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 and membership inference 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) can reveal properties of personalization that go beyond improved accuracy.

B Datasets and Hyperparameters

Stackoverflow The Stackoverflow dataset 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 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 [13] [16] [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 λ = 0.5. For APFL, mixture hyperparameter α is set to 0.25. Ditto has regularization hyperparameter λ = 0.1. There are 2 clusters by default for HypCluster. Flow and its variants were tested on the following choices of regularizing hyperparameters γ ∈ {1e-1, 1e-2, 1e-3, 1e-4}, where 1e-3 gave the best personalized accuracy.

Shakespeare The Shakespeare dataset 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 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 γ ∈ {1e-1, 1e-2, 1e-3, 1e-4} similar to Stackoverflow.

EMNIST The EMNIST dataset 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 λ = 0.4. For APFL, we have the default mixture parameter as α = 0.25. Ditto has regularization hyperparameter as λ = 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 30,000 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 $\alpha \to 0$. In this context, heterogeneity refers to the dissimilarity of the dataset instances sampled from a distribution. We conducted experiments on clients with $\alpha$ 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 $\lambda$ 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 $\alpha$. Ditto has 1e-2 as regularization parameter $\lambda$. 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 $\gamma$.

### 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

$$ Acc_g = \frac{1}{M} \sum_{m \in [M]} \sum_{(x,y) \in S_m} \mathbb{I}\{y = w_g(x)\} / S_m. $$

(6)

Personalized accuracy is formulated as

$$ Acc_p = \frac{1}{M} \sum_{m \in [M]} \sum_{(x,y) \in S_{m,p}} \mathbb{I}\{y = w_{p,m}(x)\} / S_m. $$

(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% ± 0.27%) matches the personalized accuracy of the highest performing baseline, PARTIALFED (40.18% ± 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% ± 0.32%) reaches close to the best performing baseline, PARTIALFED (34.79% ± 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Stackoverflow</th>
<th>Shakespeare</th>
<th>EMNIST</th>
<th>CIFAR10 (0.1)</th>
<th>CIFAR100 (0.1)</th>
<th>CIFAR10 (0.6)</th>
<th>CIFAR100 (0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
</tr>
<tr>
<td>LOCAL</td>
<td>-</td>
<td>15.93%</td>
<td>18.70%</td>
<td>28.18%</td>
<td>49.78%</td>
<td>36.19%</td>
<td>62.74%</td>
</tr>
<tr>
<td>FEDAVG</td>
<td>23.15%</td>
<td>-</td>
<td>50.00%</td>
<td>85.10%</td>
<td>60.98%</td>
<td>28.11%</td>
<td>67.50%</td>
</tr>
<tr>
<td>FEDAVGFT</td>
<td>23.83%</td>
<td>24.41%</td>
<td>52.12%</td>
<td>53.68%</td>
<td>89.57%</td>
<td>90.14%</td>
<td>61.23%</td>
</tr>
<tr>
<td>LITERA</td>
<td>23.16%</td>
<td>24.49%</td>
<td>51.87%</td>
<td>51.10%</td>
<td>85.20%</td>
<td>88.28%</td>
<td>59.62%</td>
</tr>
<tr>
<td>PARTIALFED</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>62.57%</td>
<td>73.20%</td>
<td>34.79%</td>
</tr>
<tr>
<td>APFL</td>
<td>22.96%</td>
<td>25.70%</td>
<td>52.38%</td>
<td>53.64%</td>
<td>88.40%</td>
<td>89.44%</td>
<td>62.87%</td>
</tr>
<tr>
<td>DITTO</td>
<td>22.59%</td>
<td>24.36%</td>
<td>52.44%</td>
<td>53.95%</td>
<td>89.08%</td>
<td>91.30%</td>
<td>62.06%</td>
</tr>
<tr>
<td>FEDREP</td>
<td>19.92%</td>
<td>21.04%</td>
<td>46.71%</td>
<td>50.09%</td>
<td>89.95%</td>
<td>89.77%</td>
<td>64.85%</td>
</tr>
<tr>
<td>LGFEDAVG</td>
<td>22.61%</td>
<td>24.03%</td>
<td>51.08%</td>
<td>51.43%</td>
<td>87.43%</td>
<td>91.70%</td>
<td>56.63%</td>
</tr>
<tr>
<td>HYPCLUSTER</td>
<td>23.75%</td>
<td>22.43%</td>
<td>51.92%</td>
<td>52.74%</td>
<td>89.47%</td>
<td>90.49%</td>
<td>63.64%</td>
</tr>
<tr>
<td>Flow (Ours)</td>
<td>26.64%</td>
<td>29.49%</td>
<td>55.99%</td>
<td>56.20%</td>
<td>90.88%</td>
<td>94.18%</td>
<td>66.26%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SO NWP</th>
<th>Shakespeare</th>
<th>EMNIST</th>
<th>CIFAR10 (0.1)</th>
<th>CIFAR100 (0.1)</th>
<th>CIFAR10 (0.6)</th>
<th>CIFAR100 (0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
</tr>
<tr>
<td>LOCAL</td>
<td>-</td>
<td>0.28%</td>
<td>-</td>
<td>0.06%</td>
<td>-</td>
<td>1.14%</td>
<td>-</td>
</tr>
<tr>
<td>FEDAVG</td>
<td>0.07%</td>
<td>-</td>
<td>0.39%</td>
<td>-</td>
<td>1.32%</td>
<td>-</td>
<td>1.12%</td>
</tr>
<tr>
<td>FEDAVGFT</td>
<td>0.09%</td>
<td>0.26%</td>
<td>0.51%</td>
<td>0.59%</td>
<td>1.16%</td>
<td>1.21%</td>
<td>0.99%</td>
</tr>
<tr>
<td>KNNPER</td>
<td>0.16%</td>
<td>0.24%</td>
<td>0.36%</td>
<td>0.41%</td>
<td>0.95%</td>
<td>1.02%</td>
<td>1.41%</td>
</tr>
<tr>
<td>PARTIALFED</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.36%</td>
<td>1.39%</td>
<td>0.29%</td>
</tr>
<tr>
<td>APFL</td>
<td>0.19%</td>
<td>0.20%</td>
<td>0.41%</td>
<td>0.53%</td>
<td>1.41%</td>
<td>1.50%</td>
<td>1.24%</td>
</tr>
<tr>
<td>DITTO</td>
<td>0.12%</td>
<td>0.15%</td>
<td>0.49%</td>
<td>0.56%</td>
<td>1.12%</td>
<td>1.22%</td>
<td>1.35%</td>
</tr>
<tr>
<td>FEDREP</td>
<td>0.15%</td>
<td>0.29%</td>
<td>0.50%</td>
<td>0.65%</td>
<td>0.89%</td>
<td>0.94%</td>
<td>0.95%</td>
</tr>
<tr>
<td>LGFEDAVG</td>
<td>0.08%</td>
<td>0.16%</td>
<td>0.32%</td>
<td>0.56%</td>
<td>1.10%</td>
<td>1.17%</td>
<td>1.21%</td>
</tr>
<tr>
<td>HYPCLUSTER</td>
<td>0.20%</td>
<td>0.19%</td>
<td>0.56%</td>
<td>0.73%</td>
<td>0.90%</td>
<td>1.13%</td>
<td>1.43%</td>
</tr>
<tr>
<td>Flow</td>
<td>0.23%</td>
<td>0.28%</td>
<td>0.40%</td>
<td>0.49%</td>
<td>1.16%</td>
<td>1.21%</td>
<td>1.23%</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Stackoverflow</th>
<th>Shakespeare</th>
<th>EMNIST</th>
<th>CIFAR10 (0.1)</th>
<th>CIFAR100 (0.1)</th>
<th>CIFAR10 (0.6)</th>
<th>CIFAR100 (0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
<td>Acc_p</td>
<td>Acc_g</td>
</tr>
<tr>
<td>LOCAL</td>
<td>79.26%</td>
<td>81.48%</td>
<td>79.00%</td>
<td>97.18%</td>
<td>91.74%</td>
<td>99.33%</td>
<td>88.54%</td>
</tr>
<tr>
<td>FEDAVGFT</td>
<td>82.73%</td>
<td>89.97%</td>
<td>68.87%</td>
<td>90.00%</td>
<td>94.71%</td>
<td>90.00%</td>
<td>96.37%</td>
</tr>
<tr>
<td>KNNPER</td>
<td>69.66%</td>
<td>93.39%</td>
<td>79.22%</td>
<td>87.48%</td>
<td>86.18%</td>
<td>90.63%</td>
<td>92.03%</td>
</tr>
<tr>
<td>PARTIALFED</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>APFL</td>
<td>74.59%</td>
<td>79.26%</td>
<td>73.74%</td>
<td>90.52%</td>
<td>91.45%</td>
<td>89.61%</td>
<td>97.45%</td>
</tr>
<tr>
<td>DITTO</td>
<td>91.53%</td>
<td>82.20%</td>
<td>79.78%</td>
<td>92.30%</td>
<td>78.81%</td>
<td>84.64%</td>
<td>99.54%</td>
</tr>
<tr>
<td>FEDREP</td>
<td>83.47%</td>
<td>66.16%</td>
<td>88.43%</td>
<td>88.41%</td>
<td>86.39%</td>
<td>89.59%</td>
<td>91.73%</td>
</tr>
<tr>
<td>LGFEDAVG</td>
<td>80.46%</td>
<td>80.76%</td>
<td>74.84%</td>
<td>95.11%</td>
<td>93.70%</td>
<td>98.18%</td>
<td>99.73%</td>
</tr>
<tr>
<td>HYPCLUSTER</td>
<td>92.74%</td>
<td>96.70%</td>
<td>89.77%</td>
<td>98.33%</td>
<td>97.29%</td>
<td>99.62%</td>
<td>99.78%</td>
</tr>
</tbody>
</table>

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

![Graphs](image1.png)

**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
Figure 6: Behavior of $\psi_g$ 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% (CIFAR10 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.
Table 7: Test (personalized) accuracy of two of the Flow variants: (a) Soft Policy variant where the probability $q$ is continuous in the range of $[0, 1]$ during inference. (b) Hard Policy variant where the probability $q$ is discrete over the set $\{0, 1\}$ during inference.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Stackoverflow</th>
<th>Shakespeare</th>
<th>EMNIST</th>
<th>CIFAR10 (0.1)</th>
<th>CIFAR100 (0.1)</th>
<th>CIFAR10 (0.6)</th>
<th>CIFAR100 (0.6)</th>
<th>CIFAR100 (0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Policy</td>
<td>29.37% ± 0.22%</td>
<td>57.01% ± 0.51%</td>
<td>94.97% ± 1.06%</td>
<td>77.24% ± 1.30%</td>
<td>42.78% ± 0.36%</td>
<td>77.02% ± 0.96%</td>
<td>39.74% ± 0.13%</td>
<td></td>
</tr>
<tr>
<td>Hard Policy</td>
<td>29.49% ± 0.28%</td>
<td>56.20% ± 0.49%</td>
<td>94.18% ± 1.21%</td>
<td>76.47% ± 1.25%</td>
<td>42.42% ± 0.36%</td>
<td>77.11% ± 0.86%</td>
<td>40.08% ± 0.27%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: Generalized Accuracy of the Ablation Study on the Regularization Term used in the Policy Learning Objective.
Figure 8: Personalized Accuracy of the Ablation Study on the Regularization Term used in the Policy Learning Objective.
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, \( M \): Set of available clients, \( p \): Client sampling rate, \( K \): Total local epoch count, \( k \in [K] \):

- Epoch index, \( \eta \): 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, \( w_{g,m}^{(r)} \): \( m \)-th client’s local 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, \( D_m \): Data distribution of \( m \)-th client, \( \mathbf{S}_m \): Dataset of \( m \)-th client, \( \zeta_m,\ell \): Dataset used to train \( w_\ell \), \( \xi_m \): Dataset used to train \( w_g \).

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

\begin{enumerate}
\item Server randomly initializes \( w_g^{(1)} \).
\item \( \text{for } r \in [R] \text{ round do} \)
\item Sample \( M \) clients from \( M \) with the rate of \( p \).
\item Send \( w_g^{(r)} \), \( \psi_g^{(r)} \) to all the clients.
\item \( \text{for } m \in [M] \text{ in parallel do} \)
\item \( w_{g,m}^{(r,0)} \leftarrow w_g^{(r)} \), \( \psi_{g,m}^{(r,0)} \leftarrow \psi_g^{(r)} \), \( w_{g,m}^{(r,0)} \leftarrow w_{g,m}^{(r,0)} \).
\item \( \zeta_m,\ell,\xi_m \leftarrow \mathbf{S}_m /\star \text{ Creating two mutually exclusive datasets} \)
\item \( \text{for } k \in [K_1] \text{ epochs do} \)
\item \( w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} \cdot \eta \nabla f_m (w_{g,m}^{(r,k-1)}; \zeta_m,\ell) \) \label{eq:local-update}
\item \( \text{for } k \in [K_2] \text{ epochs do} \)
\item \( \text{for } x_m, y_m \sim \zeta_m,\ell \text{ do} \)
\item \( \hat{w}_{p,m}^{(r,k-1)} (x_m) \leftarrow \psi_{g,m}^{(r,k-1)} (x_m) \cdot w_{g,m}^{(r,k-1)} (x_m) + (1 - \psi_{g,m}^{(r,k-1)} (x_m)) \cdot w_{g,m}^{(r,k)} (x_m) \) \label{eq:w_update}
\item \( \psi_{g,m}^{(r,k-1)} \leftarrow \eta \nabla f_m (w_{g,m}^{(r,k-1)}; \zeta_m,\ell) \) \label{eq:psi-update}
\item \( \text{for } x_m, y_m \sim \zeta_m,\ell \text{ do} \)
\item \( \hat{w}_{p,m}^{(r,k-1)} (x_m) \leftarrow \psi_{g,m}^{(r,k-1)} (x_m) \cdot w_{g,m}^{(r,k-1)} (x_m) + (1 - \psi_{g,m}^{(r,k-1)} (x_m)) \cdot w_{g,m}^{(r,k)} (x_m) \) \label{eq:w_update2}
\item \( w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} \cdot \eta \nabla f_m (w_{g,m}^{(r,k-1)}; \zeta_m,\ell) \) \label{eq:w_global_update}
\item \( \text{end} \)
\item \( w_{g,m}^{(r+1)} \leftarrow \frac{1}{n_m} \sum_{m \in [M]} n_m \cdot w_{g,m}^{(r,k)} \)
\item \( \psi_g^{(r+1)} \leftarrow \frac{1}{n} \sum_{m \in [M]} n_m \cdot \psi_{g,m}^{(r,k)} \)
\item \( \text{end} \)
\end{enumerate}
Note that $w_{p,m}$ is a combination of outputs of $w_{g,m}$ (Global parameters) and $w_{l,m}$ (Local parameters) on each layer. For tractability of analysis, we will assume that the combination is only after the last layer. Hence,

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

The local model update rule is,

$$w_{l,m}^{(r,k)} \leftarrow w_{l,m}^{(r,k-1)} - \eta_l \nabla_f m(w_{l,m}^{(r,k-1)}(x_m), y_m)$$

where $w_{l,m}^{(r,0)} = w_{l,m}^{(r)}$. Indices $r \in [R]$ and $k \in [K]$ are the global round and the local epoch indices.

The policy update rule is,

$$\psi_{g,m}^{(r,k)} \leftarrow \psi_{g,m}^{(r,k-1)} - \eta_p \nabla_f m(w_{p,m}^{(r,k-1)}(x_m), y_m).$$

The global model update rule is,

$$w_{g,m}^{(r,k)} \leftarrow w_{g,m}^{(r,k-1)} - \eta_g \nabla_f m(w_{g,m}^{(r,k-1)}(x_m), y_m).$$

We list out all the optimization problems relevant to Flow:

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

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

- **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**

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

- **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**

  $w_{p,m}^* \in \mathcal{H}$ 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**

  $$F(w_g) = \frac{1}{nM} \sum_{m \in [M]} n_m \mathbb{E}_{(x_m, y_m) \sim S_m}[f_m(w_g(x_m), y_m)]$$

  where $n = |\mathcal{S}| = | \bigcup_{m \in [M]} S_m |$

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

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

- **Local minimizer of global empirical risk**

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

We also use the following assumptions similar to [19][12][6]:

**Assumption D.1** (Strong Convexity). $f_m$ is $\mu$-convex for $\mu \geq 0$. Hence,

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

We also generalize our convergence analysis for $\mu = 0$, general convex cases.

**Assumption D.2** (Smoothness). The gradient of $f_m$ is $\beta$-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 $\sigma_f^2$.

$$E_{(x_m, y_m \sim D_m)}[|h_m(w) - \nabla f_m(w(x_m), y_m)|^2] \leq \sigma_f^2, \forall m \in [M] \text{ and } w \in \mathcal{H}.$$ 

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

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

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

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

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

$$w_{g,m}^{(r,k+1)} = w_{g,m}^{(r,k)} - \eta_f h_m(w_{g,m}^{(r,k)}).$$

(8)

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

$$w_{g,m}^{(r,K)} = w_{g,m}^{(r,0)} - \eta_f \sum_{k=1}^{K} h_m(w_{g,m}^{(r,k-1)})$$

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

(9)

(10)

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

The global model update is,

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

(11)

We first start with a lemma which binds the deviation between the local model $w_{l,m}^{(r,K)}$ 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 $\eta_f \leq \frac{1}{\beta \sqrt{2}K(1-\epsilon)}$ in Algorithm 2, the following is satisfied:

$$E[|w_{l,m}^{(r,K)} - w_{l,m}^{(r,0)}|^2] \leq 6K^2 \eta_f^2 E[\|\nabla f_m(w_g^{(r)})\|^2] + 3K^2 \eta_f^2 \sigma^2_f$$

Proof.

$$E[|w_{l,m}^{(r,K)} - w_{l,m}^{(r,0)}|^2] = E[|w_{l,m}^{(r,K)} - \eta_f \nabla f_m(w_{l,m}^{(r,K-1)}) - w_{l,m}^{(r,0)}|^2$$

$$\leq \left(1 + \frac{1}{K-1}\right)E[|w_{l,m}^{(r,K-1)} - w_{l,m}^{(r,0)}|^2 + K \eta_f^2 E[\|\nabla f_m(w_{l,m}^{(r,K-1)})\|^2 + \eta_f^2 \sigma^2_f]$$

(12)

(13)
Here we have used triangle inequality and variance separation.

\[
\begin{align*}
\leq & \left(1 + \frac{1}{K-1}\right) \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 + K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,K-1)}) \right|^2 + \eta_t^2 \sigma_t^2 \\
\leq & \left(1 + \frac{1}{K-1}\right) \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 + \eta_t^2 \sigma_t^2 \\
& + K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,K-1)}) - \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 \\
\leq & \left(1 + \frac{1}{K-1}\right) \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 + 2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,K-1)}) - \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 \\
& + 2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + \eta_t^2 \sigma_t^2 \\
\leq & \left(1 + \frac{1}{K-1}\right) \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 + 2K \eta_t^2 \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 \\
& + 2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + \eta_t^2 \sigma_t^2 \\
\end{align*}
\]

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

\[
\mathbb{E} \left| w_{t,m}^{(r,K)} - w_{t,m}^{(r,0)} \right|^2 \leq \left(1 + \frac{2}{K-1}\right) \mathbb{E} \left| w_{t,m}^{(r,K-1)} - w_{t,m}^{(r,0)} \right|^2 \\
+ 2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + \eta_t^2 \sigma_t^2
\]

Unrolling the above recursion,

\[
\begin{align*}
\mathbb{E} \left| w_{t,m}^{(r,K)} - w_{t,m}^{(r,0)} \right|^2 & \leq \sum_{i=1}^{K} \left(2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + \eta_t^2 \sigma_t^2 \right) \left(1 + \frac{2}{K-1}\right)^{i} \\
& \leq 3K \left(2K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + \eta_t^2 \sigma_t^2 \right) \\
& = 6K^2 \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{t,m}^{(r,0)}) \right|^2 + 3K \eta_t^2 \sigma_t^2
\end{align*}
\]

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 \( w_{g}^{(r,0)} \) 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 \( \eta_t \leq \frac{1}{\beta \sqrt{2K}} \) in Algorithm 2, the following is satisfied:

\[
\mathbb{E} \left| w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)} \right|^2 \leq 8K^3 \eta_t^2 \mathbb{E} \left| \psi_{g,m}^{(r,k)} \right|^2 \mathbb{E} \left| \nabla f_m(w_{g}^{(r)}) \right|^2 + 4K \eta_t^2 \sigma_t^2
\]

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

\[
\mathbb{E} \left| w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)} \right|^2 = \mathbb{E} \left| w_{g,m}^{(r,k-1)} - \eta_t \nabla_{w_{g,m}} f_m(w_{g,m}^{(r,k-1)}) - w_{g,m}^{(r,0)} \right|^2
\]

Using triangle inequality and separation of variance, we get,

\[
\begin{align*}
\leq & \left(1 + \frac{1}{k-1}\right) \mathbb{E} \left| w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)} \right|^2 + K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{g,m}^{(r,k-1)}) \right|^2 + \eta_t^2 \sigma_t^2 \\
\leq & \left(1 + \frac{1}{k-1}\right) \mathbb{E} \left| w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)} \right|^2 + \eta_t^2 \sigma_t^2 \\
& + K \eta_t^2 \mathbb{E} \left| \nabla f_m(w_{g,m}^{(r,k-1)}) - \nabla f_m(w_{g,m}^{(r,0)}) \right|^2 \\
\leq & \left(1 + \frac{1}{k-1}\right) \mathbb{E} \left| w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)} \right|^2 + \eta_t^2 \sigma_t^2 \\
& + 2K \eta_t^2 \mathbb{E} \left| \psi_{g,m}^{(r,k)} \right|^2 \mathbb{E} \left| \nabla f_m(w_{g,m}^{(r,0)}) \right|^2 \\
\leq & \left(1 + \frac{1}{k-1}\right) \mathbb{E} \left| w_{g,m}^{(r,k-1)} - w_{g,m}^{(r,0)} \right|^2 + \eta_t^2 \sigma_t^2 \\
& + 2K \eta_t^2 \mathbb{E} \left| \psi_{g,m}^{(r,k)} \right|^2 \mathbb{E} \left| \nabla f_m(w_{g,m}^{(r,0)}) \right|^2
\end{align*}
\]
Unrolling the recursion,
\[
\mathbb{E}[|w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}|^2] \leq \sum_{i=1}^{k} \left( 2k\eta_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2 \mathbb{E}[\|\nabla f_m(w_{g,m}^{(r,0)})\|^2 + \eta_i^2 \sigma_i^2] \right) \left( 1 + \frac{1}{k - 1} + 2k\eta_i^2 \beta^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2] \right)^i
\]
(28)

Assuming that \( \eta_i \leq \frac{1}{\beta \sqrt{2k(k-1)}} \) we get \( k\eta_i^2 \beta^2 \leq 1 \).
\[
\mathbb{E}[|w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}|^2] \leq \left( 2k\eta_i^2 \sum_{i=1}^{k} \mathbb{E}[\|\psi_{g,m}^{(r,i)}\|^2 \mathbb{E}[\|\nabla f_m(w_{g,m}^{(r,0)})\|^2 + \eta_i^2 \sigma_i^2] \right) \sum_{i=1}^{k} \left( 1 + \frac{1}{k - 1} + 2 \right)^i
\]
(29)
\[
\leq 4k \left( 2k\eta_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + \eta_i^2 \sigma_i^2] \right)
\]
(30)
\[
\leq 8k^3\eta_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 4k\eta_i^2 \sigma_i^2]
\]
(31)

**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 \( \eta_i \leq \min \left( \frac{1}{\beta \sqrt{2k(k-1)}}, \frac{1}{\beta \sqrt{2k(k-1)}} \right) \) in Algorithm 2, the following is satisfied:
\[
\mathbb{E}[|w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}|^2] \leq 16k^3\eta_i^2 \mathbb{E}[|1 - \psi_{g,m}^{(r,k)}| \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 8k\eta_i^2 \sigma_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2]
\]
\[
+ 12k^2\eta_i^2 \mathbb{E}[|1 - \psi_{g,m}^{(r,k)}| \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 6K\eta_i^2 \sigma_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2]
\]

Proof.
\[
\mathbb{E}[|w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}|^2] = \mathbb{E}[|\psi_{g,m}^{(r,k)}w_{g,m}^{(r,K)} - (1 - \psi_{g,m}^{(r,k)})w_{g,m}^{(r,0)}|^2]
\]
(32)
\[
= \mathbb{E}[|\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)}) + (w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)})|^2]
\]
(33)
\[
= \mathbb{E}[|\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)}) + w_{g,m}^{(r,0)} - w_{g,m}^{(r,K)} + (w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)})|^2]
\]
(34)
\[
\leq 2\mathbb{E}[|\psi_{g,m}^{(r,k)}(w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)})|^2 + 2\mathbb{E}[|(1 - \psi_{g,m}^{(r,k)})w_{g,m}^{(r,K)} - w_{g,m}^{(r,0)})|^2]
\]
(35)

Using lemmas D.6 and D.7.
\[
\mathbb{E}[|w_{g,m}^{(r,k)} - w_{g,m}^{(r,0)}|^2] \leq 2\mathbb{E}[|\psi_{g,m}^{(r,k)}|^2 \left( 8k^3\eta_i^2 \mathbb{E}[|\psi_{g,m}^{(r,k)}| \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 6K^2\eta_i^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2]
\]
\[
+ 2\mathbb{E}[|1 - \psi_{g,m}^{(r,k)}|^2 (4k\eta_i^2 \sigma_i^2 + 3K\eta_i^2 \sigma_i^2)]
\]
(36)
\[
\leq 16k^3\eta_i^2 \mathbb{E}[|1 - \psi_{g,m}^{(r,k)}| \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 8k\eta_i^2 \sigma_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2]
\]
\[
+ 12k^2\eta_i^2 \mathbb{E}[|1 - \psi_{g,m}^{(r,k)}| \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 6K\eta_i^2 \sigma_i^2 \mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2]
\]
(37)

**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 \( \frac{1}{\beta} \leq \eta_i \leq \min \left( \frac{1}{\beta \sqrt{2k(k-1)}}, \frac{1}{\beta \sqrt{2k(k-1)}} \right) \) in Algorithm 2, then the following convergence holds:

(Strong Convex Case)
\[
\mathbb{E} \left[ F(w_g^{(R)}) - F(w_g) \right] \leq \frac{\mu}{4q_K} \mathbb{E}[|w_g^{(0)} - w_g|^2 \exp \left( \frac{\eta_i\mu K R}{2M} \right) + 2\mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 40K^2\beta^2 \left( \frac{\beta^2}{\mu R} + 1 \right) \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 28K\beta \mu^2 R^2 + \frac{2\beta^2K}{\mu^2 R^2} + 1 \right) \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2]
\]

(General Convex Case)
\[
\mathbb{E} \left[ F(w_g^{(R)}) - F(w_g) \right] \leq \frac{\mu}{4q_K} \mathbb{E}[|w_g^{(0)} - w_g|^2 \exp \left( \frac{\eta_i\mu K R}{2M} \right) + 2\mathbb{E}[\|\psi_{g,m}^{(r,k)}\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 40K^2\beta^2 \left( \frac{\beta^2}{\mu R} + 1 \right) \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 28K\beta \mu^2 R^2 + \frac{2\beta^2K}{\mu^2 R^2} + 1 \right) \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2]
\]

where \( q_{K+1} \) are the probabilities of picking global/local routes averaged over all the instances sampled from the global distribution.
Taking squared norm and expectation on both sides with respect to the choice of $h_m$, we get,

$$
\mathbb{E} \left[ \|w_g^{(r+1)} - w_g^*\|^2 \right] \leq \mathbb{E} \left[ \|w_g^{(r)} - w_g^*\|^2 \right] - 2\eta T_1 + \eta^2 E \left[ \|w_g^{(r)} - w_g^*\|^2 \right] + \eta^2 \sigma^2 K \frac{M}{M}
$$

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

$$
\leq \mathbb{E} \left[ \|w_g^{(r)} - w_g^*\|^2 \right] - 2\eta \left( \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K \mathbb{E} \left[ \nabla f_m(w_g^{(r,k-1)}) \right], w_g^{(r)} - w_g^* \right)
$$

Bounding $T_1$

$$
T_1 = -2\eta \left( \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K \mathbb{E} \left[ \nabla f_m(w_g^{(r,k-1)}) \right], w_g^{(r)} - w_g^* \right)
$$

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

$$
T_1 \leq \frac{2\eta \mu K}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K \left( \mathbb{E} \left[ \nabla f_m(w_g^r) \right] - \nabla f_m(w_g^r) \right) - \frac{\mu}{4} \mathbb{E} \left[ \|w_g^{(r)} - w_g^*\|^2 \right] + \beta \mathbb{E} \left[ \|w_g^{(r,k-1)} - w_g^{(r)}\|^2 \right]
$$
Next, using Assumption D.4,

\[
T_1 \leq -2\eta K \left( E[F(w^{(r)})] - F(w_g) \right) - \frac{\eta \mu K}{2M} E\|w^{(r)} - w_g\|^2 \\
+ 32\eta K^3 \beta E\|1 - \psi^{(r)}\|^2 \left( G^2 + 2\beta B \left( E[F(w^{(r)})] - F(w_g) \right) \right) + 16\eta K^2 \beta^2 \sigma^2 E\|\psi^{(r)}\|^2 \\
+ 24\eta K^2 \beta E\|1 - \psi^{(r)}\|^2 \left( G^2 + 2\beta B \left( E[F(w^{(r)})] - F(w_g) \right) \right) + 12\eta K^2 \beta^2 \sigma^2 E\|\psi^{(r)}\|^2 \\
\leq -2\eta K \left( E[F(w^{(r)})] - F(w_g) \right) - \frac{\eta \mu K}{2M} E\|w^{(r)} - w_g\|^2 \\
+ 16\eta K^3 \beta^2 B^2 (4K + 3) E\|1 - \psi^{(r)}\|^2 \left( E[F(w^{(r)})] - F(w_g) \right) \\
+ 8\eta K^3 \beta (4K + 3) E\|1 - \psi^{(r)}\|^2 G^2 + 28\eta K^2 \beta^2 \sigma^2 E\|\psi^{(r)}\|^2 \\
\tag{48}
\]

Bounding \(T_2\)

\[
T_2 = \eta^2 E \left\| \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K \nabla w^{(r, k-1)} f_m(w^{(r, k-1)}) \right\|^2 \\
= \eta^2 E \left\| \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K (\nabla w^{(r, k-1)} f_m(w^{(r, k-1)}) - \nabla f_m(w^{(r)}) + \nabla f_m(w^{(r)})) \right\|^2 \\
\leq 2\eta^2 E \left\| \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K \nabla f_m(w^{(r)})) \right\|^2 \\
+ 2\eta^2 K \cdot \frac{1}{nM} \sum_{m \in [M]} n_m \sum_{k=1}^K E \left\| \nabla f_m(w^{(r)}) \right\|^2 \\
\leq 16\eta^4 K^4 \beta^2 B^2 (4K + 3) E\|1 - \psi^{(r)}\|^2 \left( E[F(w^{(r)})] - F(w_g) \right) + 8\eta^4 K^4 \beta^3 (4K + 3) E\|1 - \psi^{(r)}\|^2 G^2 \\
+ 56\eta^4 K^3 \beta^3 \sigma^2 E\|\psi^{(r)}\|^2 + 2\eta^2 K \left( G^2 + 2\beta B^2 E[F(w^{(r)})] - F(w_g) \right) \\
\tag{54}
\]

Plugging in \(T_1\) and \(T_2\) bounds,

\[
E \left\| w^{(r+1)} - w_g^* \right\|^2 \leq \left( 1 - \frac{\eta \mu K}{2M} \right) E \left\| w^{(r)} - w_g^* \right\|^2 \\
- (2\eta K - 80\eta^3 K^4 \beta^2 B^2 (\eta K + 1) q_0^2 - 4\eta K \beta B^2) \left( E[F(w^{(r)})] - F(w_g) \right) \\
+ 40\eta^3 K^3 \beta (2\eta K^2 + 1) q_0^2 G^2 + 2\eta^2 K G^2 + 28\eta K^2 \beta^2 (2\eta K^2 + 1) q_0^2 \sigma^2 \\
\tag{55}
\]

Rearranging the terms, and replacing \(E[\|\psi^{(r)}\|^2]\) and \(E\|1 - \psi^{(r)}\|^2\) with \(q_0^2\) (probability of picking global route) and \(q_1^2\) respectively, averaged over the instances sampled from the global distribution) and \(q_1^2\) respectively.

\[
E \left\| w^{(r+1)} - w_g^* \right\|^2 \leq \left( 1 - \frac{\eta \mu K}{2M} \right) E \left\| w^{(r)} - w_g^* \right\|^2 \\
- (2\eta K - 80\eta^3 K^4 \beta^2 B^2 (\eta K + 1) q_0^2 - 4\eta K \beta B^2) \left( E[F(w^{(r)})] - F(w_g) \right) \\
+ 40\eta^3 K^3 \beta (2\eta K^2 + 1) q_0^2 G^2 + 2\eta^2 K G^2 + 28\eta K^2 \beta^2 (2\eta K^2 + 1) q_0^2 \sigma^2 \\
\tag{57}
\]
Assuming $\eta \frac{2\mu K}{\beta} \geq 80\eta^3 K^4 \beta^2 B^2 (\eta R + 1) \implies \eta \leq \frac{1}{4\sqrt{3KBR^2}}$ and $\eta \frac{2\mu K}{\beta} \geq 4\eta^2 K \beta B^2 \implies \eta \leq \frac{1}{8\sqrt{K\beta}}$.

we get

$$
\mathbb{E} \left[ \left| w_{g}^{(r+1)} - w_{g}^{*} \right|^2 \right] \leq \left( 1 - \frac{\eta \mu K}{2M} \right) \mathbb{E} \left[ \left| w_{g}^{(r)} - w_{g}^{*} \right|^2 \right] - \eta K (1 - q_1) \left( \mathbb{E} \left[ F(w_{g}^{(r)}) \right] - F(w_{g}^{*}) \right) \\
+ 40\eta^3 K^3 \beta (\eta R + 1) q_1^2 G^2 + 2\eta^2 K \beta G^2 + 28\eta^3 K^2 (2\eta R^2 K + 1) q_0^2 \sigma^2_{\ell}
$$

(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,

$$
\eta \frac{K q_0^2}{\ell K} \mathbb{E} \left[ F(w_{g}^{(r)}) \right] - F(w_{g}^{*}) \leq \left( 1 - \frac{\eta \mu K}{2M} \right) \mathbb{E} \left[ \left| w_{g}^{(r)} - w_{g}^{*} \right|^2 \right] - \mathbb{E} \left[ \left| w_{g}^{(r+1)} - w_{g}^{*} \right|^2 \right] \\
+ 40\eta^3 K^3 \beta (\eta R + 1) q_1^2 G^2 + 2\eta^2 K \beta G^2 + 28\eta^3 K^2 (2\eta R^2 K + 1) q_0^2 \sigma^2_{\ell}
$$

(58)

$$
\therefore \mathbb{E} \left[ F(w_{g}^{(r)}) \right] - F(w_{g}^{*}) \leq \frac{1}{\eta \frac{K q_0^2}{\ell K}} \left( 1 - \frac{\eta \mu K}{2M} \right) \mathbb{E} \left[ \left| w_{g}^{(r)} - w_{g}^{*} \right|^2 \right] - \frac{1}{\eta \frac{K q_0^2}{\ell K}} \mathbb{E} \left[ \left| w_{g}^{(r+1)} - w_{g}^{*} \right|^2 \right] \\
+ 40\eta^3 K^3 \beta (\eta R + 1) q_1^2 G^2 + 2\eta^2 K \beta G^2 + 28\eta^3 K^2 (2\eta R^2 K + 1) q_0^2 \sigma^2_{\ell}
$$

(59)

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

$$
\mathbb{E} \left[ F(w_{g}^{(R)}) \right] - F(w_{g}^{*}) \leq \frac{\mu}{q_0 K} \mathbb{E} \left[ \left| w_{g}^{(0)} - w_{g}^{*} \right|^2 \right] \exp \left( -\frac{\eta \mu K R}{2M} \right) + \frac{2G^2}{q_0^2 \mu R} \\
+ \frac{40K^2 \beta}{\mu^2 R^2} \left( \frac{\beta^2}{\mu R} + 1 \right) q_1^2 G^2 + \frac{28K \beta}{\mu^2 R^2} \left( \frac{2\beta^3 K}{\mu^2 R^2} + 1 \right) \sigma^2_{\ell}
$$

(60)

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

$$
\mathbb{E} \left[ F(w_{g}^{(R)}) \right] - F(w_{g}^{*}) \leq \frac{1}{\eta K q_0^2 (R + 1)} \mathbb{E} \left[ \left| w_{g}^{(0)} - w_{g}^{*} \right|^2 \right] + \eta R \left( \frac{2G^2}{q_0^2} \right)^{1/2} + \\
+ \eta^2 \left( 40K^2 \beta \frac{q_1^2 G^2}{q_0^2} \right)^{1/2} + \eta^2 \left( 40K^2 \beta \frac{q_1^2 G^2}{q_0^2} \right)^{1/3} + \eta^2 \left( 28K \beta \sigma^2_{\ell} \right)^{1/3} + \eta^2 \left( 56K \beta^3 \sigma^2_{\ell} \right)^{1/5}
$$

(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 3 the following is satisfied:

$$
\mathbb{E} \left[ \left| w_{g(m)}^{(r)} - w_{g(m)}^{(0)} \right|^2 \right] \leq 4K^2 \eta^2 \mathbb{E} \left[ \left| \nabla f_m (w_{g(m)}^{(r)}) \right|^2 \right] + 2k\eta^2 \sigma^2_{\ell} + 4K^2 \eta^2 \beta^2 \sum_{i=1}^{k} \mathbb{E} \left[ \left| w_{p(m)}^{(r,i-1)} - w_{g(m)}^{(r)} \right|^2 \right]
$$

(62)

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

$$
\mathbb{E} \left[ \left| w_{g(m)}^{(r)} - w_{g(m)}^{(0)} \right|^2 \right] = \mathbb{E} \left[ \left| w_{g(m)}^{(r-1)} - \eta \nabla w_{g(m)}^{(r-1)} f_m (w_{g(m)}^{(r-1)}) - w_{g(m)}^{(r)} \right|^2 \right]
$$

(63)

Using triangle inequality and separation of variance, we get,

$$
\leq \left( 1 + \frac{1}{K - 1} \right) \mathbb{E} \left[ \left| w_{g(m)}^{(r-1)} - w_{g(m)}^{(0)} \right|^2 \right] + k\eta^2 \mathbb{E} \left[ \left| \nabla w_{g(m)}^{(r-1)} f_m (w_{g(m)}^{(r-1)}) \right|^2 \right] + \eta^2 \sigma^2_{\ell}
$$

$$
\leq \left( 1 + \frac{1}{K - 1} \right) \mathbb{E} \left[ \left| w_{g(m)}^{(r-1)} - w_{g(m)}^{(0)} \right|^2 \right] + \eta^2 \sigma^2_{\ell}
$$

(64)

$$
+ k\eta^2 \mathbb{E} \left[ \nabla f_m (w_{g(m)}^{(r-1)}) - \nabla f_m (w_{g(m)}^{(0)}) \right]^2 + \nabla f_m (w_{g(m)}^{(0)}) \left| \nabla f_m (w_{g(m)}^{(0)}) \right|^2
$$

(65)
Using lemmas D.6 and D.10, satisfies Assumptions D.2, D.3, D.4, using the learning rate \( \eta \), we have:

\[
\begin{align*}
&\mathbb{E}[|\psi_{u,m}(r,0) - \psi_{u,m}(r,K)\|^2] \\
&\quad \leq 2K^2 \eta_2 \mathbb{E}[|1 - \psi_{u,m}(r,K)\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 10K^2 \eta_2^2 \mathbb{E}[\|\psi_{u,m}(r,K)\|^2].
\end{align*}
\]

Proof.

\[
\begin{align*}
&\mathbb{E}[|\psi_{u,m}(r,0) - \psi_{u,m}(r,K)\|^2 = \mathbb{E}[|\psi_{u,m}(r,0) - \psi_{u,m}(r,K)\|^2 + \mathbb{E}[|\psi_{u,m}(r,K) - \psi_{u,m}(r,0)\|^2 + \mathbb{E}[|\psi_{u,m}(r,0) - \psi_{u,m}(r,K)\|^2
\end{align*}
\]

Using lemmas D.6 and D.10:

\[
\begin{align*}
&\mathbb{E}[|w_{u,m}^{(r,0)} - w_{u,m}^{(r,K)}\|^2 \leq 2K^2 \eta_2 \mathbb{E}[|1 - \psi_{u,m}(r,K)\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 10K^2 \eta_2^2 \mathbb{E}[\|\psi_{u,m}(r,K)\|^2.
\end{align*}
\]

Assuming \( 8K^2 \eta_2^{\beta_2} \leq 1 \Rightarrow \eta \leq \frac{1}{2 \sqrt{2K^2}} \) and unrolling the recursion over \( w_{u,m}^{(r,1)} - w_g^{(r)} \).

\[
\begin{align*}
&\leq \sum_{i=1}^{k} (2K^2 \eta_2^2 \mathbb{E}[|1 - \psi_{u,m}(r,K)\|^2 \mathbb{E}[\|\nabla f_m(w_g^{(r)})\|^2 + 10K^2 \eta_2^2 \mathbb{E}[\|\psi_{u,m}(r,K)\|^2]
\end{align*}
\]

Theorem D.12 (Convergence of the Global Model for Non-convex Case). If each client’s objective function satisfies Assumptions D.2, D.3, D.4 and the learning rate \( \eta = \min \left( \frac{1}{2 \sqrt{2K^2} \beta_2}, \frac{1}{2 \sqrt{40K^4 \beta_2^2}} \right) \) in Algorithm 2, then the following convergence holds:
\[ F(w_y^{(r+1)}) \leq F(w_y^{(r)}) + \langle \nabla F(w_y^{(r)}), w_y^{(r+1)} - w_y^{(r)} \rangle + \frac{\beta}{2} \|w_y^{(r+1)} - w_y^{(r)}\|^2 \]  

Taking expectation on both sides,

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

Using Equation (80) for second and third terms, and using the fact that the expectation is with respect to the choice of \( h_m \),

\[ \leq \mathbb{E}\left[F(w_y^{(r)})\right] - \eta \left\langle \nabla F(w_y^{(r)}), \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \mathbb{E}\left[h_m(w_y^{(r,k-1)})\right] \right\rangle + \frac{\beta \eta^2}{2} \mathbb{E}\left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} h_m(w_y^{(r,k-1)}) \right\|^2 \right], \]  

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

\[ \mathbb{E}\left[F(w_y^{(r+1)})\right] \leq \mathbb{E}\left[F(w_y^{(r)})\right] - \eta \left\langle \nabla F(w_y^{(r)}), \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\rangle \]  

\[ + \beta \eta^2 \mathbb{E}\left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \right] + \frac{\eta^2 \beta \sigma^2 K}{2M} \]  

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

\[ \mathbb{E}\left[F(w_y^{(r+1)})\right] \leq \mathbb{E}\left[F(w_y^{(r)})\right] - \eta \left[ -\frac{1}{2} \mathbb{E}\left\|\nabla F(w_y^{(r)}) - \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \right] \]  

\[ - \eta \left[ \frac{1}{2} \mathbb{E}\left\|\nabla F(w_y^{(r)})\right\|^2 + \frac{1}{2} \mathbb{E}\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \right] \]  

\[ + \frac{\beta \eta^2}{2} \mathbb{E}\left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \right] + \frac{\eta^2 \beta \sigma^2 K}{2M} \]  

\[ \leq \mathbb{E}\left[F(w_y^{(r)})\right] - \frac{\eta \beta}{2} \mathbb{E}\left[\right\| \nabla F(w_y^{(r)}) \right\|^2 \]  

\[ - \left( \frac{\eta}{2} - \frac{\beta \eta^2}{2} \right) \mathbb{E}\left[\left\| \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \right] \]  

\[ + \frac{\eta \beta}{2} \mathbb{E}\left[\right\| \nabla F(w_y^{(r)}) - \frac{1}{M} \sum_{m \in [M]} \alpha_m \sum_{k=1}^{K} \nabla w_y^{(r,k-1)} f_m(w_y^{(r,k-1)}) \right\|^2 \]  

\[ + \frac{\eta^2 \beta \sigma^2 K}{2M} \]  

\[ (84) \]  

\[ (85) \]
\[
\leq \mathbb{E} \left[ F(w_g^{(r)}) \right] - \frac{\eta_r}{2} \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 + \left( \frac{3\eta_r}{2} - \beta\eta_t^2 \right) \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 + \frac{\eta_r^2 \beta \sigma^2 K}{2M}
\]

760 Using Lemma [D.11]

\[
\mathbb{E} \left[ F(w_g^{(r+1)}) \right] \leq \mathbb{E} \left[ F(w_g^{(r)}) \right] - \left( \frac{3\eta_r}{2} - \beta\eta_t^2 \right) \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 + \frac{\eta_r^2 \beta \sigma^2 K}{2M} - \left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) \frac{1}{M} \sum_{m \in [M]} \alpha_m \left( 20 K^3 \eta_r^2 \mathbb{E} \left( 1 - \psi_{g,m}^\ell \right) \mathbb{E} \left\| \nabla f_m(w_g^{(r)}) \right\|^2 \right)
\]

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

\[
\mathbb{E} \left[ F(w_g^{(r+1)}) \right] \leq \mathbb{E} \left[ F(w_g^{(r)}) \right] - \left( \frac{3\eta_r}{2} - \beta\eta_t^2 \right) \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 + \frac{\eta_r^2 \beta \sigma^2 K}{2M} - \left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) \left( 20 K^3 \eta_r^2 \mathbb{E} \left( 1 - \psi_{g,m}^\ell \right) \mathbb{E} \left\| \nabla f_m(w_g^{(r)}) \right\|^2 \right)
\]

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

\[
\left( \frac{3\eta_r}{2} - \beta\eta_t^2 - 20 K^4 \beta^2 \eta_t^2 B^2 q_i^2 \left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) \right) \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 \leq \mathbb{E} \left[ F(w_g^{(r)}) \right] - \mathbb{E} \left[ F(w_g^{(r+1)}) \right] - 20 q_i^2 K^4 \beta^2 \eta_t^2 G^2 \left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) - 10 q_i^2 K^3 \beta \sigma^2 \left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) + \frac{\eta_r^2 \beta \sigma^2 K}{2M}
\]

764 Assuming \( 10 K^4 \beta^2 \eta_t^2 B^2 \leq \frac{\eta_r}{2} \Rightarrow \eta_r \leq \frac{1}{\sqrt{10 K^4 \beta^2 B^2}} \) and \( 20 K^4 \beta \eta_t^2 B^2 \leq \frac{\eta_r}{2} \Rightarrow \eta_r \leq \frac{1}{\sqrt{20 K^4 \beta^2 B^2}} \)

\[
\left( \frac{\eta_r}{2} - \beta\eta_t^2 \right) q_i^2 \mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 \leq \mathbb{E} \left[ F(w_g^{(r)}) \right] - \mathbb{E} \left[ F(w_g^{(r+1)}) \right] + \frac{\eta_r^2 \beta \sigma^2 K}{2M} + 20 q_i^2 K^4 \beta^2 \eta_t^2 G^2 \left( \frac{2\beta \eta_t^2 - \eta_r}{2} \right) + 10 q_i^2 K^3 \beta \sigma^2 \left( \beta \eta_t^2 - \frac{\eta_r}{2} \right)
\]

\[
\mathbb{E} \left\| \nabla F(w_g^{(r)}) \right\|^2 \leq \frac{2}{\eta_r q_i^2} \mathbb{E} \left[ F(w_g^{(r)}) \right] - \mathbb{E} \left[ F(w_g^{(r+1)}) \right] + \frac{\eta_r \beta \sigma^2 K}{M q_i^2} + 40 q_i^2 K^4 \beta^2 \eta_t^2 G^2 \left( \frac{2\beta \eta_t^2 - \eta_r}{2} \right) + 20 K^3 \beta \eta_t^2 \left( \beta \eta_t^2 - \frac{\eta_r}{2} \right)
\]
Taking average over all the $R$ rounds,

\[
\frac{1}{R} \sum_{r=1}^{R} \mathbb{E} \left\| \nabla F(w^{(r)}_y) \right\|^2 \leq \frac{2}{\eta Kq_0^2 R} \left[ \mathbb{E} \left[ F(w^{(1)}) \right] - \mathbb{E} \left[ F(w^{(R+1)}) \right] \right] + \frac{\eta K^2 \sigma^2}{Mq_0^2} + 40q_0^2 K^3 \beta^2 \eta G^2 \left( \frac{2\beta^2 - \eta}{2} \right) + 20K^3 \beta^2 \eta \sigma^2 \left( \frac{\beta^2 - \eta}{2} \right)
\]

(94)

\[\Box\]

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.2, D.3, and D.4 and conditioning on $\eta \leq \frac{1}{\sqrt{K}}$ in Algorithm 2 the following are satisfied:

\[
\mathbb{E}\left\| u^{(r,K)} - \tilde{u}^{(r,0)} \right\|^2 \leq 18K^2 \eta K^2 \mathbb{E}\left\| \nabla f_m(w^{(r)}_y) \right\|^2 + 108K^4 \eta^2 K^2 \mathbb{E}\left\| \nabla f_m(w^{(r)}_y) \right\|^2 + 126K^3 \eta^4 \mathbb{E}\left\| \nabla f_m(w^{(r)}_y) \right\|^2
\]

(95)

**Proof.**

\[
\mathbb{E}\left\| u^{(r,K)} - \tilde{u}^{(r,0)} \right\|^2 = \mathbb{E}\left\| (\psi^{(r,K)}_y - \eta \nabla \psi^{(r,K)}_y f_m(w^{(r,K)}_y)) (w^{(r,K)}_y - \eta \nabla w^{(r,K-1)} f_m(w^{(r,K-1)}_y)) + (1 - \psi^{(r,K)}_y + \eta \nabla \psi^{(r,K)}_y f_m(w^{(r,K)}_y)) w^{(r,K)}_y - \psi^{(r,0)}_y w^{(r,0)}_y - (1 - \psi^{(r,0)}_y) w^{(r,0)}_y \right\|^2
\]

(96)

Using the convexity of $f_m$,

\[
\nabla w^{(r,K)}_y f_m(w^{(r,K)}_y) = \nabla w^{(r,K)}_y f_m(\psi^{(r,K)}_y w^{(r,K)}_y + (1 - \psi^{(r,K+1)}_y) w^{(r,K)}_y)
\]

(97)

and

\[
\nabla \psi^{(r,K)}_y f_m(\tilde{u}^{(r,K)}_y) = \nabla \psi^{(r,K)}_y f_m(\psi^{(r,K)}_y w^{(r,k)}_y - w^{(r,K)}_y) - w^{(r,K)}_y
\]

(98)

we get,

\[
\mathbb{E}\left\| u^{(r,K)} - \tilde{u}^{(r,0)} \right\|^2 \leq \mathbb{E}\left\| \psi^{(r,K)}_y w^{(r,K)}_y - w^{(r,K)}_y \right\|^2 + \left( 1 - \psi^{(r,K)}_y \right) w^{(r,K)}_y - \psi^{(r,0)}_y w^{(r,0)}_y - (1 - \psi^{(r,0)}_y) w^{(r,0)}_y
\]

(99)

\[
\leq \left( 1 + \frac{1}{K - 1} \right) \mathbb{E}\left\| w^{(r,K)}_y - w^{(r,0)}_y \right\|^2 + 3K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,K-1)}_y) \right\|^2
\]

(100)

\[
\leq \left( 1 + \frac{1}{K - 1} \right) \mathbb{E}\left\| u^{(r,K-1)}_y - \tilde{u}^{(r,0)}_y \right\|^2 + 3K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,K-1)}_y) \right\|^2
\]

(101)

Using Lemma D.6 and conditioning on $\eta \leq \frac{1}{\sqrt{K}}$

\[
\leq \left( 1 + \frac{1}{K - 1} + 6K^2 \eta^2 \beta^2 \right) \mathbb{E}\left\| w^{(r,K)}_y - w^{(r,0)}_y \right\|^2 + 6K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,0)}_y) \right\|^2 + 6K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,0)}_y) \right\|^2
\]

(102)

(103)

Using Lemma D.6 and conditioning on $\eta \leq \frac{1}{\sqrt{K}},

\[
\leq \left( 1 + \frac{1}{K - 1} + 6K^2 \eta^2 \beta^2 \right) \mathbb{E}\left\| w^{(r,K)}_y - w^{(r,0)}_y \right\|^2 + 6K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,0)}_y) \right\|^2 + 6K^2 \eta^2 \mathbb{E}\left\| \nabla f_m(w^{(r,0)}_y) \right\|^2
\]

(104)

(105)
\[ \mathbb{E}[w_{p,m}^{(r,K)} - \tilde{w}_{p,m}^{(r,0)}]^2 \leq \sum_{i=1}^{K} \left( 6K\eta^2 \mathbb{E} \| \nabla f_m(\tilde{w}_{p,m}^{(r,0)}) \|^2 + 36K^3\eta^4 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 + 42K^2\eta^4 \sigma^2 \right. \\
+ 3K\eta^2 \mathbb{E} \| \psi_{g,m}^{(r,K)} \|^2 + 48K^4\eta^4 \mathbb{E} \| \psi_{g,m}^{(r,K)} \|^2 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 \left( 1 + \frac{1}{K-1} + 6K\eta^2 \beta^2 \right) \right) \] (106)

Assuming \( 6K\eta^2 \beta^2 \leq 1 \implies \eta \leq \frac{1}{\beta \sqrt{6K}} . \)

\[ \mathbb{E}[|u_{p,m}^{(r,K)} - \tilde{u}_{p,m}^{(r,0)}|^2] \leq 3K \left( 6K\eta^2 \mathbb{E} \| \nabla f_m(\tilde{u}_{g,m}^{(r,0)}) \|^2 + 36K^3\eta^4 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 + 42K^2\eta^4 \sigma^2 \right. \\
+ 3K\eta^2 \mathbb{E} \| \phi_{g,m}^{(r,K)} \|^2 + 48K^4\eta^4 \mathbb{E} \| \phi_{g,m}^{(r,K)} \|^2 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 \left( 1 + \frac{1}{K-1} + 6K\eta^2 \beta^2 \right) \right) \] (107)

\[ \mathbb{E}[|u_{p,m}^{(r+1,K)} - \tilde{u}_{p,m}^{(r,K)}|^2] \leq 18K^2\eta^2 \mathbb{E} \| \nabla f_m(u_{g,m}^{(r,0)}) \|^2 + 108K^4\eta^4 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 + 126K^2\eta^4 \sigma^2 \\
+ 9K^2\eta^2 \mathbb{E} \| \phi_{g,m}^{(r,K)} \|^2 + 144K^5\eta^4 \mathbb{E} \| \phi_{g,m}^{(r,K)} \|^2 \mathbb{E} \| \nabla f_m(w_g^{(r)}) \|^2 \] (108)

**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:

\[ \mathbb{E}[|u_{p,m}^{(r+1,0)} - \tilde{u}_{p,m}^{(r,0)}|^2] \leq \mathbb{E} \left[ \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right] \frac{1}{M} \sum_{c \in [M]} u_{g,c}^{(r,K)} + \left( 1 - \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right) w_{\ell,m}^{(r+1,K)} \\
- \psi_{g,m}^{(r,K)} u_{g,m}^{(r,K)} - (1 - \psi_{g,m}^{(r,K)}) w_{\ell,m}^{(r,K)} \right]^2 \] (109)

\[ \leq 2\mathbb{E} \left[ \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right] \frac{1}{M} \sum_{c \in [M]} u_{g,c}^{(r,K)} - \psi_{g,m}^{(r,K)} u_{g,m}^{(r,K)} \right|^2 \\
+ 2\mathbb{E} \left[ \left( 1 - \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right) w_{\ell,m}^{(r+1,K)} - (1 - \psi_{g,m}^{(r,K)}) w_{\ell,m}^{(r,K)} \right]^2 \] (110)

\[ \leq 2\mathbb{E} \left[ \left( 1 - \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right) \right] \left( \frac{1}{M} \sum_{c \in [M]} w_{g,c}^{(r,K)} - u_{g,c}^{(r,K)} \right)^2 \\
+ 2\mathbb{E} \left[ \left( \psi_{g,c}^{(r,K)} - \frac{1}{M} \sum_{c \in [M]} \psi_{g,c}^{(r,K)} \right) \left( w_{\ell,m}^{(r+1,K)} - w_{\ell,m}^{(r,K)} \right) \right]^2 \] (111)

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

\[ \mathbb{E}[|u_{p,m}^{(r+1,0)} - \tilde{u}_{p,m}^{(r,0)}|^2] \leq 18 \left( K\sigma^2 \eta^2 + \left( \frac{\delta^\psi}{M} + \frac{\delta^\psi}{M} \right) K^2 \eta^2 \right) \left( K\sigma^2 \eta^2 + \left( \frac{\delta^\psi}{M} + \frac{\delta^\psi}{M} \right) K^2 \eta^2 \right) \\
+ 6(1 + \eta^2 K^2 \beta^2) \eta^2 K^2 \left( K\sigma^2 \eta^2 + \left( \frac{\delta^\psi}{M} + \frac{\delta^\psi}{M} \right) K^2 \eta^2 \right) \left( \mathbb{E} + B^2 \mathbb{E} \| \nabla F(w_g^{(r)}) \|^2 \right) \] (112)

**Lemma D.15** (One epoch progress of the personalized model). If \( m^{th} \) client’s objective function \( f_m \) satisfies Assumptions D.7 - D.9 in Algorithm 2 the following are satisfied:

\[ \mathbb{E}[|u_{p,m}^{(r+1,k)} - u_{p,m}^{(r,k)}|^2] \leq 3\eta^2 \mathbb{E} \| \nabla f_m(u_{g,m}^{(r,k)}) \|^2 + 3\eta^2 \mathbb{E} \| w_{\ell,m}^{(r,k)} - w_{g,m}^{(r,k)} \|^2 + 3\eta^2 \mathbb{E} \| \psi_{g,m}^{(r,k)} \|^2 \]
and hence,
\[
E\|w_{p,m}^{(r,K)} - w_{p,m}^{(r,K)}\|^2 \leq 6\beta \eta^2 E\|E[f_m(u_{p,m}^{(r,K)})] - f(u_{p,m}^*)\|^2 + 3\eta K \sum_{i=k}^K E\|\psi_{g,m}^{(r,i)}\|^2 \\
+ 36K^3 \eta \|\nabla f_m(w_g^{(r)})\|^2 + 40K^2 \eta \|\nabla f_m(w_g^{(r)})\|^2 \\
+ 48K^4 \eta \|\nabla f_m(w_g^{(r)})\|^2 \sum_{i=k}^K E\|\psi_{g,m}^{(r,i)}\|^2
\]

Proof.
\[
E\|w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,k)}\|^2 = E\|\psi_{g,m}^{(r,k+1)} w_{g,m}^{(r,k+1)} + (1 - \psi_{g,m}^{(r,k+1)}) w_{g,m}^{(r,k)} - \psi_{g,m}^{(r,k)} w_{g,m}^{(r,k)} - (1 - \psi_{g,m}^{(r,k)}) w_{g,m}^{(r,K)}\|^2
\]

Using,
\[
\nabla \psi_{g,m}^{(r,k)} f_m(w_{p,m}^{(r,k)}) = \nabla \psi_{g,m}^{(r,k)} f_m (\psi_{g,m}^{(r,k+1)} w_{g,m}^{r,k} + (1 - \psi_{g,m}^{(r,k+1)}) w_{g,m}^{(r,K)})
\]

and,
\[
\nabla \psi_{g,m}^{(r,k)} f_m (\tilde{w}_{g,m}^{(r,k)}) = \nabla \psi_{g,m}^{(r,k)} f_m (\psi_{g,m}^{(r,k)} [w_{g,m}^{(r,k)} - w_{g,m}^{(r,K)}] - w_{g,m}^{(r,K)})
\]

we get,
\[
\leq \eta E\|w_{g,m}^{(r,K)} - w_{g,m}^{(r,K)}\|^2 \leq \eta E\|\nabla f_m(w_{g,m}^{(r,K)})\|^2 + \eta E\|\nabla f_m(w_{g,m}^{(r,K)})\|^2 + \eta E\|\nabla f_m(w_{g,m}^{(r,K)})\|^2
\]

From Lemmas D.6 and D.7
Summing over \(i = k\) to \(K\),
\[
E\|w_{p,m}^{(r,K)} - w_{p,m}^{(r,K)}\|^2 = E\|\sum_{i=k}^K w_{p,m}^{(r,k+1)} - w_{p,m}^{(r,k)}\|^2
\]
\[
\leq 3\eta^2 E\|\nabla f_m(w_{p,m}^{(r,K)})\|^2 + \eta^2 E\|\nabla f_m(w_{p,m}^{(r,K)})\|^2 + \eta^2 E\|\nabla f_m(w_{p,m}^{(r,K)})\|^2
\]
\[
+ 6\eta^2 \sum_{i=k}^K \left(6K^2 \eta^2 E\|\nabla f_m(w_g^{(r)})\|^2 + 3K \eta^2 \|\nabla f_m(w_g^{(r)})\|^2 + 4K \eta^2 \|\nabla f_m(w_g^{(r)})\|^2 \right)
\]
\[
+ 6\eta^2 \sum_{i=k}^K \left(8K^3 \eta^2 E\|\psi_{g,m}^{(r,i)}\|^2 E\|\nabla f_m(w_g^{(r)})\|^2 + 4K \eta^2 \|\nabla f_m(w_g^{(r)})\|^2 \right)
\]

34
Then, the following convergence holds:

\[
\text{(Convergence of the Personalized Model for Convex (Strong and General) Cases)}
\]

\[
\text{Proof.}\]

\[
\begin{align*}
&\leq 6\beta\eta_t^2 \left(\mathbb{E}[f_m(w^{(r,K)}_{p,m})] - f(w^*_{p,m})\right) + 3\eta_t^2 K \sum_{i=k}^{\text{R}} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \\
&+ 36K^3\eta_t^2 \mathbb{E}||\nabla f_m(w^{(r)})||^2 + 18K^2\eta_t^2 \sigma_t^2 \\
&+ 48K^4\eta_t^2 \mathbb{E}||\nabla f_m(w^{(r)})||^2 \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 + 24K^2\eta_t^2 \sigma_t^2 \\
&= (126)
\end{align*}
\]

\[
\begin{align*}
&\mathbb{E}[\|w^{(r,K)}_{p,m} - w^{(r,m)}\|^2] \leq 6\beta\eta_t^2 \left(\mathbb{E}[f_m(w^{(r,K)}_{p,m})] - f(w^*_{p,m})\right) + 3\eta_t^2 K \sum_{i=k}^{\text{R}} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \\
&+ 36K^3\eta_t^2 \mathbb{E}||\nabla f_m(w^{(r)})||^2 + 40K^2\eta_t^4 \sigma_t^2 \\
&+ 48K^4\eta_t^4 \mathbb{E}||\nabla f_m(w^{(r)})||^2 \sum_{i=k}^{K} \mathbb{E}||\psi_{g,m}^{(r,i)}||^2 \\
&= (127)
\end{align*}
\]

\[\square\]

**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 \( \frac{1}{\mu R} \leq \eta_t \leq \frac{1}{K\sigma_t} \) in Algorithm D.2 then the following convergence holds:

**(Convex Case)**

\[
\begin{align*}
\mathbb{E} \left[ f_m(w^{(0)}_{p,m}) \right] - f_m(w^*_{p,m}) &\leq \frac{36\mu^2}{K^2 R} \mathbb{E}||w^{(1,K)}_{p,m} - w^*_{p,m}||^2 \exp \left( \frac{1}{K-1} - \eta_t \mu R \right) \\
&+ 12K^2\eta_t^2 \delta_m^{w_g} + 12K^2\eta_t^2 \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}||\nabla F(w^{(r+1)}_{g})||^2 + 4K\eta_t^2 \sigma_t^2 + \frac{q_g^2}{2} + 16K^3\eta_t^4 \delta_m^{w_g} \\
&+ 16K^3\eta_t^4 \frac{q_g^2}{6} \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}||\nabla F(w^{(r+1)}_{g})||^2 + K^2\eta_t^4 \left( \frac{\sigma_t^2}{K} + \left( \delta_m^{\psi} + \frac{\delta_m^{\psi}}{M} \right) \left( \sigma_t^2 + \left( \delta_m^{w_g} + \frac{\delta_m^{w_g}}{M} \right) \right) \right) \\
&+ \eta_t^2 \left( K^2\eta_t^2 \left( \frac{\sigma_t^2}{K} + \left( \delta_m^{\psi} + \frac{\delta_m^{\psi}}{M} \right) \right)^{1/3} \right) \\
&\text{where } \frac{1}{\mu R} \sum_{r=1}^{R} \mathbb{E}||\nabla F(w^{(r)}_{g})||^2 \text{ is bounded as shown in Theorem D.12}
\end{align*}
\]

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

1. For all samples \( x_m \), define \( \tilde{w}^{(r,K)}_{p,m} (x_m) \leftarrow \psi_{g,m}^{(r,K)} (x_m) w^{(r,K)}_{g,m} (x_m) \).
2. Train policy parameters \( \psi_{g,m}^{(r+1)} \leftarrow \psi_{g,m}^{(r)} - \eta_t \nabla_{\psi_{g,m}} f_m(\tilde{w}^{(r,K)}_{p,m} (x_m), y_m) \).
3. For all samples \( x_m \), define \( w^{(r,K)}_{p,m} (x_m) \leftarrow \psi_{g,m}^{(r+1)} (x_m) w^{(r,K)}_{g,m} (x_m) + (1 - \psi_{g,m}^{(r+1)} (x_m)) u^{(r,K)}_{L,m} (x_m) \).
4. Train global parameters \( w^{(r,K)}_{g,m} \leftarrow w^{(r,K-1)}_{g,m} - \eta_t \nabla_{w^{(r,K-1)}_{g,m}} f_m(w^{(r,K)}_{p,m} (x_m), y_m) \).
And using Assumption D.4,

\[ \mathbb{E}[|w_{p,m}^{(r+1,K)} - w_{p,m}^*|^2] = \mathbb{E}[|w_{p,m}^{(r+1,K)} - w_{p,m}^{(r+1,0)} + w_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)} + w_{p,m}^{(r,K)} - w_{p,m}^*|^2] \]

(128)

\[ \leq 2K \mathbb{E}[|w_{p,m}^{(r+1,K)} - w_{p,m}^{(r,0)}|^2] + 2K \mathbb{E}[|w_{p,m}^{(r+1,0)} - w_{p,m}^{(r,K)}|^2] + \left(1 + \frac{1}{K-1}\right) \mathbb{E}[|w_{p,m}^{(r,K)} - w_{p,m}^*|^2] \]

(129)

Rearranging the terms,

\[ 36K^2 \eta_i \left[ f_m(w_{p,m}^{(r+1,0)}) - f_m(w_{p,m}^*) \right] \leq \left(1 + \frac{1}{K-1} - \mu \eta \right) \mathbb{E}[|w_{p,m}^{(r,K)} - w_{p,m}^*|^2] - \mathbb{E}[|w_{p,m}^{(r+1,K)} - w_{p,m}^*|^2] \]

(130)

For strong convex (\( \mu > 0 \)) case, using the linear convergence rate lemma from [12] (Lemma 1) and Definition D.5.
We proceed with a lemma which bounds the deviation of the personalized model \( w_{D.3}, D.4, \) and \( w_{D.6} \).

The last inequality follows from Assumption D.4.

### 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:

\[
\mathbb{E}[\|w_{(r+1), m}^{(r)} - w_{(r), m}^{(r)}\|_2^2] \leq (1 - 2\eta_i K \beta^2 + \eta_i^2 K^2 \beta^4) \eta_i^2 K^2 \left( G^2 + B^2 \mathbb{E}[\|\nabla F(w_{(r)}^r)\|^2]\right)
\]

**Proof.**

\[
\mathbb{E}[\|w_{(r+1), m}^{(r)} - w_{(r), m}^{(r)}\|_2^2] = \mathbb{E}[\|w_{(r+1)}^{(r)} - w_{(r)}^{(r)} + \eta_i \sum_{k=1}^K g_m f_m(w_{g}^{(r)})\|_2^2]
\]

\[
= \mathbb{E}[\|w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\|_2^2] + \eta_i \sum_{k=1}^K \mathbb{E}[\|g_m f_m(w_{g}^{(r)})\|_2^2]
\]

\[
\leq (1 - \eta_i K \beta^2) \left( w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\right)\|_2^2 \leq (1 - \eta_i K \beta^2) \left( w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\right)\|_2^2
\]

\[
= (1 - \eta_i K \beta^2)^2 \mathbb{E}\left[ \left( w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\right)\|_2^2 \right]
\]

\[
= (1 - \eta_i K \beta^2)^2 \mathbb{E}\left[ \left( \left( w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\right)\|_2^2 \right) \right]
\]

\[
= (1 - \eta_i K \beta^2)^2 \mathbb{E}\left[ \left( \left( w_{(r+1)}^{(r)} - w_{(r)}^{(r)}\right)\|_2^2 \right) \right]
\]

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 \( \eta_i \leq \frac{1}{K \sqrt{d(1 - \beta^2)}} \), the following is satisfied:

\[
\mathbb{E}[\|w_{p,m}^{(r+1)} - w_{p,m}^{(r)}\|_2^2] \leq 18K^5 \eta_i^2 \mathbb{E}[\|\nabla f_m(w_{g}^{(r)})\|^2] + 9K^4 \eta_i^2 \sigma_1^2 + 36K^3 \eta_i^2 \sigma_2^2 \mathbb{E}[\|1 - \psi_{y,m}^{(r)}\|^2]
\]

\[
+ 24K^4 \eta_i^2 \mathbb{E}[\nabla f_m(w_{g}^{(r)})\|^2] \mathbb{E}[\|\psi_{y,m}^{(r)}\|^2]
\]
Proof. We start with using the update rule stated for the personalized model at the beginning of Theorem D.16
\[
\mathbb{E}||w_m^{(r,k+1)} - w_m^{(r,0)}||^2 = \mathbb{E}||w_m^{(r,k+1)} - w_m^{(r,K)} + (1 - \psi_m^{(r,k+1)})w_m^{(r,K)} - w_m^{(r,0)}||^2
\] (145)
Expanding by one iterate,
\[
= \mathbb{E}|| \left( \psi_m^{(r,k)} - \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \left( w_m^{(r,k)} - \eta_l \nabla w_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \\
+ \left( 1 - \psi_m^{(r,k)} + \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \left( w_m^{(r,K)} - w_m^{(r,0)} \right) ||^2 \] (146)
\[
= \mathbb{E}|| \left( \psi_m^{(r,k)} - w_m^{(r,k)} - \psi_m^{(r,k)} \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) + \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \left( w_m^{(r,k)} - \psi_m^{(r,k)} \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) + \eta_l \nabla w_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \left( w_m^{(r,K)} - w_m^{(r,0)} \right)||^2 \] (147)
\[
= \mathbb{E}||w_m^{(r,k)} - \psi_m^{(r,k)} \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) ||^2 \\
+ \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) \left( w_m^{(r,k)} - \psi_m^{(r,k)} \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) + \eta_l \nabla w_m^{(r,k)} f_m(w_m^{(r,k)}) \right) \left( w_m^{(r,K)} - w_m^{(r,0)} \right)||^2 \] (148)
\[
\leq 3\mathbb{E}||w_m^{(r,k)} - \psi_m^{(r,k)} \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) ||^2 + 3\mathbb{E}||w_m^{(r,K)} - w_m^{(r,0)} ||^2 + 3\eta_l^2 \mathbb{E}||\nabla w_m^{(r,k)} f_m(w_m^{(r,k)}) ||^2 \] (149)
The inequality was derived from the fact that \( \mathbb{E}|| - \eta_l \nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)}) ||^2 = \mathbb{E}||\psi_m^{(r,k+1)} - \psi_m^{(r,k)}||^2 \leq 1. \)
Unrolling the recursion across \( r \in [R] \), then using Lemmas D.6 and D.10 and Assumption D.4
\[
\mathbb{E}||w_m^{(r,K)} - w_m^{(r,0)}||^2 \leq \frac{K}{K-1} \left( 1 + \frac{1}{K-1} \right) \mathbb{E}||u_m^{(r,K)} - w_m^{(r,k)}||^2 + K\eta_l^2 \mathbb{E}||\nabla \psi_m^{(r,k)} f_m(w_m^{(r,k)})||^2 \right) \] (150)
\[
\leq \left( 6K^4 \eta_l^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K^3 \eta_l^2 \sigma_t^2 + 12K^2 \eta_l^2 \sigma_t^2 \mathbb{E}||1 - \psi_m^{(r,k)}||^2 \right) \right) \left( 1 + \frac{1}{K-1} \right) \frac{K}{K-1} \left( 1 + \frac{1}{K-1} \right) \] (151)
Assuming \( \frac{1}{\sqrt{K}} \geq 12K^2 \eta_l^2 \beta \Rightarrow \eta_l \leq \frac{1}{K \sqrt{12(K-1)}} \),
\[
\mathbb{E}||w_m^{(r,k)} - w_m^{(r,0)}||^2 \leq \left( 6K^4 \eta_l^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 + 3K^3 \eta_l^2 \sigma_t^2 + 12K^2 \eta_l^2 \mathbb{E}||1 - \psi_m^{(r,k)}||^2 \right) \] (152)
\[
+ 4 \left( 1 + \frac{1}{K-1} \right) K^3 \eta_l^2 \mathbb{E}||\nabla f_m(w_g^{(r)})||^2 \mathbb{E}||\psi_m^{(r,k)}||^2 \sigma_t^2 \right) \] (153)

\[ \text{Lemma D.19 (One round progress of personalized model). If } m^{th} \text{ client’s objective function } f_m \text{ satisfies } \]
Assumptions D.3, D.4 in Algorithm D.4, the following is satisfied:
\[
\mathbb{E}||w_m^{(r,K)} - w_m^{(r,0)}||^2 \leq 72(1 + \eta_l^2) K^2 \eta_l^2 \left( 5K(G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2) + 12\sigma_t^2 \right) \]
\[
+ 36 \left( K \sigma_t^2 + \delta_m^\psi + \delta_m^\psi M \right) K^2 \eta_l^2 \left( K \sigma_t^2 + \delta_m^\psi + \delta_m^\psi M \right) K^2 \eta_l^2 \]
\[
+ 12(1 + \eta_l^2) K^2 \beta^4 \delta_t^\gamma \left( K \sigma_t^2 + \delta_m^\psi + \delta_m^\psi M \right) K^2 \eta_l^2 \left( G^2 + B^2 \mathbb{E}||\nabla F(w_g^{(r)})||^2 \right) \]
\[
\text{Proof.} \]
\[
\mathbb{E}||w_m^{(r,K)} - w_m^{(r,0)}||^2 = \mathbb{E}||w_m^{(r+1,K)} - w_m^{(r+1,0)} + w_m^{(r+1,0)} - w_m^{(r,K)}||^2 \] (154)
\[
\leq 2\mathbb{E}||w_m^{(r+1,K)} - w_m^{(r+1,0)}||^2 + 2\mathbb{E}||w_m^{(r+1,0)} - w_m^{(r,K)}||^2 \] (155)

38
Using the Lemmas D.18 and D.14 we proceed as

\[
\leq 72(1 + \eta_l^2)K^3\eta_l^2 \left( 5K(G^2 + B^2\mathbb{E}\|\nabla F(w_y^{(r)})\|^2) + 12\sigma_l^2 \right) \\
+ 36 \left( K\sigma_l^2\eta_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K^2\eta_l^2 \right) \left( K\sigma_l^2\eta_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K^2\eta_l^2 \right) \\
+ 12(1 + \eta_l^2)K^2\beta^4\eta_l^2 \left( K\sigma_l^2\eta_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K^2\eta_l^2 \right) \left( G^2 + B^2\mathbb{E}\|\nabla F(w_y^{(r)})\|^2 \right)
\]

(156)

**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 \( \eta \leq \frac{1}{K\sqrt{t}} \) in Algorithm 2, then the following convergence holds:

\[
\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}\|\nabla f_m(w_{r,m}^{(R)})\|^2 \leq \frac{2}{R} \left( \mathbb{E} \left[ f_m(w_{1,m}^{(R)}) \right] - \mathbb{E} \left[ f_m(w_{R,m}^{(R)}) \right] \right) \\
+ 6(1 + \eta)^2 K \left( 5K(G^2 + B^2) \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}\|\nabla F(w_y^{(r)})\|^2 \right) + 12\sigma_l^2 \\
+ 3K\eta_l^2 \left( \sigma_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K \right) \left( \sigma_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K \right) \\
+ (1 + \eta_l^2)K^2\beta^4\eta_l^2 \left( \sigma_l^2 + \left( \frac{\delta_m}{M} + \frac{\delta_y}{M} \right) K \right) \left( G^2 + B^2 \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}\|\nabla F(w_y^{(r)})\|^2 \right)
\]

(157)

**Proof.** According to the update rule of Equation 10 and \( \beta \)-smoothness of \( f_m \), we have,

\[
f_m(w_{r+1,m}^{(r+1,K)}) \leq f_m(w_{r,m}^{(r,K)}) + \langle \nabla f_m(w_{r,m}^{(r,K)}), w_{r+1,m}^{(r+1,K)} - w_{r,m}^{(r,K)} \rangle + \frac{\beta}{2} \|w_{r+1,m}^{(r+1,K)} - w_{r,m}^{(r+K)}\|^2
\]

(158)

Taking expectation on both sides,

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

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

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

(159)

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

(160)

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

(161)

(162)
Rearranging the terms to put $\frac{1}{2} \mathbb{E}[\|\nabla f_m(u_{p,m}^{(r,K)})\|^2]$ at LHS,

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

(163)

$$
\mathbb{E}[\|\nabla f_m(u_{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^2) K^3 \eta^2 \left( 5K(G^2 + B^2 \mathbb{E}[\|\nabla F(w_y^{(r)})\|^2]) + 12 \sigma^2 \right)
+ 36 \beta \left( K \sigma_i^2 \eta_i + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K^2 \eta_i \right)
+ 12 \beta (1 + \eta^2 K^2 \beta^4) \eta_i K^3 \left( \sigma_i^2 + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K \right)

(164)

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

$$
\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}[\|\nabla f_m(u_{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^2) K^3 \eta^2 \left( 5K(G^2 + B^2 \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}[\|\nabla F(w_y^{(r)})\|^2]) + 12 \sigma^2 \right)
+ 36 \beta K^3 \eta_i \left( \sigma_i^2 + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K \right)
+ 12 \beta (1 + \eta^2 K^2 \beta^4) \eta_i K^3 \left( \sigma_i^2 + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K \right)

(165)

Assuming $12 K^2 \eta_i^2 \beta \leq 1 \leq \frac{1}{K \sqrt{27}} \implies \eta_i \leq \frac{1}{K \sqrt{27}}.$

$$
\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^2) K \left( 5K(G^2 + B^2 \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}[\|\nabla F(w_y^{(r)})\|^2]) + 12 \sigma^2 \right)
+ 3K \eta^2 \left( \sigma_i^2 + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K \right)
+ (1 + \eta^2 K^2 \beta^4) \eta_i K \left( \sigma_i^2 + \left( \delta_m^2 + \frac{\delta^v}{M} \right) K \right)

(166)

Plugging in Theorem $\textbf{D.12}$ to get bounds on $\sum_{r=1}^{R} \mathbb{E}[\|\nabla F(w_y^{(r)})\|^2]$ would get us bounds on

$$
\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}[\|\nabla f_m(u_{p,m}^{(r,K)})\|^2].$$