Personalized) accuracy is calculated based on the global (personalized) model, where each participating client's test dataset is used to compute accuracy of the global (personalized) model.

Generalized accuracy is formulated as

$$Acc_g = \frac{1}{M} \sum_{m \in [M]} \frac{\sum_{(x,y) \in \mathcal{S}_m^{\text{test}}} \mathbb{1}\{y = w_g(x)\}}{\mathcal{S}_m^{\text{test}}}. \tag{6}$$

Personalized accuracy is formulated as

$$Acc_p = \frac{1}{M} \sum_{m \in [M]} \frac{\sum_{(x,y) \in \mathcal{S}_m^{\text{test}}} \mathbb{1}\{y = w_{p,m}(x)\}}{\mathcal{S}_m^{\text{test}}}. \tag{7}$$

We have reported Generalized (Personalized) Accuracy Acc_g (Acc_p) of *Flow*, averaged across 1000 clients in Table 4, for all the datasets. Similarly, variance of accuracies across 3 different runs (based on seeds 0, 44, 56) is reported in Table 5.

Flow sees an improvement of 1.11-3.46% in Acc_g and 1.33-4.58% in Acc_p over the best performing baseline. Besides the main observations listed in Section 5, we discuss results on the CIFAR100 dataset here. For CIFAR100 (0.6), *Flow* ($40.08\% \pm 0.27\%$) matches the personalized accuracy of the highest performing baseline, PARTIALFED ($40.18\% \pm 0.19\%$), while achieving 1.98% point increase in generalized accuracy. And for CIFAR100 (0.1), *Flow* improves personalized accuracy by 1.78% points. For generalized accuracy, *Flow* ($34.00\% \pm 0.32\%$) reaches close to the best performing baseline, PARTIALFED ($34.79\% \pm 0.29\%$). The reason behind the on-par performance of *Flow* with PARTIALFED can be attributed to the statefulness of PARTIALFED. With the assumption of full device participation, PARTIALFED makes use of each client's previous state of the personalized model to further train its layer-wise model building policy. With *Flow*, both the assumptions of full device participation and statefulness of the personalized model are not necessary. Since the clients do

Table 4: Generalized (Acc_g) and Personalized (Acc_p) accuracy (the higher, the better) for *Flow* and baselines. Variance across different runs is reported in Appendix C, Table 5.

Datasets	Stackoverflow		Shakespeare		EMNIST		CIFAR10 (0.1)		CIFAR100 (0.1)		CIFAR10 (0.6)		CIFAR100 (0.6)	
Baselines	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p
LOCAL	-	15.93%	-	18.70%	-	28.18%	-	49.78%	-	36.19%	-	62.74%	-	21.31%
FEDAVG	23.15%	-	52.00%	-	85.10%	-	60.98%	-	28.11%	-	67.50%	-	30.33%	-
FEDAVGFT	23.83%	24.41%	52.12%	53.68%	89.57%	90.14%	61.23%	73.03%	29.60%	31.02%	68.19%	72.21%	31.15%	37.24%
KNNPER	23.16%	24.49%	51.87%	53.10%	85.20%	88.28%	59.62%	75.14%	28.08%	33.62%	69.22%	70.14%	30.66%	34.39%
PARTIALFED	-	-	-	-	-	-	62.57%	73.20%	**34.79%**	40.64%	66.93%	70.38%	37.72%	**40.18%**
APFL	22.96%	25.70%	52.38%	53.64%	88.40%	89.44%	62.87%	72.86%	31.05%	32.56%	69.53%	72.53%	36.37%	36.74%
DITTO	22.59%	24.36%	52.44%	53.95%	89.08%	91.30%	62.06%	72.06%	28.14%	35.45%	68.12%	70.31%	35.11%	36.07%
FEDREP	18.92%	21.04%	46.71%	50.09%	89.95%	89.77%	64.85%	68.62%	26.10%	33.72%	69.77%	63.61%	28.42%	31.02%
LGFEDAVG	22.61%	24.03%	51.08%	51.43%	87.43%	91.70%	56.63%	73.19%	31.65%	39.63%	67.48%	68.94%	35.01%	33.90%
HYPCLUSTER	23.75%	22.43%	51.92%	52.74%	89.47%	90.49%	63.64%	71.55%	31.57%	33.04%	65.44%	72.40%	34.76%	36.22%
Flow (Ours)	**26.64%**	**29.49%**	**55.90%**	**56.20%**	**90.88%**	**94.18%**	**66.26%**	**76.47%**	34.00%	**42.42%**	**70.88%**	**77.11%**	**39.70%**	40.08%

Table 5: Variance of generalized and personalized accuracies across 3 different runs (seeds = 0, 44, 56) for *Flow* and its baselines.

Datasets	SO NWP		Shakespeare		EMNIST		CIFAR10 (0.1)		CIFAR100 (0.1)		CIFAR10 (0.6)		CIFAR100 (0.6)	
Baselines	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p	Acc_g	Acc_p
LOCAL	-	0.25%	-	0.46%	-	1.14%	-	1.56%	-	0.43%	-	0.89%	-	0.25%
FEDAVG	0.07%	-	0.39%	-	1,32%	-	1.12%	-	0.31%	-	0.82%	-	0.15%	-
FEDAVGFT	0.09%	0.26%	0.51%	0.59%	1.16%	1.21%	0.99%	0.89%	0.46%	0.62%	1.10%	1.26%	0.33%	0.42%
KNNPER	0.16%	0.24%	0.36%	0.41%	0.95%	1.02%	1.41%	1,57%	0.34%	0.57%	0.91%	1.06%	0.24%	0.29%
PARTIALFED	-	-	-	-	-	-	1.36%	1.39%	0.29%	0.46%	0.96%	1.97%	0.09%	0.19%
APFL	0.19%	0.20%	0.41%	0.53%	1.41%	1.50%	1.24%	1.31%	0.36%	0.72%	0.70%	0.97%	0.42%	0.59%
DITTO	0.12%	0.15%	0.49%	0.56%	1.12%	1.22%	1.35%	1.41%	0.43%	0.69%	0.84%	0.87%	0.28%	0.34%
FEDREP	0.15%	0.29%	0.50%	0.65%	0.89%	0.94%	0.95%	1.02%	0.59%	0.79%	0.96%	1.14%	0.14%	0.10%
LGFEDAVG	0.08%	0.16%	0.32%	0.56%	1.10%	1.17%	1.21%	1.24%	0.47%	0.51%	0.82%	0.96%	0.23%	0.21%
HYPCLUSTER	0.20%	0.19%	0.56%	0.73%	0.90%	1.13%	1.43%	1.49%	0.39%	0.47%	0.98%	0.76%	0.35%	0.46%
Flow	0.23%	0.28%	0.40%	0.49%	1.16%	1.21%	1.23%	1.25%	0.32%	0.36%	0.78%	0.86%	0.21%	0.27%

not necessarily have to carry their personalized model states to the upcoming rounds, the personalized model recreated by *Flow* might be unable to compete against stateful approaches like PARTIALFED. Although because of the per-instance routing, *Flow* still manages to outperform PARTIALFED for the CIFAR10 (0.1/0.6) datasets, and gives comparable performance for the CIFAR100 (0.1/0.6) datasets.

C.2 Percentage of Clients Benefiting from Personalization

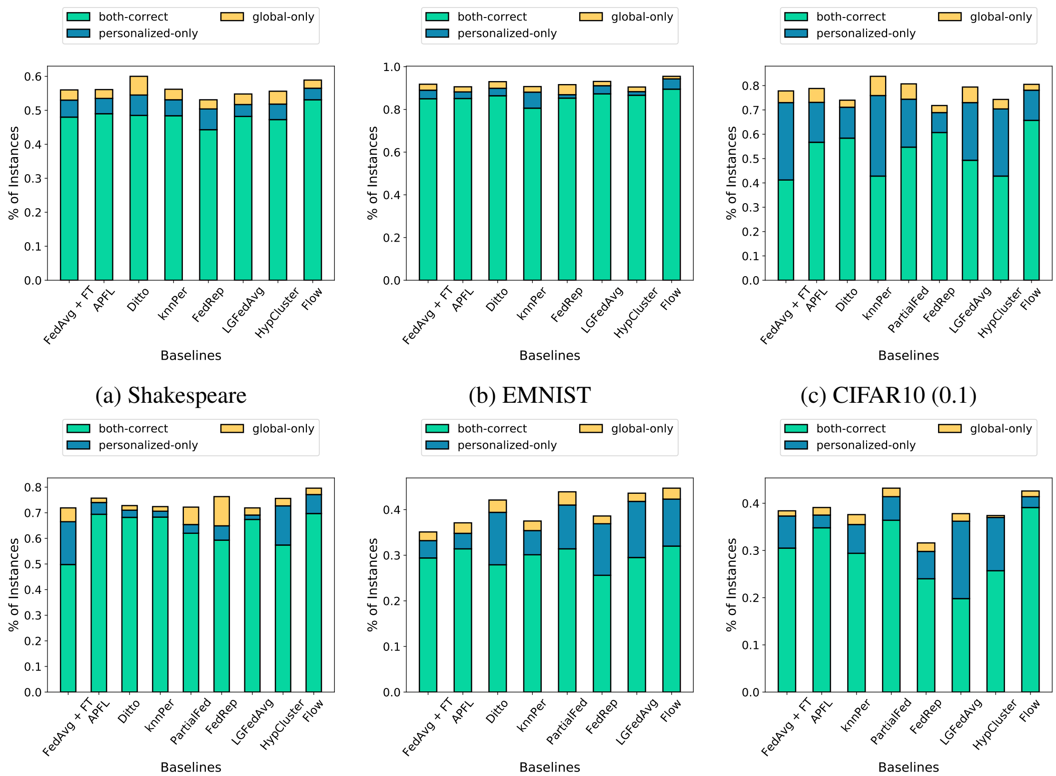
In this section we discuss the effect of personalization, by comparing each client's performance on their individual personalized models with their performance on the global model. The evaluation, just as in section C.1, is done on the test datasets of all the clients. The goal with any personalization method is to make each client's personalized model more beneficial (for us, in terms of accuracy) compared to the global model. Hence we want $Acc_p > Acc_g$, to incentivize personalization for each client. As shown in Table 6, compared to the best performing baseline, *Flow* improves the utility of personalization by up to 3.31% points.

Table 6: % of clients for which $Acc_p > Acc_g$ (the higher, the better).

	Stackoverflow	EMNIST	Shakespeare	CIFAR10 (0.1)	CIFAR100 (0.1)	CIFAR10 (0.6)	CIFAR100 (0.6)
FEDAVGFT	79.26%	81.48%	79.00%	97.18%	91.74%	99.33%	88.54%
KNNPER	82.73%	89.97%	68.87%	90.00%	94.71%	90.00%	96.37%
PARTIALFED	-	-	-	88.30%	90.32%	84.80%	98.64%
APFL	69.66%	93.39%	79.22%	87.48%	86.18%	90.63%	92.03%
DITTO	74.59%	79.26%	73.74%	90.52%	91.45%	89.61%	97.45%
FEDREP	91.53%	82.20%	79.78%	92.30%	78.81%	84.64%	99.54%
LGFEDAVG	83.47%	66.16%	88.43%	88.41%	86.39%	89.59%	91.73%
HYPCLUSTER	80.46%	80.70%	74.84%	95.11%	93.70%	98.18%	99.73%
Flow (Ours)	**92.74%**	**96.70%**	**89.77%**	**98.33%**	**97.29%**	**99.62%**	**99.75%**

C.3 Breakdown of Correctly Classified Instances

Here we show a detailed view of how instances (across all the clients) get classified correctly between global and personalized models for each of the baselines. For the plots in Figures 5, y-axis represent % of instances correctly classified by (a) Both the global and the personalized models (**both-correct**), (b) Only the global model (**global-only**), and (c) Only the personalized model (**personalized-only**). This % of instances metric is averaged across all clients, and is based on their test datasets. The goal here is to increase the % of instances for **both-correct** and **personalized-only**, and reduce the % of instances for **global-only**. We make the following



(a) Shakespeare (b) EMNIST (c) CIFAR10 (0.1)

(d) CIFAR 10 (0.6) (e) CIFAR100 (0.1) (f) CIFAR100 (0.6)

Figure 5: Different combinations of w_g and w_p accuracies.

observations for each of the datasets: Since *Flow* improves both the generalized and personalized accuracies, we see higher **both-correct** for Stackoverflow (by 2.75% points), Shakespeare (by 4.34% points), EMNIST (by 3.17% points), CIFAR10 (0.1) (by 5.24% points), CIFAR10 (0.6) (by 0.03% points), CIFAR100 (0.1) (by 0.63% points) and CIFAR100 (0.6) (by 2.78% points).

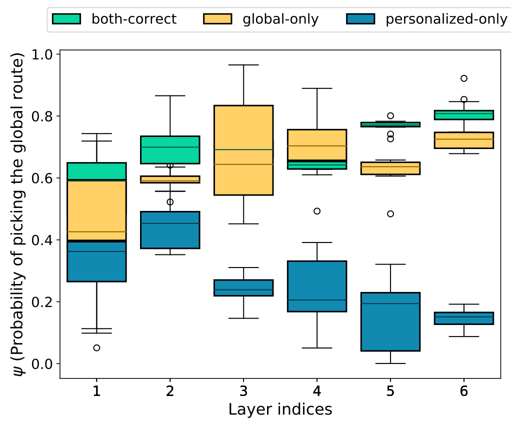
Due per-instance personalization, we see improvements in personalized accuracy, but those improvements are also included in the **both-correct** bars, so solely comparing **personalized-only** bar lengths is not a right comparison. Similarly, we see fewer instances in **global-only** bars due to the increase in instances which fall under **both-correct**.

C.4 Analysis of Routing Decisions

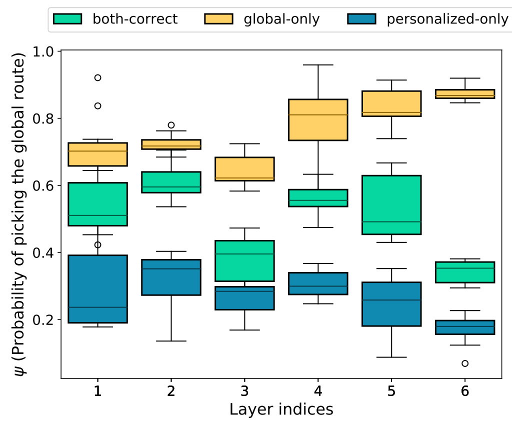
Now we show probability value analysis of the routing policy for CIFAR10/100 datasets. Here we have fixed the client as the client which had the highest loss difference between its global and personalized models for *Flow*. This analysis was done during the inference stage, on the test dataset of the above-mentioned client. The box plots show statistics on the probability of picking the global route for all the instances. Echoing the observations made in Section 5, in Figure 6, we see a trend in increasing probability for the global parameters for the instances which are correctly classified by only the global model. In the contrary, for the instances which can only be classified by the personalized model, the probability for taking the global route is lower as the input passes through more layers.

C.5 Ablation Study: Regularization

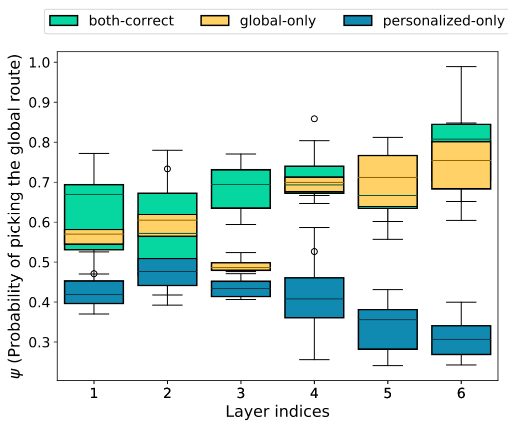
Figures 7 and 8 show the validation curves for generalized and personalized accuracy with and without the regularization term used in the policy learning objective as shown in Equation 4. With regularization, we see an improvement of 2.18% (Stackoverflow), 1.86% (Shakespeare), 3.98% (EMNIST), 2.55% (CIFAR10 0.1), 4.36% (CIFAR10 0.6), 0.91% (CIFAR100 0.1), 3.46% (CIFAR100 0.6) for the generalized accuracy. And for



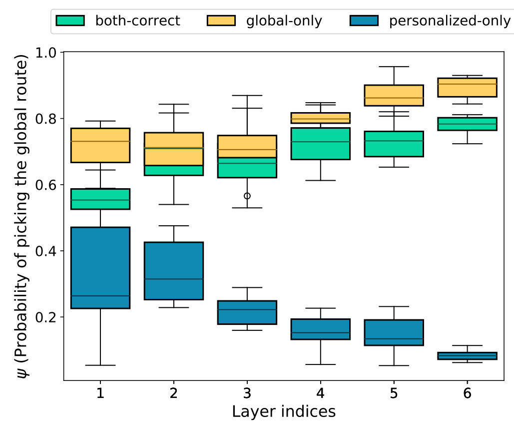
(a) CIFAR10 (0.1)



(b) CIFAR 10 (0.6)



(c) CIFAR100 (0.1)



(d) CIFAR100 (0.6)

Figure 6: Behavior of ψ_g for all instances with respect to each layer of a client with highest loss difference between personalized and global models.

the personalized accuracy, we see an improvement of 1.92% (Stackoverflow), 2.02% (Shakespeare), 3.01% (EMNIST), 0.65% (CIFAR10 0.1), 3.98% (CIFAR10 0.6), 2.42% (CIFAR100 0.1), 2.19% (CIFAR100 0.6).

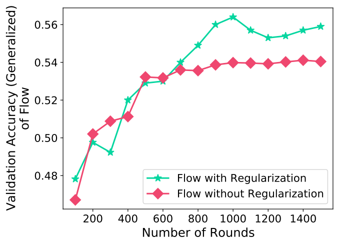
C.6 Ablation Study: Per-instance Personalization

Figures 9 show the validation curves for 3 *Flow* variants: (a) Per-instance Per-client *Flow*, which is the primary design proposed in this work, (b) Per-instance *Flow*, which makes choices between two global routes solely based on each client's instances, (c) Per-client *Flow*, which is simply FEDAVGFT where the personalization only depends on a client, and not on any specific instances.

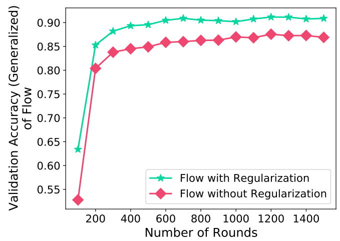
With all the datasets, we see a trend of Per-instance *Flow* outperforming Per-client *Flow* by 1.88% (Stackoverflow), 0.82% (Shakespeare), 5.07% (EMNIST), 2.90% (CIFAR10 0.1), 2.41% (CIFAR10 0.6), 7.52% (CIFAR100 0.1), 1.09% (CIFAR100 0.6). We also see the trend of Per-instance *Flow* outperforming Per-Instance Per-Client *Flow* by 3.19% (Stackoverflow), 1.24% (Shakespeare), 0.94% (EMNIST), 0.55% (CIFAR10 0.1), 4.49% (CIFAR10 0.6), 3.88% (CIFAR100 0.1), 1.37% (CIFAR100 0.6).

C.7 Ablation Study: Soft versus Hard Policy

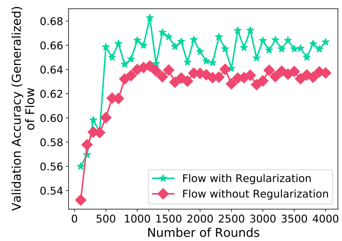
Table 7 shows the personalized accuracy of the test clients while using soft and hard policies during inference. We see that the accuracy difference between the two designs are statistically insignificant. Hence, using a hard policy for inference not only saves half the compute resources, but also doesn't affect the personalized model's performance.



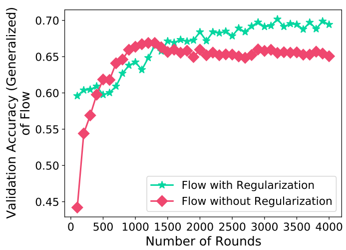
(a) Shakespeare



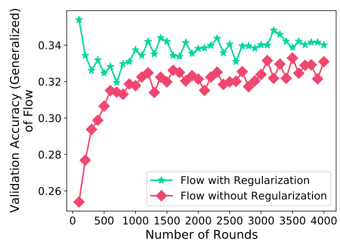
(b) EMNIST



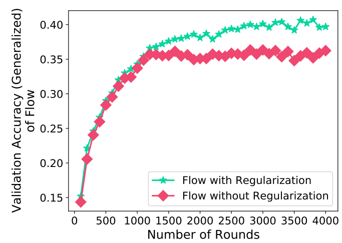
(c) CIFAR10 (0.1)



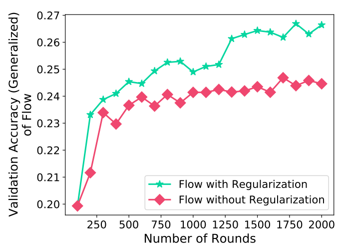
(d) CIFAR 10 (0.6)



(e) CIFAR100 (0.1)



(f) CIFAR100 (0.6)

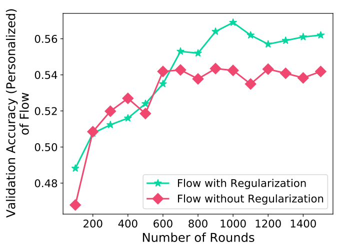


(g) Stackoverflow

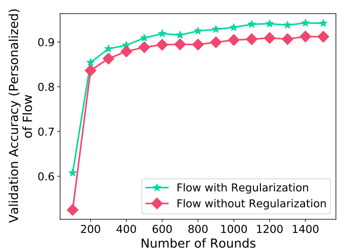
Figure 7: Generalized Accuracy of the Ablation Study on the Regularization Term used in the Policy Learning Objective.

Table 7: Test (personalized) accuracy of two of the *Flow* variants: (a) Soft Policy variant where the probability $\mathbf{q}$ is continuous in the range of $[0, 1]$ during inference. (b) Hard Policy variant where the probability $\mathbf{q}$ is discrete over the set $\{0,1\}$ during inference.

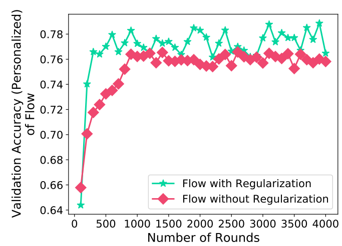
Datasets	Stackoverflow	Shakespeare	EMNIST	CIFAR 10 (0.1)	CIFAR 100 (0.1)	CIFAR 10 (0.6)	CIFAR 100 (0.6)
Soft Policy	29.57% $\pm$ 0.22%	57.01% $\pm$ 0.53%	94.97% $\pm$ 1.06%	77.24% $\pm$ 1.30%	42.75% $\pm$ 0.30%	77.02% $\pm$ 0.90%	39.74% $\pm$ 0.13%
Hard Policy	29.49% $\pm$ 0.28%	56.20% $\pm$ 0.49%	94.18% $\pm$ 1.21%	76.47% $\pm$ 1.25%	42.42% $\pm$ 0.36%	77.11% $\pm$ 0.86%	40.08% $\pm$ 0.27%



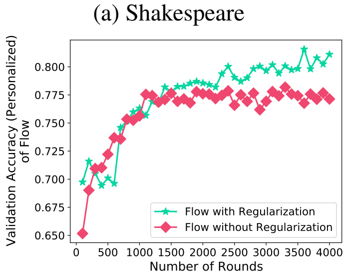
(a) Shakespeare



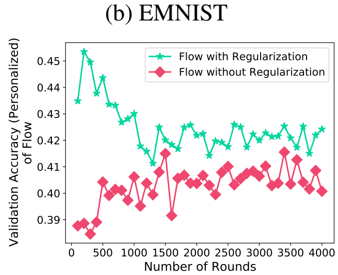
(b) EMNIST



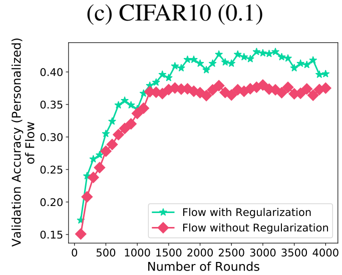
(c) CIFAR10 (0.1)



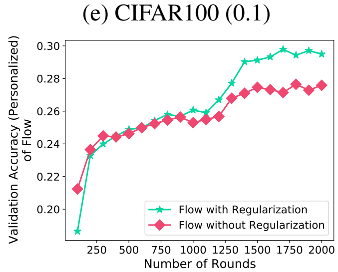
(d) CIFAR 10 (0.6)



(e) CIFAR100 (0.1)

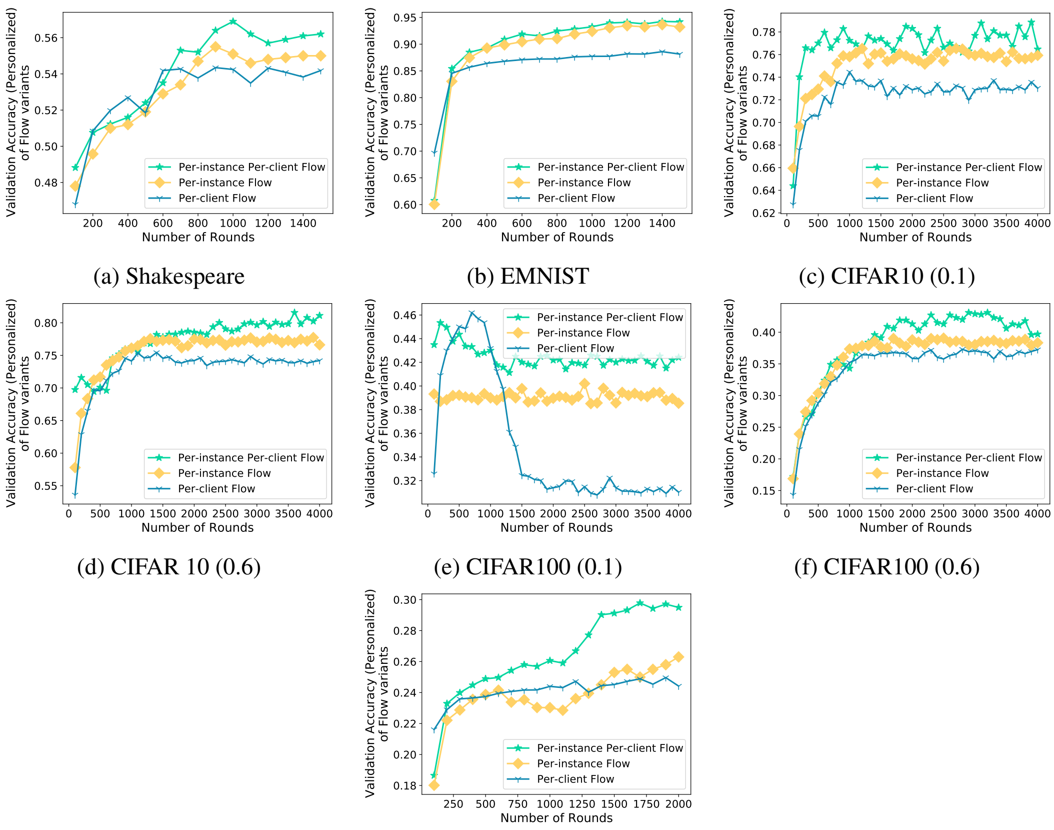


(f) CIFAR100 (0.6)



(g) Stackoverflow

Figure 8: Personalized Accuracy of the Ablation Study on the Regularization Term used in the Policy Learning Objective.



(a) Shakespeare (b) EMNIST (c) CIFAR10 (0.1)

(d) CIFAR 10 (0.6) (e) CIFAR100 (0.1) (f) CIFAR100 (0.6)

(g) Stackoverflow

Figure 9: Ablation of the dynamic routing component (Per-client *Flow*), and the local component (Per-instance *Flow*).

D Proofs

D.1 *Flow*: Detailed

Here we give a detailed version of *Flow* (Algorithm 2) for proving its convergence properties. Here we are assuming that the global and local model output interpolation is model-wise (after the final layer), not layer-wise.

Algorithm 2: *Flow*

Input: R: Total number of rounds, $r \in [R]$: Round index, M: Total number of clients, $m \in [M]$: Client index, $\mathcal{M}$: Set of available clients , p: Client sampling rate, K: Total local epoch count, $k \in [K]$: Epoch index, η_ℓ: Local learning rate, $w_g^{(r)}$: Global model at r^{th} round, $w_{g,m}^{(r,k)}$: m^{th} client's local update of the global model for r^{th} round and k^{th} epoch, $w_{\ell,m}^{(r,k)}$: m^{th} client's local model for r^{th} round and k^{th} epoch, $w_{p,m}^{(r,k)}$: m^{th} client's personalized model for r^{th} round and k^{th} epoch, $\psi_g^{(r)}$: Global policy model at r^{th} round, $\psi_{g,m}^{(r,k)}$: m^{th} client's routing policy for r^{th} round and k^{th} epoch, $\mathcal{D}_m$: Data distribution of m^{th} client, $\mathcal{S}_m$: Dataset of m^{th} client, $\zeta_{m,\ell}$: Dataset used to train w_ℓ, $\zeta_{m,g}$: Dataset used to train w_g and ψ_g

Output: $w_g^{(R+1)}$: Global model at the end of the training

Server randomly initializes $w_g^{(1)}$
for $r \in [R]$ *round* **do**
  Sample M clients from $\mathcal{M}$ with the rate of p
  Send $w_g^{(r)}, \psi_g^{(r)}$ to all the clients
  for $m \in [M]$ *in parallel* **do**
    $w_{g,m}^{(r,0)} \leftarrow w_g^{(r)}; \psi_{g,m}^{(r,0)} \leftarrow \psi_g^{(r)}; w_{\ell,m}^{(r,0)} \leftarrow w_{g,m}^{(r,0)}$
    $\zeta_{m,\ell}, \zeta_{m,g} \leftarrow \mathcal{S}_m$ `/* Creating two mutually exclusive datasets */`
    for $k \in [K_1]$ *epochs* **do**
      $w_{\ell,m}^{(r,k)} \leftarrow w_{\ell,m}^{(r,k-1)} - \eta_\ell \nabla f_m(w_{\ell,m}^{(r,k-1)}; \zeta_{m,\ell})$
    end
    for $k \in [K_2]$ *epochs* **do**
      $\forall\ (x_m, y_m) \sim \zeta_{m,g}$, define
      $\tilde{w}_{p,m}^{(r,k-1)}(x_m) \leftarrow \psi_{g,m}^{(r,k-1)}(x_m) \cdot w_{g,m}^{(r,k-1)}(x_m) + (1 - \psi_{g,m}^{(r,k-1)}(x_m)) \cdot w_{\ell,m}^{(r,K)}(x_m)$
      $\psi_{g,m}^{(r,k)} \leftarrow \psi_{g,m}^{(r,k-1)} - \eta_\ell \nabla_{\psi_{g,m}^{(r,k-1)}} \left[f_m(\tilde{w}_{p,m}^{(r,k-1)}; \zeta_{m,g}) \right]$
      $\forall\ (x_m, y_m) \sim \zeta_{m,g}$, define
      $w_{p,m}^{(r,k-1)}(x_m) \leftarrow \psi_{g,m}^{(r,k)}(x_m) \cdot w_{g,m}^{(r,k-1)}(x_m) + (1 - \psi_{g,m}^{(r,k)}(x_m)) \cdot w_{\ell,m}^{(r,K)}(x_m)$
      $w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} - \eta_\ell \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}; \zeta_{m,g})$
    end
    Send back $w_{g,m}^{(r,K)}, \psi_{g,m}^{(r,K)}, n_m := |\zeta_{m,g}|$
  end
  $w_g^{(r+1)} \leftarrow \frac{1}{nM} \sum_{m \in [M]} n_m w_{g,m}^{(r,K)}$
  $\psi_g^{(r+1)} \leftarrow \frac{1}{nM} \sum_{m \in [M]} n_m \psi_{g,m}^{(r,K)}$
end

D.2 Basics

We perform theoretic analysis of *Flow* based on the following setup: There are total M clients. A client is denoted by a unique integer m associated with it where $m \in [M]$. Each client m has a dataset $\mathcal{S}_m = \{(x_m^{(i)}, y_m^{(i)}); i \in [n_m]\}$ where $(x_m^{(i)}, y_m^{(i)})$ has been sampled from $\mathcal{D}_m$ distribution of the m^{th} client. $n_m = |\mathcal{S}_m|$ is the total sample count of the m^{th} client. Total sample count across all the participating client is $n = \sum_{m \in [M]} n_m$. The ratio of m^{th} client's sample count to total sample count is $\alpha = \frac{n_m}{n}$.

The global distribution is defined as $\mathcal{D} = \sum_{m \in [M]} q_m \mathcal{D}_m$ where q_m is the weight associated with m^{th} client and $\sum_{m \in [M]} q_m = 1$.

Note that $w_{p,m}$ is a combination of outputs of $w_{g,m}$ (Global parameters) and $w_{\ell,m}$ (Local parameters) on each layer. For tractability of analysis, we will assume that the combination is only after the last layer. Hence,

$$w_{p,m}(x_m) \leftarrow \psi_{g,m}(x_m) w_{g,m}(x_m) + (1 - \psi_{g,m}(x_m)) w_{\ell,m}(x_m).$$

The local model update rule is,

$$w_{\ell,m}^{(r,k)} \leftarrow w_{\ell,m}^{(r,k-1)} - \eta_\ell \nabla f_m(w_{\ell,m}^{(r,k-1)}(x_m), y_m)$$

677 where $w_{\ell,m}^{(r,0)} = w_{g,m}^{(r,0)} = w_g^{(r)}$. Indices $r \in [R]$ and $k \in [K]$ are the global round and the local epoch indices.
The policy update rule is,

$$\psi_{g,m}^{(r,k)} \leftarrow \psi_{g,m}^{(r,k-1)} - \eta_\ell \nabla_{\psi_g} f_m(w_{p,m}^{(r,k-1)}(x_m), y_m).$$

The global model update rule is,

$$w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} - \eta_\ell \nabla_{w_g} f_m(w_{p,m}^{(r,k-1)}(x_m), y_m).$$

678 We list out all the optimization problems relevant to *Flow*:

- **Local true risk of the personalized model**

$$F_m(w_{p,m}) := \mathbb{E}_{(x_m, y_m \sim \mathcal{D}_m)}[f_m(w_{p,m}(x_m), y_m)]$$

679 where f_m is a loss function associated with the m^{th} client.

- **Local empirical risk of the personalized model**

$$\hat{F}_m(w_{p,m}) := \frac{1}{n_m} \sum_{i \in [n_m]} f_m(w_{p,m}(x_m^{(i)}), y_m^{(i)})$$

- **Local true risk of the global model**

$$F_m(w_{g,m}) := \mathbb{E}_{(x_m, y_m \sim \mathcal{D}_m)}[f_m(w_{g,m}(x_m), y_m)]$$

- **Local empirical risk of the global model**

$$\hat{F}_m(w_{g,m}) := \frac{1}{n_m} \sum_{i \in [n_m]} f_m(w_{g,m}(x_m^{(i)}), y_m^{(i)})$$

- **Local minimizer of local empirical risk of the personalized model**

$$w_{p,m}^* \in \mathcal{H} \text{ such that } \hat{F}_m(w_{p,m}) \geq \hat{F}_m(w_{p,m}^*);\ \forall w_{p,m} \in \mathcal{H},\ \exists \epsilon > 0,\ ||w_{p,m} - w_{p,m}^*|| < \epsilon$$

- **Global true risk of the global model**

$$F(w_g) = \frac{1}{nM} \sum_{m \in [M]} n_m \mathbb{E}_{(x_m, y_m) \sim \mathcal{S}_m}[f_m(w_{g,m}(x_m), y_m)] \text{ where } n = |\mathcal{S}| = |\bigcup_{m \in [M]} \mathcal{S}_m|$$

- **Global empirical risk of the global model**

$$\hat{F}(w_g) = \frac{1}{nM} \sum_{m \in [M]} n_m \hat{F}_m(w_{g,m}(x_m), y_m) = \frac{1}{nM} \sum_{m \in [M]} \sum_{i \in [n_m]} f_m(w_{g,m}(x_m^{(i)}), y_m^{(i)})$$

- **Local minimizer of global empirical risk**

$$w_g^* \in \mathcal{H} \text{ such that } \hat{F}(w_g) \geq \hat{F}(w_g^*);\ \forall w_g \in \mathcal{H},\ \exists \epsilon > 0,\ ||w_g - w_g^*|| < \epsilon$$

680 We also use the following assumptions similar to [19, 12, 6]:
Assumption D.1 (Strong Convexity)**.** f_m is μ-convex for $\mu \geq 0$. Hence,

$$\langle \nabla f_m(w), v - w \rangle \leq f_m(v) - f_m(w) - \frac{\mu}{2}||w - v||^2,\ \forall m \in [M] \text{ and } w, v \in \mathcal{H}.$$

681 We also generalize our convergence analysis for $\mu = 0$, general convex cases.
Assumption D.2 (Smoothness)**.** The gradient of f_m is β-Lipschitz,

$$||\nabla f_m(w) - \nabla f_m(v)|| \leq \beta ||w - v||,\ \forall m \in [M] \text{ and } w, v \in \mathcal{H}.$$

Assumption D.3 (Bounded Local Variance)**.** $h_m(w) := \nabla f_m(w(x_m), y_m)$ is an unbiased stochastic gradient of f_m with variance bounded by σ_ℓ^2.

$$\mathbb{E}_{(x_m, y_m \sim \mathcal{D}_m)} ||h_m(w) - \nabla f_m(w(x_m), y_m)||^2 \le \sigma_\ell^2, \ \forall m \in [M] \text{ and } w \in \mathcal{H}.$$

Assumption D.4 ((G, B)-Bounded Gradient Dissimilarity)**.** There exists constants $G \ge 0$ and $B \ge 1$ such that

$$\frac{1}{M} \sum_{m \in [M]} ||\nabla f_m(w)||^2 \le G^2 + 2\beta B^2 (F(w) - F(w^*))$$

for a convex f_m. And for a non-convex f_m,

$$\frac{1}{M} \sum_{m \in [M]} ||\nabla f_m(w)||^2 \le G^2 + B^2 ||\nabla F(w)||^2.$$

The derivation is given in Section D.1 of Scaffold [12].

We also use a definition to quantify the diversity of a client's gradient with respect to the global gradient as defined in [29]:

Definition D.5 (Gradient Diversity)**.** The difference between gradients of the m^{th} client's true risk and the global true risk based on the global model w is,

$$\delta_m = \sup_{w \in \mathcal{H}} ||\nabla f_m(w) - \nabla F(w)||^2$$

D.3 Convergence Proof for the Global Model: Convex (Strong and General) Cases

A client's local update for one local epoch on the global model, starting with $w_{g,m}^{(r,0)} \leftarrow w_g^{(r)}$, is

$$w_{g,m}^{(r,k+1)} = w_{g,m}^{(r,k)} - \eta_\ell h_m(w_{p,m}^{(r,k)}). \tag{8}$$

And a client's local update for K epochs on the global model, would be

$$w_{g,m}^{(r,K)} = w_{g,m}^{(r,0)} - \eta_\ell \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) \tag{9}$$

$$= w_{g,m}^{(r,0)} - \eta_\ell \sum_{k=1}^{K} h_m(\psi_{g,m}^{(r,k)}(x_m) w_{g,m}^{(r,k-1)}(x_m) + (1 - \psi_{g,m}^{(r,k)}(x_m)) w_{\ell,m}^{(r,K)}(x_m), y_m). \tag{10}$$

In both the above cases, the gradient is with respect to w_g parameters.

The global model update is,

$$w_g^{(r+1)} = \frac{1}{nM} \sum_{m \in [M]} n_m w_{g,m}^{(r,K)} \tag{11}$$

We first start with a lemma which binds the deviation between the local model $w_{\ell,m}^{(r,K)}$ and the global model starting point $w_g^{(r)}$ for it at round r.

Lemma D.6 (Local model progress)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.2, D.3, and condition $\eta_\ell \le \frac{1}{\beta\sqrt{2K(K-1)}}$ in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{\ell,m}^{(r,K)} - w_{\ell,m}^{(r,0)}||^2 \le 6K^2 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K\eta_\ell^2 \sigma_\ell^2$$

Proof.

$$\mathbb{E}||w_{\ell,m}^{(r,K)} - w_{\ell,m}^{(r,0)}||^2 = \mathbb{E}||w_{\ell,m}^{(r,K-1)} - \eta_\ell \nabla f_m(w_{\ell,m}^{(r,K-1)}) - w_{\ell,m}^{(r,0)}||^2 \tag{12}$$

$$\le \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 + K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,K-1)})||^2 + \eta_\ell^2 \sigma_\ell^2 \tag{13}$$

Here we have used triangle inequality and variance separation.

$$\leq \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 + K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,K-1)})||^2 + \eta_\ell^2\sigma_\ell^2 \tag{14}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 + \eta_\ell^2\sigma_\ell^2 \\ &\quad + K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,K-1)}) - \nabla f_m(w_{\ell,m}^{(r,0)}) + \nabla f_m(w_{\ell,m}^{(r,0)})||^2 \end{aligned} \tag{15}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 + 2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,K-1)}) - \nabla f_m(w_{\ell,m}^{(r,0)})||^2 \\ &\quad + 2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2 \end{aligned} \tag{16}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 + 2K\beta^2\eta_\ell^2 \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 \\ &\quad + 2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2 \end{aligned} \tag{17}$$

Assuming $\eta_\ell \leq \frac{1}{\beta\sqrt{2K(K-1)}}$, we get

$$\begin{aligned} \mathbb{E}||w_{\ell,m}^{(r,K)} - w_{\ell,m}^{(r,0)}||^2 &\leq \left(1 + \frac{2}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K-1)} - w_{\ell,m}^{(r,0)}||^2 \\ &\quad + 2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2 \end{aligned} \tag{18}$$

Unrolling the above recursion,

$$\mathbb{E}||w_{\ell,m}^{(r,K)} - w_{\ell,m}^{(r,0)}||^2 \leq \sum_{i=1}^{K} \left(2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2\right) \left(1 + \frac{2}{K-1}\right)^i \tag{19}$$

$$\leq 3K \left(2K\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2\right) \tag{20}$$

$$= 6K^2\eta_\ell^2 \mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + 3K\eta_\ell^2\sigma_\ell^2 \tag{21}$$

□

Now we move forward to a lemma which binds the deviation between the local version of the global model $w_{g,m}^{(r,k)}$ and the global model starting point $w_g^{(r)}$ for it round r.

Lemma D.7 (Local version of the global model progress)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.1, D.2, D.3, and condition $\eta_\ell \leq \frac{1}{\beta\sqrt{2k}}$ in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 \leq 8k^3\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 4k\eta_\ell^2\sigma_\ell^2$$

Proof. We start by expanding $w_{g,m}^{(r,k)}$ in terms of its previous epoch iterate.

$$\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 = \mathbb{E}||w_{g,m}^{(r,k-1)} - \eta_\ell \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) - w_{g,m}^{(r,0)}||^2 \tag{22}$$

Using triangle inequality and separation of variance, we get,

$$\leq \left(1 + \frac{1}{k-1}\right) \mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + k\eta_\ell^2 \mathbb{E}||\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})||^2 + \eta_\ell^2\sigma_\ell^2 \tag{23}$$

Using the convex property of f_m, we get

$$\leq \left(1 + \frac{1}{k-1}\right) \mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)} \nabla f_m(w_{g,m}^{(r,k-1)})||^2 + \eta_\ell^2\sigma_\ell^2 \tag{24}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{k-1}\right) \mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + \eta_\ell^2\sigma_\ell^2 \\ &\quad + k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)} (\nabla f_m(w_{g,m}^{(r,k-1)}) - \nabla f_m(w_{g,m}^{(r,0)}) + \nabla f_m(w_{g,m}^{(r,0)}))||^2 \end{aligned} \tag{25}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{k-1}\right) \mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + \eta_\ell^2\sigma_\ell^2 + 2k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_{g,m}^{(r,0)}))||^2 \\ &\quad + 2k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||(\nabla f_m(w_{g,m}^{(r,k-1)}) - \nabla f_m(w_{g,m}^{(r,0)})||^2 \end{aligned} \tag{26}$$

$$\begin{aligned} &\leq \left(1 + \frac{1}{k-1} + 2k\eta_\ell^2\beta^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2\right) \mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + \eta_\ell^2\sigma_\ell^2 \\ &\quad + 2k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_{g,m}^{(r,0)}))||^2 \end{aligned} \tag{27}$$

Unrolling the recursion,

$$\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 \leq \sum_{i=1}^{k} \left(2k\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_{g,m}^{(r,0)})||^2 + \eta_\ell^2 \sigma_\ell^2\right) \left(1 + \frac{1}{k-1} + 2k\eta_\ell^2 \beta^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2\right)^i \tag{28}$$

Assuming that $\eta_\ell \leq \frac{1}{\beta\sqrt{2k}}$ we get $k\eta_\ell^2\beta^2 \leq 1$,

$$\begin{aligned}
\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 &\leq \left(2k\eta_\ell^2 \sum_{i=1}^{k} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \mathbb{E}||\nabla f_m(w_{g,m}^{(r,0)})||^2 + \eta_\ell^2\sigma_\ell^2\right) \sum_{i=1}^{k} \left(1 + \frac{1}{k-1} + 2\right)^i && (29)\\
&\leq 4k\left(2k^2\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + \eta_\ell^2\sigma_\ell^2\right) && (30)\\
&\leq 8k^3\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 4k\eta_\ell^2\sigma_\ell^2 && (31)
\end{aligned}$$

□

Lemma D.8 (Deviation of the personalized model from the global model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.1, D.2, D.3, and condition $\eta_\ell \leq \min\left(\frac{1}{\beta\sqrt{2K(K-1)}}, \frac{1}{\beta\sqrt{2K}}\right)$ in Algorithm 2, the following is satisfied:*

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 \leq\ & 16k^3\eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 8k\eta_\ell^2\sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2\\
&+ 12K^2\eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 6K\eta_\ell^2\sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 &= \mathbb{E}||\psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} + (1 - \psi_{g,m}^{(r,k)}) w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,0)}||^2 && (32)\\
&= \mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)} - w_{\ell,m}^{(r,k)}) + (w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,0)})||^2 && (33)\\
&= \mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)} + w_{g,m}^{(r,0)} - w_{\ell,m}^{(r,k)}) + (w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,0)})||^2 && (34)\\
&\leq 2\mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)})||^2 + 2\mathbb{E}||(1 - \psi_{g,m}^{(r,k)})(w_{\ell,m}^{(r,K)} - w_{\ell,m}^{(r,0)})||^2 && (35)
\end{aligned}$$

Using lemmas D.6 and D.7,

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 \leq\ & 2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \left(8k^3\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 6K^2\eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2\right)\\
&+ 2\mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \left(4k\eta_\ell^2\sigma_\ell^2 + 3K\eta_\ell^2\sigma_\ell^2\right) && (36)\\
\leq\ & 16k^3\eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 8k\eta_\ell^2\sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2\\
&+ 12K^2\eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 6K\eta_\ell^2\sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 && (37)
\end{aligned}$$

□

Theorem D.9 (Convergence of the Global Model for Convex Cases)**.** *If each client's objective function f_m satisfies Assumptions D.1, D.2, D.3, D.4 using the learning rate $\frac{1}{\mu R} \leq \eta_\ell \leq \min\left(\frac{1}{4\sqrt{10}\beta B K^2}, \frac{1}{8B^2\beta}\right)$ in Algorithm 2, then the following convergence holds:*
(Strong Convex Case)

$$\begin{aligned}
\mathbb{E}\left[F(w_g^{(R)})\right] - F(w_g^*) \leq\ & \frac{\mu}{\mathbf{q}_0^2 K} \mathbb{E}||w_g^{(0)} - w_g^*||^2 \exp\left(-\frac{\eta_\ell \mu K R}{2M}\right) + \frac{2G^2}{\mathbf{q}_0^2 \mu R}\\
&+ \frac{40K^2\beta}{\mu^2 R^2}\left(\frac{\beta^2}{\mu R} + 1\right)\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2} G^2 + \frac{28K\beta}{\mu^2 R^2}\left(\frac{2\beta^2 K}{\mu^2 R^2} + 1\right)\sigma_\ell^2
\end{aligned}$$

(General Convex Case)

$$\begin{aligned}
\mathbb{E}\left[F(w_g^{(R)})\right] - F(w_g^*) \leq\ & \frac{1}{\eta_\ell K \mathbf{q}_0^2 (R+1)} \mathbb{E}||w_g^{(0)} - w_g^*||^2 + \eta_\ell \left(\frac{2G^2}{\mathbf{q}_0^2}\right)^{1/2} +\\
&+ \eta_\ell^2 \left(40K^2\beta\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2} G^2\right)^{1/3} + \eta_\ell^3 \left(40K^2\beta^3\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2} G^2\right)^{1/4} + \eta_\ell^2 \left(28K\beta\sigma_\ell^2\right)^{1/3} + \eta_\ell^4 \left(56K\beta^3\sigma_\ell^2\right)^{1/5}
\end{aligned}$$

where $\mathbf{q}_{0/1}^2$ are the probabilities of picking global/local routes averaged over all the instances sampled from the global distribution.

Proof. From the update rules stated in Equations 10 and 11:

$$w_g^{(r+1)} - w_g^* = \frac{1}{nM} \sum_{m \in [M]} n_m \left[w_{g,m}^{(r)} - \eta_\ell \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) \right] - w_g^* \tag{38}$$

$$= \frac{1}{nM} \sum_{m \in [M]} n_m w_{g,m}^{(r)} - \frac{\eta_\ell}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) - w_g^* \tag{39}$$

$$= w_g^{(r)} - w_g^* - \frac{\eta_\ell}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) \tag{40}$$

Taking squared norm and expectation on both sides with respect to the choice of h_m,

$$\mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2 \right] \leq \mathbb{E}\left[||w_g^{(r)} - w_g^*||^2 \right] - 2\eta_\ell \left\langle \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \mathbb{E}[h_m(w_{p,m}^{(r,k-1)})], w_g^{(r)} - w_g^* \right\rangle$$
$$+ \eta_\ell^2 \mathbb{E}\left[\left|\left| \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) \right|\right|^2 \right] \tag{41}$$

Separating mean and variance according to Lemma 4 of Scaffold [12],

$$\leq \mathbb{E}\left[||w_g^{(r)} - w_g^*||^2 \right] \underbrace{-2\eta_\ell \left\langle \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \mathbb{E}[\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})], w_g^{(r)} - w_g^* \right\rangle}_{T_1}$$
$$\underbrace{+\eta_\ell^2 \mathbb{E}\left[\left|\left| \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right|\right|^2 \right]}_{T_2} + \frac{\eta_\ell^2 \sigma_\ell^2 K}{M} \tag{42}$$

Bounding T_1

$$T_1 = -2\eta_\ell \left\langle \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \mathbb{E}[\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})], w_g^{(r)} - w_g^* \right\rangle \tag{43}$$

$$= 2\eta_\ell \left\langle \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \mathbb{E}[\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})], w_g^* - w_g^{(r)} \right\rangle \tag{44}$$

Using perturbed strong convexity lemma (Lemma 5) from [12], we get,

$$T_1 \leq \frac{2\eta_\ell}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \left(\mathbb{E}[\nabla f_m(w_g^*)] - \nabla f_m(w_g^{(r)}) - \frac{\mu}{4} \mathbb{E}||w_g^{(r)} - w_g^*||^2 + \beta \underbrace{\mathbb{E}||w_{p,m}^{(r,k-1)} - w_g^{(r)}||^2}_{\text{Lemma D.8}} \right) \tag{45}$$

$$\leq -2\eta_\ell K \left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*) \right) - \frac{\eta_\ell \mu K}{2M} \mathbb{E}||w_g^{(r)} - w_g^*||^2$$
$$+ \frac{2\eta_\ell \beta}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^{K} \Big(16k^3 \eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 8k \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2$$
$$+ 12K^2 \eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 6K \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \Big) \tag{46}$$

$$\leq -2\eta_\ell K \left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*) \right) - \frac{\eta_\ell \mu K}{2M} \mathbb{E}||w_g^{(r)} - w_g^*||^2$$
$$+ \frac{2\eta_\ell \beta}{nM} \sum_{m \in [M]} n_m \Big(16K^4 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||1 - \psi_{g,m}^{(r,K)}||^2 + 8K^2 \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,K)}||^2$$
$$+ 12K^3 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||1 - \psi_{g,m}^{(r,K)}||^2 + 6K^2 \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,K)}||^2 \Big), \tag{47}$$

Next, using Assumption D.4,

$$\begin{aligned}
T_1 \le& -2\eta_\ell K\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) - \frac{\eta_\ell \mu K}{2M}\mathbb{E}||w_g^{(r)} - w_g^*||^2 \\
&+ 32\eta_\ell^3 K^4 \beta \mathbb{E}||1-\psi_g^{(r)}||^2\left(G^2 + 2\beta B^2\left(\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)\right)\right) + 16\eta_\ell^3 K^2\beta\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2 \\
&+ 24\eta_\ell^3 K^3 \beta \mathbb{E}||1-\psi_g^{(r)}||^2\left(G^2 + 2\beta B^2\left(\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)\right)\right) + 12\eta_\ell^3 K^2\beta\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2
\end{aligned} \tag{48}$$

$$\begin{aligned}
\le& -2\eta_\ell K\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) - \frac{\eta_\ell \mu K}{2M}\mathbb{E}||w_g^{(r)} - w_g^*||^2 \\
&+ 16\eta_\ell^3 K^3\beta^2 B^2(4K+3)\mathbb{E}||1-\psi_g^{(r)}||^2\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) \\
&+ 8\eta_\ell^3 K^3\beta(4K+3)\mathbb{E}||1-\psi_g^{(r)}||^2 G^2 + 28\eta_\ell^3 K^2\beta\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2
\end{aligned} \tag{49}$$

Bounding T_2

$$T_2 = \eta_\ell^2\mathbb{E}\left[\left|\left|\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})\right|\right|^2\right] \tag{50}$$

$$= \eta_\ell^2\mathbb{E}\left[\left|\left|\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}(\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) - \nabla f_m(w_g^{(r)}) + \nabla f_m(w_g^{(r)}))\right|\right|^2\right] \tag{51}$$

$$\begin{aligned}
\le& 2\eta_\ell^2\mathbb{E}\left[\left|\left|\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}(\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) - \nabla f_m(w_g^{(r)}))\right|\right|^2\right] \\
&+ 2\eta_\ell^2\mathbb{E}\left[\left|\left|\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}\nabla f_m(w_g^{(r)}))\right|\right|^2\right]
\end{aligned} \tag{52}$$

$$\le 2\eta_\ell^2\beta^2 K\cdot\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}\underbrace{\mathbb{E}\left[\left|\left|w_{p,m}^{(r,k-1)} - w_g^{(r)}\right|\right|^2\right]}_{\text{Lemma D.8}} + 2\eta_\ell^2 K\cdot\frac{1}{nM}\sum_{m\in[M]} n_m\sum_{k=1}^{K}\mathbb{E}\left[\left|\left|\nabla f_m(w_g^{(r)}))\right|\right|^2\right] \tag{53}$$

$$\begin{aligned}
\le& 16\eta_\ell^4 K^4\beta^3 B^2(4K+3)\mathbb{E}||1-\psi_g^{(r)}||^2\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) + 8\eta_\ell^4 K^4\beta^3(4K+3)\mathbb{E}||1-\psi_g^{(r)}||^2 G^2 \\
&+ 56\eta_\ell^5 K^3\beta^3\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2 + 2\eta_\ell^2 K\left(G^2 + 2\beta B^2\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right)
\end{aligned} \tag{54}$$

Plugging in T_1 and T_2 bounds,

$$\begin{aligned}
\mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2\right] \le& \mathbb{E}\left[||w_g^{(r)} - w_g^*||^2\right] - 2\eta_\ell K\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) - \frac{\eta_\ell\mu K}{2M}\mathbb{E}||w_g^{(r)} - w_g^*||^2 \\
&+ 16\eta_\ell^3 K^3\beta^2 B^2(4K+3)(\eta_\ell\beta+1)\mathbb{E}||1-\psi_g^{(r)}||^2\left(\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right) \\
&+ 8\eta_\ell^3 K^3\beta(4K+3)(\eta_\ell\beta^2+1)\mathbb{E}||1-\psi_g^{(r)}||^2 G^2 + 28\eta_\ell^3 K^2\beta\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2 \\
&+ 56\eta_\ell^5 K^3\beta^3\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2 + 2\eta_\ell^2 K\left(G^2 + 2\beta B^2\mathbb{E}[F(w_g^{(r)})] - F(w_g^*)\right)
\end{aligned} \tag{55}$$

Rearranging the terms, and replacing $\mathbb{E}||\psi_g^{(r)}||^2$ and $\mathbb{E}||1-\psi_g^{(r)}||^2$ with $\mathbf{q}_0^2$ (probability of picking global route averaged over the instances sampled from the global distribution) and $\mathbf{q}_1^2$ respectively,

$$\begin{aligned}
\mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2\right] \le& \left(1 - \frac{\eta_\ell\mu K}{2M}\right)\mathbb{E}\left[||w_g^{(r)} - w_g^*||^2\right] \\
&- \left(2\eta_\ell K - 80\eta_\ell^3 K^4\beta^2 B^2(\eta_\ell\beta+1)\mathbf{q}_1^2 - 4\eta_\ell^2 K\beta B^2\right)\left(\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)\right) \\
&+ 40\eta_\ell^3 K^3\beta(\eta_\ell\beta^2+1)\mathbf{q}_1^2 G^2 + 2\eta_\ell^2 K G^2 + 28\eta_\ell^3 K^2\beta(2\eta_\ell^2\beta^2 K+1)\mathbf{q}_0^2\sigma_\ell^2
\end{aligned} \tag{56}$$

Assuming $\frac{\eta_\ell K}{2} \geq 80\eta_\ell^3 K^4\beta^2 B^2(\eta_\ell\beta+1) \implies \eta_\ell \leq \frac{1}{4\sqrt{10}\beta B K^2}$ and $\frac{\eta_\ell K}{2} \geq 4\eta_\ell^2 K\beta B^2 \implies \eta_\ell \leq \frac{1}{8B^2\beta}$, we get

$$\begin{aligned}\mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2\right] \leq \left(1 - \frac{\eta_\ell \mu K}{2M}\right)\mathbb{E}\left[||w_g^{(r)} - w_g^*||^2\right] - \eta_\ell K(1-\mathbf{q}_1)^2\left(\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)\right) \\ + 40\eta_\ell^3 K^3\beta(\eta_\ell\beta^2+1)\mathbf{q}_1^2 G^2 + 2\eta_\ell^2 KG^2 + 28\eta_\ell^3 K^2\beta(2\eta_\ell^2\beta^2 K+1)\mathbf{q}_0^2\sigma_\ell^2\end{aligned} \tag{57}$$

Moving $\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)$ to the left-hand side, and rest of the terms on right-hand side,

$$\begin{aligned}\eta_\ell K\mathbf{q}_0^2\left(\mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*)\right) &\leq \left(1 - \frac{\eta_\ell \mu K}{2M}\right)\mathbb{E}\left[||w_g^{(r)} - w_g^*||^2\right] - \mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2\right] \\ &+ 40\eta_\ell^3 K^3\beta(\eta_\ell\beta^2+1)\mathbf{q}_1^2 G^2 + 2\eta_\ell^2 KG^2 + 28\eta_\ell^3 K^2\beta(2\eta_\ell^2\beta^2 K+1)\mathbf{q}_0^2\sigma_\ell^2 \qquad (58) \\ \therefore \mathbb{E}\left[F(w_g^{(r)})\right] - F(w_g^*) &\leq \frac{1}{\eta_\ell K\mathbf{q}_0^2}\left(1 - \frac{\eta_\ell \mu K}{2M}\right)\mathbb{E}\left[||w_g^{(r)} - w_g^*||^2\right] - \frac{1}{\eta_\ell K\mathbf{q}_0^2}\mathbb{E}\left[||w_g^{(r+1)} - w_g^*||^2\right] \\ &+ 40\eta_\ell^2 K^2\beta(\eta_\ell\beta^2+1)\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}G^2 + \frac{2\eta_\ell G^2}{\mathbf{q}_0^2} + 28\eta_\ell^2 K\beta(2\eta_\ell^2\beta^2 K+1)\sigma_\ell^2 \qquad (59)\end{aligned}$$

Unrolling the recursion over R rounds and then using the linear convergence lemma (Lemma 1) for strong convex case from Scaffold [12],

$$\begin{aligned}\mathbb{E}\left[F(w_g^{(R)})\right] - F(w_g^*) \leq \frac{\mu}{\mathbf{q}_0^2 K}\mathbb{E}||w_g^{(0)} - w_g^*||^2\exp\left(-\frac{\eta_\ell\mu KR}{2M}\right) + \frac{2G^2}{\mathbf{q}_0^2\mu R} \\ + \frac{40K^2\beta}{\mu^2R^2}\left(\frac{\beta^2}{\mu R}+1\right)\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}G^2 + \frac{28K\beta}{\mu^2R^2}\left(\frac{2\beta^2 K}{\mu^2R^2}+1\right)\sigma_\ell^2\end{aligned} \tag{60}$$

Unrolling the recursion over R rounds and then using the sublinear convergence lemma (Lemma 2) for general convex case from Scaffold [12],

$$\begin{aligned}\mathbb{E}\left[F(w_g^{(R)})\right] - F(w_g^*) \leq \frac{1}{\eta_\ell K\mathbf{q}_0^2(R+1)}\mathbb{E}||w_g^{(0)} - w_g^*||^2 + \eta_\ell\left(\frac{2G^2}{\mathbf{q}_0^2}\right)^{1/2} + \\ + \eta_\ell^2\left(40K^2\beta\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}G^2\right)^{1/2} + \eta_\ell^3\left(40K^2\beta^3\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}G^2\right)^{1/3} + \eta_\ell^2\left(28K\beta\sigma_\ell^2\right)^{1/3} + \eta_\ell^4\left(56K\beta^3\sigma_\ell^2\right)^{1/5}\end{aligned} \tag{61}$$

□

D.4 Convergence Proof for the Global Model: Non-convex Case

We start with a non-convex version of Lemmas D.7 and D.8,

Lemma D.10 (Local version of the global model progress)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.2, D.3, in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 \leq 4k^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 2k\eta_\ell^2\sigma_\ell^2 + 4k^2\eta_\ell^2\beta^2\sum_{i=1}^{k}\mathbb{E}||w_{p,m}^{(r,i-1)} - w_g^{(r)}||^2$$

Proof. We start by expanding $w_{g,m}^{(r,k)}$ in terms of its previous epoch iterate.

$$\mathbb{E}||w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}||^2 = \mathbb{E}||w_{g,m}^{(r,k-1)} - \eta_\ell\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) - w_{g,m}^{(r,0)}||^2 \tag{62}$$

Using triangle inequality and separation of variance, we get,

$$\begin{aligned}&\leq \left(1+\frac{1}{k-1}\right)\mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + k\eta_\ell^2\mathbb{E}||\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})||^2 + \eta_\ell^2\sigma_\ell^2 \qquad (63) \\ &\leq \left(1+\frac{1}{k-1}\right)\mathbb{E}||w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)}||^2 + \eta_\ell^2\sigma_\ell^2 \\ &\quad + k\eta_\ell^2\mathbb{E}||\nabla f_m(w_{p,m}^{(r,k-1)}) - \nabla f_m(w_{g,m}^{(r,0)}) + \nabla f_m(w_{g,m}^{(r,0)})||^2 \qquad (64)\end{aligned}$$

(65)

$$\leq \left(1+\frac{1}{k-1}\right)\mathbb{E}||w_{g,m}^{(r,k-1)}-w_{g,m}^{(r,0)}||^2+\eta_\ell^2\sigma_\ell^2+2k\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)}))||^2$$
$$+2k\eta_\ell^2\mathbb{E}||\nabla f_m(w_{p,m}^{(r,k-1)})-\nabla f_m(w_g^{(r)})||^2 \tag{66}$$
$$\leq \left(1+\frac{1}{k-1}\right)\mathbb{E}||w_{g,m}^{(r,k-1)}-w_{g,m}^{(r,0)}||^2+\eta_\ell^2\sigma_\ell^2+2k\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)}))||^2$$
$$+2k\eta_\ell^2\beta^2\mathbb{E}||w_{p,m}^{(r,k-1)}-w_g^{(r)}||^2 \tag{67}$$
$$\tag{68}$$

Unrolling the recursion,

$$\mathbb{E}||w_{g,m}^{(r,k)}-w_{g,m}^{(r,0)}||^2\leq\sum_{i=1}^{k}\left(2k\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+\eta_\ell^2\sigma_\ell^2+2k\eta_\ell^2\beta^2\mathbb{E}||w_{p,m}^{(r,k-1)}-w_g^{(r)}||^2\right)$$
$$\cdot\left(1+\frac{1}{k-1}\right)^i \tag{69}$$

$$\mathbb{E}||w_{g,m}^{(r,k)}-w_{g,m}^{(r,0)}||^2\leq 2k\left(2k\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+\eta_\ell^2\sigma_\ell^2+2k\eta_\ell^2\beta^2\sum_{i=1}^{k}\mathbb{E}||w_{p,m}^{(r,i-1)}-w_g^{(r)}||^2\right) \tag{70}$$
$$=4k^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+2k\eta_\ell^2\sigma_\ell^2+4k^2\eta_\ell^2\beta^2\sum_{i=1}^{k}\mathbb{E}||w_{p,m}^{(r,i-1)}-w_g^{(r)}||^2 \tag{71}$$

□

Lemma D.11 (Deviation of the personalized model from the global model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.2, D.3, and condition $\eta_\ell\leq\frac{1}{2\sqrt{2}\beta K}$ in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{p,m}^{(r,k)}-w_{g,m}^{(r,0)}||^2\leq 20K^3\eta_\ell^2\mathbb{E}||1-\psi_{g,m}^{(r,k)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+10K^2\eta_\ell^2\sigma_\ell^2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2.$$

Proof.

$$\mathbb{E}||w_{p,m}^{(r,k)}-w_{g,m}^{(r,0)}||^2=\mathbb{E}||\psi_{g,m}^{(r,k)}w_{g,m}^{(r,k)}+(1-\psi_{g,m}^{(r,k)})w_{\ell,m}^{(r,K)}-w_{g,m}^{(r,0)}||^2 \tag{72}$$
$$=\mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)}-w_{\ell,m}^{(r,k)})+(w_{\ell,m}^{(r,K)}-w_{g,m}^{(r,0)})||^2 \tag{73}$$
$$=\mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)}-w_{g,m}^{(r,0)}+w_{g,m}^{(r,0)}-w_{\ell,m}^{(r,k)})+(w_{\ell,m}^{(r,K)}-w_{g,m}^{(r,0)})||^2 \tag{74}$$
$$\leq 2\mathbb{E}||\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,k)}-w_{g,m}^{(r,0)})||^2+2\mathbb{E}||(1-\psi_{g,m}^{(r,k)})(w_{\ell,m}^{(r,K)}-w_{\ell,m}^{(r,0)})||^2 \tag{75}$$

Using lemmas D.6 and D.10,

$$\mathbb{E}||w_{p,m}^{(r,k)}-w_{g,m}^{(r,0)}||^2\leq 2\mathbb{E}||1-\psi_{g,m}^{(r,k+1)}||^2\left(4K^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+6K^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2\right)$$
$$+2\mathbb{E}||\psi_{g,m}^{(r,k+1)}||^2\left(2K\eta_\ell^2\sigma_\ell^2+3K\eta_\ell^2\sigma_\ell^2+4K^2\eta_\ell^2\beta^2\sum_{i=1}^{k}\mathbb{E}||w_{p,m}^{(r,i-1)}-w_g^{(r)}||^2\right) \tag{76}$$

Assuming $8K^2\eta_\ell^2\beta^2\leq 1 \implies \eta\leq\frac{1}{2\sqrt{2}\beta K}$ and unrolling the recursion over $w_{p,m}^{(r,i-1)}-w_g^{(r)}$,

$$\leq\sum_{i=1}^{k}\left(20K^2\eta_\ell^2\mathbb{E}||1-\psi_{g,m}^{(r,k)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+10K\eta_\ell^2\sigma_\ell^2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2\right) \tag{77}$$
$$\leq 20K^3\eta_\ell^2\mathbb{E}||1-\psi_{g,m}^{(r,k)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2+10K^2\eta_\ell^2\sigma_\ell^2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \tag{78}$$

□

Theorem D.12 (Convergence of the Global Model for Non-convex Case)**.** *If each client's objective function f_m satisfies Assumptions D.2, D.3, D.4, using the learning rate $\frac{1}{2\beta}\leq\eta_\ell\leq\min\left(\frac{1}{2\sqrt{5}\beta BK^2},\frac{1}{\sqrt[3]{40K^4\beta^3B^2}}\right)$ in Algorithm 2, then the following convergence holds:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2\leq\frac{2}{\eta_\ell\mathbf{q}_0^2R}\left[\mathbb{E}\left[F(w_g^{(1)})\right]-\mathbb{E}\left[F(w_g^{(R+1)})\right]\right]+\frac{\eta_\ell\beta\sigma_\ell^2K}{M\mathbf{q}_0^2}$$
$$+40\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}K^4\beta^2\eta_\ell G^2\left(\frac{2\beta\eta_\ell^2-\eta_\ell}{2}\right)+20K^3\beta^2\eta_\ell\sigma_\ell^2\left(\beta\eta_\ell^2-\frac{\eta_\ell}{2}\right).$$

Proof. From the update rule stated in Equation 11, and β-smoothness of f_m, we have

$$F(w_g^{(r+1)}) \leq F(w_g^{(r)}) + \left\langle \nabla F(w_g^{(r)}), w_g^{(r+1)} - w_g^{(r)} \right\rangle + \frac{\beta}{2} ||w_g^{(r+1)} - w_g^{(r)}||^2 \tag{79}$$

Taking expectation on both sides,

$$\mathbb{E}\left[F(w_g^{(r+1)})\right] \leq \mathbb{E}\left[F(w_g^{(r)})\right] + \mathbb{E}\left[\left\langle \nabla F(w_g^{(r)}), w_g^{(r+1)} - w_g^{(r)} \right\rangle\right] + \frac{\beta}{2} \|w_g^{(r+1)} - w_g^{(r)}\|^2 \tag{80}$$

Using Equation 10 for second and third terms, and using the fact that the expectation is with respect to the choice of h_m,

$$\begin{aligned} &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \eta_\ell \left\langle \nabla F(w_g^{(r)}), \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \mathbb{E}\left[h_m(w_{p,m}^{(r,k-1)})\right] \right\rangle \\ &\quad + \frac{\beta \eta_\ell^2}{2} \mathbb{E} \left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} h_m(w_{p,m}^{(r,k-1)}) \right\|^2, \end{aligned} \tag{81}$$

where $\alpha_m = \frac{n_m}{n}$, which are the weights for weighted aggregation according to the sample count, as shown in Equation 11.

Separating mean and variance according to Assumption D.3,

$$\begin{aligned} \mathbb{E}\left[F(w_g^{(r+1)})\right] &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \eta_\ell \left\langle \nabla F(w_g^{(r)}), \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \mathbb{E}\left[\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})\right] \right\rangle \\ &\quad + \frac{\beta \eta_\ell^2}{2} \mathbb{E} \left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 \right] + \frac{\eta_\ell^2 \beta \sigma_\ell^2 K}{2M} \end{aligned} \tag{82}$$

Using $\langle a, b \rangle = -\frac{1}{2}||a-b||^2 + \frac{1}{2}||a||^2 + \frac{1}{2}||b||^2$,

$$\begin{aligned} \mathbb{E}\left[F(w_g^{(r+1)})\right] &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \eta_\ell \left[-\frac{1}{2} \mathbb{E} \left\| \nabla F(w_g^{(r)}) - \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 \right] \\ &\quad - \eta_\ell \left[\frac{1}{2} \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 + \frac{1}{2} \mathbb{E} \left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 \right] \\ &\quad + \frac{\beta \eta_\ell^2}{2} \mathbb{E} \left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 \right] + \frac{\eta_\ell^2 \beta \sigma_\ell^2 K}{2M} \qquad (83) \\ &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \frac{\eta_\ell}{2} \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 \\ &\quad - \left(\frac{\eta_\ell}{2} - \frac{\beta \eta_\ell^2}{2} \right) \mathbb{E} \left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 \\ &\quad + \frac{\eta_\ell}{2} \mathbb{E} \left\| \nabla F(w_g^{(r)}) - \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) \right\|^2 + \frac{\eta_\ell^2 \beta \sigma_\ell^2 K}{2M} \qquad (84) \end{aligned}$$

(85)

$$
\begin{aligned}
&\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \frac{\eta_\ell}{2}\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 \\
&\quad - \left(\frac{\eta_\ell}{2} - \frac{\beta\eta_\ell^2}{2}\right)\mathbb{E}\left\|\nabla F(w_g^{(r)}) - \frac{1}{M}\sum_{m\in[M]}\alpha_m\sum_{k=1}^{K}\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)}) - \nabla F(w_g^{(r)}))\right\|^2 \\
&\quad + \frac{\eta_\ell}{2}\mathbb{E}\left\|\nabla F(w_g^{(r)}) - \frac{1}{M}\sum_{m\in[M]}\alpha_m\sum_{k=1}^{K}\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})\right\|^2 + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} &(86)\\
&\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \left(\frac{3\eta_\ell}{2} - \beta\eta_\ell^2\right)\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} \\
&\quad - \left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)\mathbb{E}\left\|\nabla F(w_g^{(r)}) - \frac{1}{M}\sum_{m\in[M]}\alpha_m\sum_{k=1}^{K}\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k-1)})\right\|^2 &(87)\\
&\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \left(\frac{3\eta_\ell}{2} - \beta\eta_\ell^2\right)\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} \\
&\quad - \left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)\beta^2 K\cdot\frac{1}{M}\sum_{m\in[M]}\alpha_m\sum_{k=1}^{K}\mathbb{E}\left\|w_g^{(r)} - w_{p,m}^{(r,k-1)}\right\|^2 &(88)
\end{aligned}
$$

Using Lemma D.11,

$$
\begin{aligned}
\mathbb{E}\left[F(w_g^{(r+1)})\right] &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \left(\frac{3\eta_\ell}{2} - \beta\eta_\ell^2\right)\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} \\
&\quad - \left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)\beta^2 K\cdot\frac{1}{M}\sum_{m\in[M]}\alpha_m\sum_{k=1}^{K}\Bigg(20K^3\eta_\ell^2\mathbb{E}||1-\psi_{g,m}^{(r,k)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \\
&\quad + 10K^2\eta_\ell^2\sigma_\ell^2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2\Bigg) &(89)
\end{aligned}
$$

Using Assumption D.4 for non-convex case, we get,

$$
\begin{aligned}
\mathbb{E}\left[F(w_g^{(r+1)})\right] &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \left(\frac{3\eta_\ell}{2} - \beta\eta_\ell^2\right)\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} \\
&\quad - \left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)20\beta^2K^4\eta_\ell^2(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2)\mathbb{E}||1-\psi_g^{(r)}||^2 \\
&\quad - \left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)\left(10\beta^2K^3\eta_\ell^2\sigma_\ell^2\mathbb{E}||\psi_g^{(r)}||^2\right) &(90)
\end{aligned}
$$

Rearranging the terms to put $\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2$ on left-hand side,

$$
\begin{aligned}
&\left(\frac{3\eta_\ell}{2} - \beta\eta_\ell^2 - 20K^4\beta^2\eta_\ell^2B^2\mathbf{q}_1^2\left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right)\right)\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 \leq \mathbb{E}\left[F(w_g^{(r)})\right] - \mathbb{E}\left[F(w_g^{(r+1)})\right] \\
&\quad - 20\mathbf{q}_1^2K^4\beta^2\eta_\ell^2G^2\left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right) - 10\mathbf{q}_0^2K^3\beta^2\eta_\ell^2\sigma_\ell^2\left(\frac{\eta_\ell}{2} - \beta\eta_\ell^2\right) + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} &(91)
\end{aligned}
$$

Assuming $10K^4\beta^2\eta_\ell^3B^2 \leq \frac{\eta_\ell}{2} \implies \eta_\ell \leq \frac{1}{2\sqrt{5}\beta BK^2}$ and $20K^4\beta^3\eta_\ell^4B^2 \leq \frac{\eta_\ell}{2} \implies \eta_\ell \leq \frac{1}{\sqrt[3]{40K^4\beta^3B^2}}$,

$$
\begin{aligned}
\left(\frac{\eta_\ell}{2}\right)\mathbf{q}_0^2\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 &\leq \mathbb{E}\left[F(w_g^{(r)})\right] - \mathbb{E}\left[F(w_g^{(r+1)})\right] + \frac{\eta_\ell^2\beta\sigma_\ell^2 K}{2M} \\
&\quad + 20\mathbf{q}_1^2K^4\beta^2\eta_\ell^2G^2\left(\frac{2\beta\eta_\ell^2 - \eta_\ell}{2}\right) + 10\mathbf{q}_0^2K^3\beta^2\eta_\ell^2\sigma_\ell^2\left(\beta\eta_\ell^2 - \frac{\eta_\ell}{2}\right) &(92)\\
\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 &\leq \frac{2}{\eta_\ell\mathbf{q}_0^2}\left[\mathbb{E}\left[F(w_g^{(r)})\right] - \mathbb{E}\left[F(w_g^{(r+1)})\right]\right] + \frac{\eta_\ell\beta\sigma_\ell^2 K}{M\mathbf{q}_0^2} \\
&\quad + 40\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}K^4\beta^2\eta_\ell G^2\left(\frac{2\beta\eta_\ell^2 - \eta_\ell}{2}\right) + 20K^3\beta^2\eta_\ell\sigma_\ell^2\left(\beta\eta_\ell^2 - \frac{\eta_\ell}{2}\right) &(93)
\end{aligned}
$$

Taking average over all the R rounds,

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F(w_g^{(r)})\right\|^2 \leq \frac{2}{\eta_\ell \mathbf{q}_0^2 R}\left[\mathbb{E}\left[F(w_g^{(1)})\right] - \mathbb{E}\left[F(w_g^{(R+1)})\right]\right] + \frac{\eta_\ell \beta \sigma_\ell^2 K}{M\mathbf{q}_0^2}$$
$$+ 40\frac{\mathbf{q}_1^2}{\mathbf{q}_0^2}K^4\beta^2\eta_\ell G^2\left(\frac{2\beta\eta_\ell^2 - \eta_\ell}{2}\right) + 20K^3\beta^2\eta_\ell\sigma_\ell^2\left(\beta\eta_\ell^2 - \frac{\eta_\ell}{2}\right) \tag{94}$$

□

D.5 Convergence Proof for the Personalized Model: Convex (Strong and General) Cases

Lemma D.13 (Local progress of the personalized model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.1, D.2, D.3, and D.4 and conditioning on $\eta_\ell \leq \frac{1}{\beta\sqrt{6K}}$ in Algorithm 2, the following are satisfied:*

$$\mathbb{E}||w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}||^2 \leq 18K^2\eta_\ell^2\mathbb{E}\left\|\nabla f_m(\tilde{w}_{p,m}^{(r,0)})\right\|^2 + 108K^4\eta_\ell^4\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 126K^3\eta_\ell^4\sigma_\ell^2$$
$$+ 9K^2\eta_\ell^2\mathbb{E}\left\|\psi_{g,m}^{(r,K)}\right\|^2 + 144K^5\eta_\ell^4\mathbb{E}||\psi_{g,m}^{(r,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2$$

Proof.

$$\mathbb{E}||w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}||^2 \leq \mathbb{E}||\psi_{g,m}^{(r,K+1)}w_{g,m}^{(r,K)} + (1-\psi_{g,m}^{(r,K+1)})w_{\ell,m}^{(r,K)} - \psi_{g,m}^{(r,0)}w_{g,m}^{(r,0)} - (1-\psi_{g,m}^{(r,0)})w_{\ell,m}^{(r,K)}||^2 \tag{95}$$

$$= \mathbb{E}\Big\|\left(\psi_{g,m}^{(r,K)} - \eta_\ell\nabla_{\psi_{g,m}^{(r,K)}}f_m(\tilde{w}_{p,m}^{(r,K)})\right)\left(w_{g,m}^{(r,K-1)} - \eta_\ell\nabla_{w_{g,m}^{(r,K-1)}}f_m(w_{p,m}^{(r,K-1)})\right)$$
$$+ \left(1 - \psi_{g,m}^{(r,K)} + \eta_\ell\nabla_{\psi_{g,m}^{(r,K)}}f_m(\tilde{w}_{p,m}^{(r,K)})\right)w_{\ell,m}^{(r,K)} - \psi_{g,m}^{(r,0)}w_{g,m}^{(r,0)} - (1-\psi_{g,m}^{(r,0)})w_{\ell,m}^{(r,K)}\Big\|^2 \tag{96}$$

$$= \mathbb{E}||\psi_{g,m}^{(r,K)}w_{g,m}^{(r,K-1)} - \psi_{g,m}^{(r,K)}\eta_\ell\nabla_{w_{g,m}^{(r,K-1)}}f_m(w_{p,m}^{(r,K-1)}) - w_{g,m}^{(r,K-1)}\eta_\ell\nabla_{\psi_{g,m}^{(r,K)}}f_m(\tilde{w}_{p,m}^{(r,K)})$$
$$+ \eta_\ell^2\nabla_{\psi_{g,m}^{(r,K)}}f_m(\tilde{w}_{p,m}^{(r,K)})\nabla_{w_{g,m}^{(r,K-1)}}f_m(w_{p,m}^{(r,K-1)}) + \left(1-\psi_{g,m}^{(r,K)}\right)w_{\ell,m}^{(r,K)}$$
$$+ w_{\ell,m}^{(r,K)}\eta_\ell\nabla_{\psi_{g,m}^{(r,K)}}f_m(\tilde{w}_{p,m}^{(r,K)}) - \psi_{g,m}^{(r,0)}w_{g,m}^{(r,0)} - (1-\psi_{g,m}^{(r,0)})w_{\ell,m}^{(r,K)}||^2 \tag{97}$$

Using the convexity of f_m,

$$\nabla_{w_{g,m}^{(r,k)}}f_m(w_{p,m}^{(r,k)}) = \nabla_{w_{g,m}^{(r,k)}}f_m(\psi_{g,m}^{(r,k+1)}w_{g,m}^{(r,k)} + (1-\psi_{g,m}^{(r,k+1)})w_{\ell,m}^{(r,K)}) \tag{98}$$
$$\leq \psi_{g,m}^{(r,k+1)}\nabla f_m(w_{g,m}^{(r,k)}) \tag{99}$$

and

$$\nabla_{\psi_{g,m}^{(r,k)}}f_m(\tilde{w}_{p,m}^{(r,k)}) = \nabla_{\psi_{g,m}^{(r,k)}}f_m(\psi_{g,m}^{(r,k)}[w_{g,m}^{(r,k)} - w_{\ell,m}^{(r,K)}] - w_{\ell,m}^{(r,K)}) \tag{100}$$
$$\leq (w_{g,m}^{(r,k)} - w_{\ell,m}^{(r,K)})\nabla f_m(\psi_{g,m}^{(r,k)}) \tag{101}$$

we get,

$$\mathbb{E}||w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}||^2 \leq \mathbb{E}||\psi_{g,m}^{(r,K)}w_{g,m}^{(r,K-1)} + \left(1-\psi_{g,m}^{(r,K)}\right)w_{\ell,m}^{(r,K)} - \psi_{g,m}^{(r,0)}w_{g,m}^{(r,0)} - (1-\psi_{g,m}^{(r,0)})w_{\ell,m}^{(r,K)}$$
$$- \eta_\ell(\psi_{g,m}^{(r,K)})^2\nabla f_m(w_{g,m}^{(r,K-1)}) + \eta_\ell(w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,K-1)})^2\nabla f_m(w_{g,m}^{(r,K)})||^2 \tag{102}$$

$$\leq \left(1+\frac{1}{K-1}\right)\mathbb{E}||w_{p,m}^{(r,K-1)} - \tilde{w}_{p,m}^{(r,0)}||^2 + 3K\eta_\ell^2\mathbb{E}\left\|\nabla f_m(w_{p,m}^{(r,K-1)})\right\|^2$$
$$+ 3K\eta_\ell^2\mathbb{E}\left\|w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,K-1)}\right\|^2 + 3K\eta_\ell^2\mathbb{E}\left\|\psi_{g,m}^{(r,K)}\right\|^2 \tag{103}$$

$$\leq \left(1+\frac{1}{K-1}\right)\mathbb{E}||w_{p,m}^{(r,K-1)} - \tilde{w}_{p,m}^{(r,0)}||^2 + 3K\eta_\ell^2\mathbb{E}\left\|\nabla f_m(w_{p,m}^{(r,K-1)}) - \nabla f_m(\tilde{w}_{p,m}^{(r,0)}) + \nabla f_m(\tilde{w}_{p,m}^{(r,0)})\right\|^2$$
$$+ 6K\eta_\ell^2\mathbb{E}\left\|w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,0)}\right\|^2 + 6K\eta_\ell^2\mathbb{E}\left\|w_{g,m}^{(r,0)} - w_{g,m}^{(r,K-1)}\right\|^2 + 3K\eta_\ell^2\mathbb{E}\left\|\psi_{g,m}^{(r,K)}\right\|^2 \tag{104}$$

Using Lemma D.6 and D.7, and smoothness property,

$$\leq \left(1+\frac{1}{K-1}+6K\eta_\ell^2\beta^2\right)\mathbb{E}||w_{p,m}^{(r,K-1)} - \tilde{w}_{p,m}^{(r,0)}||^2 + 6K\eta_\ell^2\mathbb{E}\left\|\nabla f_m(\tilde{w}_{p,m}^{(r,0)})\right\|^2$$
$$+ 6K\eta_\ell^2\left(6K^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_{\ell,m}^{(r,0)})||^2 + 3K\eta_\ell^2\sigma_\ell^2\right) + 3K\eta_\ell^2\mathbb{E}\left\|\psi_{g,m}^{(r,K)}\right\|^2$$
$$+ 6K\eta_\ell^2\left(8K^3\eta_\ell^2\mathbb{E}||\psi_{g,m}^{(r,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 4K\eta_\ell^2\sigma_\ell^2\right) \tag{105}$$

$$
\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}||^2 \leq \sum_{i=1}^{K} \Bigg(& 6K\eta_\ell^2 \mathbb{E}\big\|\nabla f_m(\tilde{w}_{p,m}^{(r,0)})\big\|^2 + 36K^3\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 42K^2\eta_\ell^4\sigma_\ell^2 \\
& + 3K\eta_\ell^2\mathbb{E}\big\|\psi_{g,m}^{(r,K)}\big\|^2 + 48K^4\eta_\ell^4\mathbb{E}||\psi_{g,m}^{(r,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \Bigg) \left(1 + \frac{1}{K-1} + 6K\eta_\ell^2\beta^2\right)^i
\end{aligned} \tag{106}
$$

Assuming $6K\eta_\ell^2\beta^2 \leq 1 \implies \eta_\ell \leq \frac{1}{\beta\sqrt{6K}}$,

$$
\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}||^2 \leq 3K \Bigg(& 6K\eta_\ell^2 \mathbb{E}\big\|\nabla f_m(\tilde{w}_{p,m}^{(r,0)})\big\|^2 + 36K^3\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 42K^2\eta_\ell^4\sigma_\ell^2 \\
& + 3K\eta_\ell^2\mathbb{E}\big\|\psi_{g,m}^{(r,K)}\big\|^2 + 48K^4\eta_\ell^4\mathbb{E}||\psi_{g,m}^{(r,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \Bigg) \qquad (107) \\
= 18K^2\eta_\ell^2 & \mathbb{E}\big\|\nabla f_m(\tilde{w}_{p,m}^{(r,0)})\big\|^2 + 108K^4\eta_\ell^4\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 126K^3\eta_\ell^4\sigma_\ell^2 \\
& + 9K^2\eta_\ell^2\mathbb{E}\big\|\psi_{g,m}^{(r,K)}\big\|^2 + 144K^5\eta_\ell^4\mathbb{E}||\psi_{g,m}^{(r,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \qquad (108)
\end{aligned}
$$

□

Lemma D.14 (Deviation of local parameters from the aggregated global parameters)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.3, D.4, in Algorithm 2, the following is satisfied:*

$$
\begin{aligned}
\mathbb{E}||\tilde{w}_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}||^2 \leq 18 & \left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right) \\
& + 6(1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)
\end{aligned}
$$

Proof. Stating the aggregate rule from Algorithm 2, Lines 12 ,19 and 20,

$$
\begin{aligned}
\mathbb{E}||\tilde{w}_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}||^2 &= \mathbb{E}\Big\|\frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)}\frac{1}{M}\sum_{c\in[M]}w_{g,c}^{(r,K)} + \left(1 - \frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)}\right)w_{\ell,m}^{(r+1,K)} \\
&\qquad - \psi_{g,m}^{(r,K)}w_{g,m}^{(r,K)} - (1-\psi_{g,m}^{(r,K)})w_{\ell,m}^{(r,K)}\Big\|^2 \qquad (109) \\
&\leq 2\mathbb{E}\Big\|\frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)}\frac{1}{M}\sum_{c\in[M]}w_{g,c}^{(r,K)} - \psi_{g,m}^{(r,K)}w_{g,m}^{(r,K)}\Big\|^2 \\
&\qquad + 2\mathbb{E}\Big\|\left(1 - \frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)}\right)w_{\ell,m}^{(r+1,K)} - (1-\psi_{g,m}^{(r,K)})w_{\ell,m}^{(r,K)}\Big\|^2 \qquad (110) \\
&\leq 2\mathbb{E}\Big\|\left(\frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)} - \psi_{g,m}^{(r,K)}\right)\left(\frac{1}{M}\sum_{c\in[M]}w_{g,c}^{(r,K)} - w_{g,m}^{(r,K)}\right)\Big\|^2 \\
&\qquad + 2\mathbb{E}\Big\|\left(\psi_{g,m}^{(r,K)} - \frac{1}{M}\sum_{c\in[M]}\psi_{g,c}^{(r,K)}\right)\left(w_{\ell,m}^{(r+1,K)} - w_{\ell,m}^{(r,K)}\right)\Big\|^2 \qquad (111)
\end{aligned}
$$

Using Lemma 8 from [29] and Lemma D.17,

$$
\begin{aligned}
\mathbb{E}||\tilde{w}_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}||^2 \leq 18 & \left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right) \\
& + 6(1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)
\end{aligned} \tag{112}
$$

□

Lemma D.15 (One epoch progress of the personalized model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.1, D.2, D.3, and D.4 in Algorithm 2, the following are satisfied:*

$$
\mathbb{E}||w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,k)}||^2 \leq 3\eta_\ell^2\mathbb{E}\big\|\nabla f_m(w_{p,m}^{(r,k)})\big\|^2 + 3\eta_\ell^2\mathbb{E}\big\|w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}\big\|^2 + 3\eta_\ell^2\mathbb{E}||\psi_{g,m}^{(r,k)}||^2
$$

and hence,

$$\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{(r,k)}||^2 \leq 6\beta\eta_\ell^2 \left(\mathbb{E}[f_m(w_{p,m}^{(r,K)})] - f(w_{p,m}^*)\right) + 3\eta_\ell^2 K \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2$$
$$+ 36K^3\eta_\ell^4\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 40K^2\eta_\ell^4\sigma_\ell^2$$
$$+ 48K^4\eta_\ell^4\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2$$

Proof.

$$\mathbb{E}||w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,k)}||^2 = \mathbb{E}||\psi_{g,m}^{(r,k+1)} w_{g,m}^{(r,k+1)} + (1-\psi_{g,m}^{(r,k+1)})w_{\ell,m}^{(r,K)} - \psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - (1-\psi_{g,m}^{(r,k)})w_{\ell,m}^{(r,K)}||^2 \tag{113}$$

$$= \mathbb{E}\Big\| \left(\psi_{g,m}^{(r,k)} - \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)})\right) \left(w_{g,m}^{(r,k)} - \eta_\ell \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\right)$$
$$+ \left(1 - \psi_{g,m}^{(r,k)} + \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)})\right) w_{\ell,m}^{(r,K)} - \psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - (1-\psi_{g,m}^{(r,k)}) w_{\ell,m}^{(r,K)} \Big\|^2 \tag{114}$$

$$= \mathbb{E}\Big\|\psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - \eta_\ell w_{g,m}^{(r,k)} \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}) - \eta_\ell \psi_{g,m}^{(r,k)} \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})$$
$$+ \eta_\ell^2 \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}) \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) + (1-\psi_{g,m}^{(r,k)}) w_{\ell,m}^{(r,K)}$$
$$+ \eta_\ell w_{\ell,m}^{(r,K)} \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}) - \psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - (1-\psi_{g,m}^{(r,k)}) w_{\ell,m}^{(r,K)} \Big\|^2 \tag{115}$$

$$= \mathbb{E}\Big\|\eta_\ell \left(w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}\right) \nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}) - \eta_\ell \left(\psi_{g,m}^{(r,k+1)}\right) \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\Big\|^2 \tag{116}$$

Using,

$$\nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) = \nabla_{w_{g,m}^{(r,k)}} f_m(\psi_{g,m}^{(r,k+1)} w_{g,m}^{(r,k)} + (1-\psi_{g,m}^{(r,k+1)}) w_{\ell,m}^{(r,K)}) \tag{117}$$
$$\leq \psi_{g,m}^{(r,k+1)} \nabla f_m(w_{g,m}^{(r,k)}) \tag{118}$$

and,

$$\nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}) = \nabla_{\psi_{g,m}^{(r,k)}} f_m(\psi_{g,m}^{(r,k)}[w_{g,m}^{(r,k)} - w_{\ell,m}^{(r,K)}] - w_{\ell,m}^{(r,K)}) \tag{119}$$
$$\leq [w_{g,m}^{(r,k)} - w_{\ell,m}^{(r,K)}] \nabla f_m(\psi_{g,m}^{(r,k)}), \tag{120}$$

we get,

$$\leq \eta_\ell^2 \mathbb{E}\Big\| - \left(w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}\right)^2 \nabla f_m(\psi_{g,m}^{(r,k)}) - \left(\psi_{g,m}^{(r,k+1)}\right)^2 \nabla f_m(w_{g,m}^{(r,k)})\Big\|^2 \tag{121}$$
$$\leq \eta_\ell^2 \mathbb{E}\|\nabla f_m(w_{p,m}^{(r,k)}) + \left(w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}\right) + \psi_{g,m}^{(r,k+1)}\|^2 \tag{122}$$
$$\leq 3\eta_\ell^2 \mathbb{E}\|\nabla f_m(w_{p,m}^{(r,k)})\|^2 + 3\eta_\ell^2 \mathbb{E}\|w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}\|^2 + 3\eta_\ell^2 \mathbb{E}\|\psi_{g,m}^{(r,k+1)}\|^2 \tag{123}$$

From Lemmas D.6 and D.7.

Summing over $i = k$ to K,

$$\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{(r,k)}||^2 = \mathbb{E}||\sum_{i=k}^{K} w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,k)}||^2 \tag{124}$$
$$\leq 3\eta_\ell^2 \sum_{i=k}^{K} \mathbb{E}||\nabla f_m(w_{p,m}^{(r,i)})||^2 + 3\eta_\ell^2 K \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2$$
$$+ 6\eta_\ell^2 \sum_{i=k}^{K} \left(6K^2\eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K\eta_\ell^2\sigma_\ell^2\right)$$
$$+ 6\eta_\ell^2 \sum_{i=k}^{K} \left(8K^3\eta_\ell^2 \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 4K\eta_\ell^2\sigma_\ell^2\right) \tag{125}$$

$$
\begin{aligned}
&\leq 6\beta\eta_\ell^2 \left(\mathbb{E}[f_m(w_{p,m}^{(r,K)})] - f(w_{p,m}^*)\right) + 3\eta_\ell^2 K \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \\
&\quad + 36K^3\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 18K^2\eta_\ell^4\sigma_\ell^2 \\
&\quad + 48K^4\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 + 24K^2\eta_\ell^4\sigma_\ell^2
\end{aligned}
\tag{126}
$$

$$
\begin{aligned}
\therefore \mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{(r,k)}||^2 &\leq 6\beta\eta_\ell^2 \left(\mathbb{E}[f_m(w_{p,m}^{(r,K)})] - f(w_{p,m}^*)\right) + 3\eta_\ell^2 K \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \\
&\quad + 36K^3\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 40K^2\eta_\ell^4\sigma_\ell^2 \\
&\quad + 48K^4\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2
\end{aligned}
\tag{127}
$$

□

Theorem D.16 (Convergence of the Personalized Model for Convex (Strong and General) Cases)**.** *If each client's objective function f_m satisfies Assumptions D.2, D.3, D.4, using the learning rate $\frac{1}{\mu R} \leq \eta_\ell \leq \frac{1}{K\beta^2}$ in Algorithm 2, then the following convergence holds:*
(Strong Convex Case)

$$
\begin{aligned}
\mathbb{E}\left[f_m(w_{p,m}^{(R,0)})\right] - f_m(w_{p,m}^*) &\leq \frac{36\mu^2}{RK^3}\mathbb{E}||w_{p,m}^{(1,K)} - w_{p,m}^*||^2 \exp\left(\frac{1}{K-1} - \eta_\ell \mu K R\right) \\
&+ 12K^2\eta_\ell^2\delta_m^{w_g} + 12K^2\eta_\ell^2 \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2 + 4K\eta_\ell^2\sigma_\ell^2 + \frac{\mathbf{q}_0^2}{2} + 16K^3\eta_\ell^2\mathbf{q}_0^2\delta_m^{w_g} \\
&+ 16K^3\eta_\ell^2\mathbf{q}_0^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2 + \frac{K^2\eta_\ell^2}{2}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)\right)\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)\right) \\
&+ \frac{1+\eta_\ell^2K^2\beta^4}{6}\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(\frac{2}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2 + 2\delta_m^{w_g}\right)
\end{aligned}
$$

(General Convex Case)

$$
\begin{aligned}
\mathbb{E}\left[f_m(w_{p,m}^{(R,0)})\right] - f_m(w_{p,m}^*) &\leq \frac{1}{36\eta_\ell^2K^2R}\left(1 + \frac{1}{K-1}\right)\mathbb{E}||w_{p,m}^{(1,K)} - w_{p,m}^*||^2 \\
&+ \eta_\ell^2(12K^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2)^{1/3} + \eta_\ell^2(12K^2\delta_m^{w_g})^{1/3} + \eta_\ell^2(4K\sigma_\ell^2)^{1/3} + \frac{\mathbf{q}_0^2}{2} + \eta_\ell^2(16K^3\mathbf{q}_0^2\delta_m^{w_g})^{1/3} \\
&+ \eta_\ell^2(16K^3\mathbf{q}_0^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2)^{1/3} + \eta_\ell^2\left(\frac{K^2}{2}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)\right)\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)\right)\right)^{1/3} \\
&+ \eta_\ell^2\left(\frac{K^2}{3}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)\right)\left(\frac{2}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2 + 2\delta_m^{w_g}\right)\right)^{1/3}
\end{aligned}
$$

where $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2$ is bounded as shown in Theorem D.12.

Proof. We restate the update rules of the personalized model in Algorithm 2,

1. For all samples x_m, define $\tilde{w}_{p,m}^{(r,k)}(x_m) \leftarrow \psi_{g,m}^{(r,k)}(x_m)w_{g,m}^{(r,k)}(x_m) + (1 - \psi_{g,m}^{(r,k)}(x_m))w_{\ell,m}^{(r,K)}(x_m)$
2. Train policy parameters $\psi_{g,m}^{(r,k+1)} \leftarrow \psi_{g,m}^{(r,k)} - \eta_\ell\nabla_{\psi_{g,m}^{(r,k)}} f_m(\tilde{w}_{p,m}^{(r,k)}(x_m), y_m)$
3. For all samples x_m, define $w_{p,m}^{(r,k)}(x_m) \leftarrow \psi_{g,m}^{(r,k+1)}(x_m)w_{g,m}^{(r,k)}(x_m) + (1 - \psi_{g,m}^{(r,k+1)}(x_m))w_{\ell,m}^{(r,K)}(x_m)$
4. Train global parameters $w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} - \eta_\ell\nabla_{w_{g,m}^{(r,k-1)}} f_m(w_{p,m}^{(r,k)}(x_m), y_m)$

$$
\begin{aligned}
\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{*}||^2 &= \mathbb{E}||w_{p,m}^{(r+1,K)} - \tilde{w}_{p,m}^{(r+1,0)} + \tilde{w}_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)} + w_{p,m}^{(r,K)} - w_{p,m}^{*}||^2 \qquad (128)\\
&\leq 2K \underbrace{\mathbb{E}||w_{p,m}^{(r+1,K)} - \tilde{w}_{p,m}^{(r+1,0)}||^2}_{\text{Lemma D.13}} + 2K \underbrace{\mathbb{E}||\tilde{w}_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}||^2}_{\text{Lemma D.14}}\\
&\quad + \left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{*}||^2 \qquad (129)
\end{aligned}
$$

And using Assumption D.4,

$$
\begin{aligned}
&\leq 36K^2\eta_\ell^2 \left[f_m(w_{p,m}^{*}) - \mathbb{E}\left[f_m(w_{p,m}^{(r+1,0)})\right]\right] + 216K^4\eta_\ell^4 \mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2 + 126K^3\eta_\ell^4\sigma_\ell^2\\
&\quad + 18K^2\eta_\ell^2 \mathbb{E}\left\|\psi_{g,m}^{(r+1,K)}\right\|^2 + 288K^5\eta_\ell^4 \mathbb{E}||\psi_{g,m}^{(r+1,K)}||^2 \mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2\\
&\quad + 18\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right)\\
&\quad + 6(1 + \eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\mathbb{E}||\nabla f_m(w_g^{(r)})||^2\\
&\quad + \left(1 + \frac{1}{K-1} - \mu\eta_\ell\right)\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{*}||^2 \qquad (130)
\end{aligned}
$$

Rearranging the terms,

$$
\begin{aligned}
36K^2\eta_\ell^2\left[\mathbb{E}\left[f_m(w_{p,m}^{(r+1,0)})\right] - f_m(w_{p,m}^{*})\right] &\leq \left(1 + \frac{1}{K-1} - \mu\eta_\ell\right)\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{*}||^2 - \mathbb{E}||w_{p,m}^{(r,K+1)} - w_{p,m}^{*}||^2\\
&\quad + 216K^4\eta_\ell^4\mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2 + 126K^3\eta_\ell^4\sigma_\ell^2 + 18K^2\eta_\ell^2\mathbb{E}\left\|\psi_{g,m}^{(r+1,K)}\right\|^2\\
&\quad + 288K^5\eta_\ell^4\mathbb{E}||\psi_{g,m}^{(r+1,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2\\
&\quad + 18\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right)\\
&\quad + 6(1 + \eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \qquad (131)
\end{aligned}
$$

$$
\begin{aligned}
\therefore \mathbb{E}\left[f_m(w_{p,m}^{(r+1,0)})\right] - f_m(w_{p,m}^{*}) &\leq \frac{1}{36\eta_\ell^2K^2}\left(1 + \frac{1}{K-1} - \mu\eta_\ell\right)\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{*}||^2 - \frac{1}{36\eta_\ell^2K^2}\mathbb{E}||w_{p,m}^{(r,K+1)} - w_{p,m}^{*}||^2\\
&\quad + 6K^2\eta_\ell^2\mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2 + 4K\eta_\ell^2\sigma_\ell^2 + \frac{1}{2}\mathbb{E}\left\|\psi_{g,m}^{(r+1,K)}\right\|^2 + 8K^3\eta_\ell^2\mathbb{E}||\psi_{g,m}^{(r+1,K)}||^2\mathbb{E}||\nabla f_m(w_g^{(r+1)})||^2\\
&\quad + \frac{K^2\eta_\ell^2}{2}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)\right)\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)\right)\\
&\quad + \frac{1 + \eta_\ell^2K^2\beta^4}{6}\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \qquad (132)
\end{aligned}
$$

For strong convex ($\mu > 0$) case, using the linear convergence rate lemma from [12] (Lemma 1) and Definition D.5,

$$
\begin{aligned}
\mathbb{E}\left[f_m(w_{p,m}^{(R,0)})\right] - f_m(w_{p,m}^{*}) &\leq \frac{36\mu^2}{RK^3}\mathbb{E}||w_{p,m}^{(1,K)} - w_{p,m}^{*}||^2 \exp\left(\frac{1}{K-1} - \eta_\ell\mu KR\right)\\
&\quad + 12K^2\eta_\ell^2\delta_m^{w_g} + 12K^2\eta_\ell^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2 + 4K\eta_\ell^2\sigma_\ell^2 + \frac{\mathbf{q}_0^2}{2} + 16K^3\eta_\ell^2\mathbf{q}_0^2\delta_m^{w_g}\\
&\quad + 16K^3\eta_\ell^2\mathbf{q}_0^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2 + \frac{K^2\eta_\ell^2}{2}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)\right)\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)\right)\\
&\quad + \frac{1 + \eta_\ell^2K^2\beta^4}{6}\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{\psi} + \frac{\delta^{\psi}}{M}\right)K^2\eta_\ell^2\right)\left(\frac{2}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2 + 2\delta_m^{w_g}\right) \qquad (133)
\end{aligned}
$$

For general convex ($\mu = 0$) case, using the sublinear convergence rate lemma from [12] (Lemma 2), and conditioning on $\eta_\ell^2 K^2 \beta^4 \le 1 \implies \eta_\ell \le \frac{1}{K\beta^2}$,

$$\mathbb{E}\left[f_m(w_{p,m}^{(R,0)})\right] - f_m(w_{p,m}^*) \le \frac{1}{36\eta_\ell^2 K^2 R}\left(1 + \frac{1}{K-1}\right)\mathbb{E}||w_{p,m}^{(1,K)} - w_{p,m}^*||^2$$
$$+ \eta_\ell^2 (12K^2 \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2)^{1/3} + \eta_\ell^2 (12K^2 \delta_m^{w_g})^{1/3} + \eta_\ell^2 (4K\sigma_\ell^2)^{1/3} + \frac{\mathbf{q}_0^2}{2} + \eta_\ell^2 (16K^3 \mathbf{q}_0^2 \delta_m^{w_g})^{1/3}$$
$$+ \eta_\ell^2 (16K^3 \mathbf{q}_0^2 \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r+1)})||^2)^{1/3} + \eta_\ell^2 \left(\frac{K^2}{2}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)\right)\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)\right)\right)^{1/3}$$
$$+ \eta_\ell^2 \left(\frac{K^2}{3}\left(\frac{\sigma_\ell^2}{K} + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)\right)(\frac{2}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2 + 2\delta_m^{w_g})\right)^{1/3} \tag{134}$$

□

D.6 Convergence Proof for the Personalized Model: Non-convex Case

Lemma D.17 (One round progress of the local model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.3, D.4, in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{\ell,m}^{(r+1,K)} - w_{\ell,m}^{(r,K)}||^2 \le (1 - 2\eta_\ell K\beta^2 + \eta_\ell^2 K^2 \beta^4)\eta_\ell^2 K^2 \left(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2\right)$$

Proof.

$$\mathbb{E}||w_{\ell,m}^{(r+1,K)} - w_{\ell,m}^{(r,K)}||^2 = \mathbb{E}||w_g^{(r+1)} - \eta_\ell \sum_{k=1}^{K}\nabla f_m(w_g^{(r+1)}) - w_g^{(r)} + \eta_\ell \sum_{k=1}^{K}\nabla f_m(w_g^{(r)})||^2 \tag{135}$$
$$= \mathbb{E}||w_g^{(r+1)} - w_g^{(r)} - \eta_\ell \sum_{k=1}^{K}\left[\nabla f_m(w_g^{(r+1)}) - \nabla f_m(w_g^{(r)})\right]||^2 \tag{136}$$
$$\le \mathbb{E}||w_g^{(r+1)} - w_g^{(r)} - \eta_\ell K\beta^2 \left(w_g^{(r+1)} - w_g^{(r)}\right)||^2 \tag{137}$$
$$= \mathbb{E}||(1 - \eta_\ell K\beta^2)\left(w_g^{(r+1)} - w_g^{(r)}\right)||^2 \tag{138}$$
$$\le (1 - \eta_\ell K\beta^2)^2 \mathbb{E}\left\|\frac{1}{M}\sum_{c\in[M]} w_{g,m}^{(r,K)} - w_g^{(r)}\right\|^2 \tag{139}$$
$$= (1 - \eta_\ell K\beta^2)^2 \mathbb{E}\left\|\frac{1}{M}\sum_{c\in[M]}(w_g^{(r)} - \eta_\ell \sum_{k=1}^{K}\nabla f_m(w_g^{(r)})) - w_g^{(r)}\right\|^2 \tag{140}$$
$$= (1 - \eta_\ell K\beta^2)^2 \mathbb{E}\left\| - \frac{\eta_\ell}{M}\sum_{c\in[M]}\sum_{k=1}^{K}\nabla f_m(w_g^{(r)})\right\|^2 \tag{141}$$
$$\le (1 - \eta_\ell K\beta^2)^2 \eta_\ell^2 \mathbb{E}\left\|\frac{K}{M}\sum_{c\in[M]}\nabla f_m(w_g^{(r)})\right\|^2 \tag{142}$$
$$\le (1 - \eta_\ell K\beta^2)^2 \eta_\ell^2 K^2 \left(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2\right) \tag{143}$$
$$= (1 - 2\eta_\ell K\beta^2 + \eta_\ell^2 K^2 \beta^4)\eta_\ell^2 K^2 \left(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2\right) \tag{144}$$

The last inequality follows from Assumption D.4. □

We proceed with a lemma which binds the deviation of the personalized model w_p of an arbitrary client m over one round, i.e., $w_{p,m}^{(r+1)}$ and $w_{p,m}^{(r)}$, for non-convex case.

Lemma D.18 (Local progress of personalized model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.3, D.4, and $\eta_\ell \le \frac{1}{K\sqrt{12\beta(K-1)}}$, in Algorithm 2, the following is satisfied:*

$$\mathbb{E}||w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,0)}||^2 \le 18K^5\eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 9K^4\eta_\ell^2\sigma_\ell^2 + 36K^3\eta_\ell^2\sigma_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2$$
$$+ 24K^4\eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2$$

Proof. We start with using the update rule stated for the personalized model at the beginning of Theorem D.16,

$$\mathbb{E}||w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,0)}||^2 = \mathbb{E}||\psi_{g,m}^{(r,k+1)} w_{g,m}^{(r,k+1)} + (1 - \psi_{g,m}^{(r,k+1)}) w_{\ell,m}^{(r,K)} - w_{p,m}^{(r,0)}||^2 \tag{145}$$

Expanding by one iterate,

$$\begin{aligned}
&= \mathbb{E}|| \left(\psi_{g,m}^{(r,k)} - \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\right) \left(w_{g,m}^{(r,k)} - \eta_\ell \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\right) \\
&\quad + \left(1 - \psi_{g,m}^{(r,k)} + \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\right) w_{\ell,m}^{(r,K)} - w_{p,m}^{(r,0)}||^2 &(146) \\
&= \mathbb{E}||\psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - w_{g,m}^{(r,k)} \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) - \psi_{g,m}^{(r,k)} \eta_\ell \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) \\
&\quad + \eta_\ell^2 \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) + (1 - \psi_{g,m}^{(r,k)}) w_{\ell,m}^{(r,K)} \\
&\quad + w_{\ell,m}^{(r,K)} \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) - w_{p,m}^{(r,0)}||^2 &(147) \\
&= \mathbb{E}||w_{p,m}^{(r,k)} - w_{p,m}^{(r,0)} + (w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}) \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)}) \\
&\quad \left(-\psi_{g,m}^{(r,k)} + \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})\right) \eta_\ell \nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})||^2 &(148) \\
&\leq 3\mathbb{E}||w_{p,m}^{(r,k)} - w_{p,m}^{(r,0)}||^2 + 3\mathbb{E}||w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}||^2 + 3\eta_\ell^2 \mathbb{E}||\nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})||^2 &(149)
\end{aligned}$$

The inequality was derived from the fact that $\mathbb{E}|| - \eta_\ell \nabla_{\psi_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})||^2 = \mathbb{E}||\psi_{g,m}^{(r,k+1)} - \psi_{g,m}^{(r,k)}||^2 \leq 1$. Unrolling the recursion across $r \in [R]$, then using Lemmas D.6 and D.10 and Assumption D.4,

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{(r,0)}||^2 &\leq \sum_{k=1}^{K} \left(\left(1 + \frac{1}{K-1}\right) \mathbb{E}||w_{\ell,m}^{(r,K)} - w_{g,m}^{(r,k)}||^2 + K\eta_\ell^2 \mathbb{E}||\nabla_{w_{g,m}^{(r,k)}} f_m(w_{p,m}^{(r,k)})||^2 \right) &(150) \\
&\leq \Bigg(6K^4 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K^3 \eta_\ell^2 \sigma_\ell^2 + 12K^2 \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||1 - \psi_{g,m}||^2 \\
&\quad + 4\left(1 + \frac{1}{K-1}\right) K^3 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||\psi_{g,m}||^2 \Bigg) \sum_{k=1}^{K} \left(1 + \frac{1}{K-1} + 12K^2 \eta_\ell^2 \beta\right)^k &(151)
\end{aligned}$$

Assuming $\frac{1}{K-1} \geq 12K^2 \eta_\ell^2 \beta \implies \eta_\ell \leq \frac{1}{K\sqrt{12(K-1)\beta}}$,

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r,K)} - w_{p,m}^{(r,0)}||^2 &\leq \Bigg(6K^4 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K^3 \eta_\ell^2 \sigma_\ell^2 + 12K^2 \eta_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \\
&\quad + 4\left(1 + \frac{1}{K-1}\right) K^3 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 \sigma_\ell^2 \Bigg) 3K &(152) \\
&= 18K^5 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 9K^4 \eta_\ell^2 \sigma_\ell^2 + 36K^3 \eta_\ell^2 \sigma_\ell^2 \mathbb{E}||1 - \psi_{g,m}^{(r,k)}||^2 \\
&\quad + 24K^4 \eta_\ell^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||\psi_{g,m}^{(r,k)}||^2 &(153)
\end{aligned}$$

□

Lemma D.19 (One round progress of personalized model)**.** *If m^{th} client's objective function f_m satisfies Assumptions D.3, D.4, in Algorithm 2, the following is satisfied:*

$$\begin{aligned}
\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 &\leq 72(1 + \eta_\ell^2) K^3 \eta_\ell^2 \left(5K(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\
&\quad + 36 \left(K\sigma_\ell^2 \eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right) K^2 \eta_\ell^2\right) \left(K\sigma_\ell^2 \eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right) K^2 \eta_\ell^2\right) \\
&\quad + 12(1 + \eta_\ell^2 K^2 \beta^4) \eta_\ell^2 K^2 \left(K\sigma_\ell^2 \eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right) K^2 \eta_\ell^2\right) \left(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2\right)
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}\|w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}\|^2 &= \mathbb{E}\|w_{p,m}^{(r+1,K)} - w_{p,m}^{(r+1,0)} + w_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}\|^2 &(154) \\
&\leq 2\mathbb{E}\|w_{p,m}^{(r+1,K)} - w_{p,m}^{(r+1,0)}\|^2 + 2\mathbb{E}\|w_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}\|^2 &(155)
\end{aligned}$$

Using the Lemmas D.18 and D.14, we proceed as

$$
\begin{aligned}
&\leq 72(1+\eta_\ell^2)K^3\eta_\ell^2 \left(5K(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\
&\quad + 36\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right) \\
&\quad + 12(1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2\right) \qquad (156)
\end{aligned}
$$

□

Theorem D.20 (Convergence of the Personalized Model for Non-convex Cases)**.** *If each client's objective function f_m satisfies Assumptions D.2, D.3, D.4 using the learning rate $\eta_\ell \leq \frac{1}{K\sqrt{12\beta}}$ in Algorithm 2, then the following convergence holds:*

$$
\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 &\leq \frac{2}{R}\left(\mathbb{E}\left[f_m(w_{p,m}^{(1,K)})\right] - \mathbb{E}\left[f_m(w_{p,m}^{(R,K)})\right]\right) \\
&\quad + 6(1+\eta_\ell^2)K\left(5K(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\
&\quad + 3K\eta_\ell^2\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(\sigma_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K\right) \\
&\quad + (1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)
\end{aligned}
$$

Proof. According to the update rule of Equation 10 and β-smoothness of f_m, we have,

$$
f_m(w_{p,m}^{(r+1,K)}) \leq f_m(w_{p,m}^{(r,K)}) + \left\langle \nabla f_m(w_{p,m}^{(r,K)}), w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}\right\rangle + \frac{\beta}{2}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 \quad (157)
$$

Taking expectation on both sides,

$$
\mathbb{E}\left[f_m(w_{p,m}^{(r+1,K)})\right] \leq \mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] + \mathbb{E}\left\langle \nabla f_m(w_{p,m}^{(r,K)}), w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}\right\rangle + \frac{\beta}{2}\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 \quad (158)
$$

Using $\langle a,b\rangle = \frac{1}{2}||a||^2 + \frac{1}{2}||b||^2 - \frac{1}{2}||a-b||^2$

$$
\begin{aligned}
\mathbb{E}\left[f_m(w_{p,m}^{(r+1,K)})\right] &\leq \mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] + \frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 + \left(\frac{\beta+1}{2}\right)\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 \\
&\quad - \frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)}) - (w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)})||^2 && (159) \\
&\leq \mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] + \frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 + \left(\frac{\beta+1}{2}\right)\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 \\
&\quad - \mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 - \mathbb{E}||(w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)})||^2 && (160) \\
&\leq \mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] - \frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 + \left(\frac{\beta-1}{2}\right)\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2 \\
& && (161) \\
& && (162)
\end{aligned}
$$

Rearranging the terms to put $\frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2$ at LHS,

$$\frac{1}{2}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 \leq \mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] - \mathbb{E}\left[f_m(w_{p,m}^{(r+1,K)})\right] + \left(\frac{\beta-1}{2}\right)\underbrace{\mathbb{E}||w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,K)}||^2}_{\text{Lemma D.19}} \tag{163}$$

$$\begin{aligned}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 \leq\ & 2\left(\mathbb{E}\left[f_m(w_{p,m}^{(r,K)})\right] - \mathbb{E}\left[f_m(w_{p,m}^{(r+1,K)})\right]\right) \\ & + 72\beta(1+\eta_\ell^2)K^3\eta_\ell^2\left(5K(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\ & + 36\beta\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K^2\eta_\ell^2\right) \\ & + 12\beta(1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K^2\left(K\sigma_\ell^2\eta_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K^2\eta_\ell^2\right)\left(G^2 + B^2\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)\end{aligned} \tag{164}$$

Taking an average over all the rounds $r \in [R]$,

$$\begin{aligned}\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 \leq\ & \frac{2}{R}\left(\mathbb{E}\left[f_m(w_{p,m}^{(1,K)})\right] - \mathbb{E}\left[f_m(w_{p,m}^{(R,K)})\right]\right) \\ & + 72\beta(1+\eta_\ell^2)K^3\eta_\ell^2\left(5K(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\ & + 36\beta K^2\eta_\ell^4\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(\sigma_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K\right) \\ & + 12\beta(1+\eta_\ell^2K^2\beta^4)\eta_\ell^4K^3\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)\end{aligned} \tag{165}$$

Assuming $12K^2\eta_\ell^2\beta \leq 1 \leq 1 \implies \eta_\ell \leq \frac{1}{K\sqrt{12\beta}}$,

$$\begin{aligned}\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2 \leq\ & \frac{2}{R}\left(\mathbb{E}\left[f_m(w_{p,m}^{(1,K)})\right] - \mathbb{E}\left[f_m(w_{p,m}^{(R,K)})\right]\right) \\ & + 6(1+\eta_\ell^2)K\left(5K(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_\ell^2\right) \\ & + 3K\eta_\ell^2\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(\sigma_\ell^2 + \left(\delta_m^{w_g} + \frac{\delta^{w_g}}{M}\right)K\right) \\ & + (1+\eta_\ell^2K^2\beta^4)\eta_\ell^2K\left(\sigma_\ell^2 + \left(\delta_m^\psi + \frac{\delta^\psi}{M}\right)K\right)\left(G^2 + B^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2\right)\end{aligned} \tag{166}$$

Plugging in Theorem D.12 to get bounds on $\sum_{r=1}^{R}\mathbb{E}||\nabla F(w_g^{(r)})||^2$ would get us bounds on $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}||\nabla f_m(w_{p,m}^{(r,K)})||^2$. $\square$