Submodel Extraction for Efficient and Personalized Federated Learning via Optimal Transport

Supplementary Material

8. Experimental Setup

8.1. Datasets

Our evaluation utilizes a diverse suite of seven datasets to ensure a comprehensive assessment of our method's performance across various data modalities, including computer vision (CV), natural language processing (NLP), and Internet of Things (IoT) sensor data. The benchmark includes standard single-domain datasets such as CIFAR-10/100, Tiny-ImageNet, AG News, and HAR. To specifically evaluate robustness against feature distribution shifts, we also incorporate two multi-domain datasets: **Digit-5**, comprising five distinct handwritten digit domains (MNIST, MNIST-M, USPS, SVHN, and SYN), and **PACS**, which includes four artistic domains (Photo, Art, Cartoon, and Sketch). Key statistics for each dataset are summarized in Table 9.

8.2. Data Partition

To simulate realistic Federated Learning (FL) environments, we construct three distinct statistical heterogeneity scenarios [5, 24]. For the *label skew* scenario, we implement two common settings: the pathological setting and the practical setting [32]. For the *pathological label skew*, we sample data with label amount 2/10/20 for each client on Cifar10/Cifar100/TinyImageNet from a total of 10/100/200 categories. For the *practical label skew*, we employ a Dirichlet distribution (default $\beta = 0.1$) to generate realistic partially-overlapping class distributions for Cifar10, Cifar100, TinyImageNet and AG News. For the *feature shift* scenario, We utilize the Digit5 (5 domains) and PACS (4 domains) datasets. Each client participating in the FL system is assigned data from one of these distinct domains. Finally, we use the HAR dataset to represent a *real-world* scenario, which provides a natural partitioning of sensor data from 30 users performing six activities.

9. Method Details

Algorithm 1 delineates the procedural details of our Optimal Transport-based Pruning (OTP) module. The Optimal Transport-enhanced Aggregation (OTA) module then adapts this layer-wise mechanism, performing a conceptually inverse operation to map the updated client submodels back into the global parameter space for aggregation.

To accommodate modern network architectures, we incorporate specialized handling for specific layer types.

- **Residual Blocks:** To maintain the continuity of the transport map across skip connections, the transport matrices derived from parallel branches are averaged prior to propagation to subsequent layers.
- **Batch Normalization Layers:** Since these layers perform channel-wise normalization without altering the dimensional permutation of the feature space, the incoming transport matrices are passed through without modification.

The complete end-to-end workflow of the SubFLOT framework is summarized in Algorithm 2.

Algorithm 1: Optimal Transport-based Parameter Alignment and Fusion (OTP)

Input: Global model M_G, client model M_i, number of layers L, fusion ratio α
Output: Personalized fused model $\widetilde{M}_i$
for *each client* $i \in \{1, 2, \ldots, N\}$ **do**
  Initialize transport matrix for the input layer: $T_i^{(0)} = I$;
  for *each layer* $l \in \{1, 2, \ldots, L\}$ **do**
    $\widehat{W}_G^{(l,l-1)} = W_G^{(l,l-1)} T_i^{(l-1)}$;
    $C_{jk}^{(l)} = \|\widehat{W}_G^{(l,l-1)}[j] - W_i^{(l,l-1)}[k]\|$;
    $T_i^{(l)} = \arg\min_T \langle C^{(l)}, T \rangle_F \quad \text{s.t.} \quad T\mathbf{1}_{|\nu|} = \mu,\ T^\top \mathbf{1}_{|\mu|} = \nu$;
    $T_i^{(l)} = T_i^{(l)} \odot \left(\frac{1}{\mathbf{1}_m^\top T_i^{(l)}} \right)$;
    $\widetilde{W}_{\text{aligned}}^{(l,l-1)} = {T_i^{(l)}}^\top \widehat{W}_G^{(l,l-1)}$;
    $\widetilde{W}_i^{(l,l-1)} = \alpha \cdot \widetilde{W}_{\text{aligned}}^{(l,l-1)} + (1-\alpha) \cdot W_i^{(l,l-1)}$;
  end
  Construct the personalized model $\widetilde{M}_i$ from the fused weights $\{\widetilde{W}_i^{(l,l-1)}\}_{l=1}^L$;
end
return *Personalized models* $\{\widetilde{M}_i\}_{i=1}^N$;

10. Additional Experimental Results

10.1. Server-Side Latency and Scalability Analysis

A major concern regarding our framework is whether the Optimal Transport (OT) computations introduced by SubFLOT incur prohibitive server-side latency. To address this concern, we conducted a wall-clock time analysis on CIFAR-10, as summarized in Fig. 3. We report two complementary metrics: (i) the total elapsed time required to complete a fixed number of communication rounds, and (ii)

Table 9. Dataset Specifications. Our evaluation covers diverse data modalities and includes multi-domain benchmarks (Digit-5, PACS) to rigorously assess model generalization and robustness against feature distribution shifts.

Dataset	Classes	Training Samples	Test Samples	Domains	Modality
CIFAR-10	10	50,000	10,000	1	CV
CIFAR-100	100	50,000	10,000	1	CV
Tiny-ImageNet	200	100,000	10,000	1	CV
Digit-5	10	130,288	32,572	5	CV
PACS	7	7,988	2,003	4	CV
AG News	4	120,000	7,600	1	NLP
HAR	6	7,352	2,947	1	IoT Sensor

Algorithm 2: SubFLOT: Federated Submodel Learning via Optimal Transport

Input: Communication rounds T, local epochs R, client number N, learning rate η, fusion ratio α, regularization factor λ, local dataset D_i and pruning ratio ρ_i for each client i

Initialize global model W_G^0 and client submodels $\{W_i^0\}_{i=1}^N$;

for *each communication round* $t \in \{1, \ldots, T\}$ **do**

  for *each client* $i \in \{1, \ldots, N\}$ ***in parallel*** **do**

    $\widetilde{W}_i^t \leftarrow \text{OTP}(W_G^{t-1}, W_i^{t-1}, \alpha)$ `// See Algorithm 1;`

    $W_i^t \leftarrow \text{ClientUpdate}(i, \widetilde{W}_i^t, D_i, R, \eta)$;

  end

  $W_G^t \leftarrow \text{OTA}(\{W_i^t\}_{i=1}^N)$;

end

Procedure *ClientUpdate(i,* $\widetilde{W}_i^t$*,* D_i*, R,* η*)*

  for *each local epoch* $r \in \{1, \ldots, R\}$ **do**

    for *each batch* $(x, y) \in D_i$ **do**

      Update weights of $\widetilde{W}_i^t$ using SGD:

      $\widetilde{W}_i^t \leftarrow \widetilde{W}_i^t - \eta \nabla \mathcal{L}_i(\widetilde{W}_i^t; x, y)$;

    end

  end

  return *Updated local model* $\widetilde{W}_i^t$;

the total elapsed time required to reach a target test accuracy (80% in our experiment).

The results indicate that the additional OT-related computation constitutes only a modest fraction of the total training time. More importantly, this overhead is compensated by the improved optimization efficiency of SubFLOT, which reaches a desirable accuracy level in fewer effective rounds and with a better overall time-to-accuracy trade-off. In other words, although OT introduces extra computation on the server, the resulting gains in convergence behavior make the full training process more efficient from a practical deployment perspective.

This observation is consistent with the intended design of SubFLOT. First, OT is executed on the server rather than on clients, which is desirable in federated settings because the server generally has substantially stronger computational resources than edge devices. Second, the OT computations for different clients are naturally parallelizable, since each client-specific transport plan can be computed independently. Therefore, in a realistic distributed infrastructure, the latency introduced by OT can be further amortized through parallel processing. Third, unlike client-side personalized pruning methods that require full-model training before pruning, our server-side personalization mechanism avoids imposing heavy burdens on resource-constrained clients, which is often a more critical bottleneck in practical FL systems.

Taken together, these results suggest that the overhead of OT is not a limiting factor for SubFLOT in realistic federated deployments. Instead, the method achieves a favorable balance between server-side computation and end-to-end training efficiency.

10.2. Hyperparameter Sensitivity and Practical Tuning Guidelines

We further investigate the sensitivity of SubFLOT to its two key hyperparameters, namely the fusion coefficient α used in OTP and the adaptive regularization coefficient λ used in SAR. The corresponding results on multiple datasets and under different non-IID settings are shown in Fig. 4.

Effect of α. The parameter α controls the extent to which the server-side personalized pruning process is influenced by the historical client model. Empirically, we observe that a relatively larger value of α can accelerate convergence in moderately heterogeneous scenarios, as it enables stronger client-specific adaptation. However, under severe statistical heterogeneity, excessive reliance on historical client information may amplify instability, especially when local distributions drift substantially from one another. In such cases, a smaller α often yields more stable optimization by preserving a stronger anchor to the global

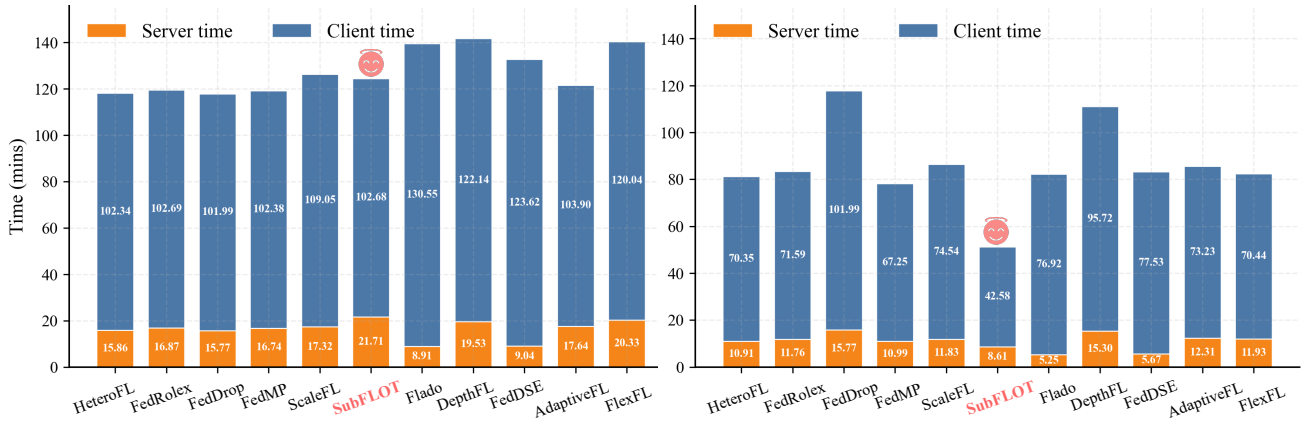

Figure 3. Comparison of total wall-clock time required to complete 200 communication rounds (left) and to reach 80% test accuracy (right) on CIFAR-10. Although OT introduces additional server-side computation, SubFLOT achieves a superior time-to-accuracy trade-off due to faster convergence.

model. This phenomenon is especially evident on challenging benchmarks such as Digit5, where domain gaps are more pronounced.

Effect of λ. The regularization coefficient λ controls the strength of the Scaling-based Adaptive Regularization (SAR) term. We find that increasing λ is generally beneficial for heavily pruned submodels, because stronger regularization can effectively suppress pruning-induced parametric divergence and stabilize local optimization. Nevertheless, overly large values may over-constrain local training and weaken the ability of clients to adapt to their own data distributions, resulting in reduced personalization performance.

Practical tuning strategy. Based on these observations, we recommend a simple yet effective tuning rule in practice:

- **Tune α inversely with statistical heterogeneity.** When data distributions are highly non-IID, a smaller α is preferred to improve robustness and avoid overfitting to unstable historical representations.
- **Tune λ proportionally with system heterogeneity.** When pruning rates vary widely across clients or when some clients are assigned highly sparse submodels, a larger λ is helpful for controlling the resulting parametric drift.

In our main experiments, the default settings $\alpha = 0.5$ and $\lambda = 1.0$ provide a robust trade-off across datasets and non-IID conditions.

10.3. Necessity of OT-Enhanced Aggregation

The motivation stems from the well-known permutation invariance of deep neural networks. Two subnetworks may realize highly similar functions while arranging semantically similar neurons or channels at different indices. This mismatch becomes especially pronounced in federated settings with heterogeneous data distributions and personalized pruning patterns. As a result, directly averaging client parameters without alignment may lead to severe neuron mismatch, thereby degrading aggregation quality.

OTP addresses personalization by aligning the global model to the client's local feature space before training. However, different clients' feature spaces are themselves not necessarily aligned with one another. Consequently, even if each client receives a personalized submodel that is suitable for local optimization, their updated parameters may reside in distinct local coordinate systems after training. OTA is therefore essential for mapping these heterogeneous local updates back into a shared global canonical space before aggregation.

This perspective also clarifies the boundary conditions of OTA. The alignment may deteriorate under extreme scenarios, such as abrupt distribution shifts, excessively aggressive local training, or very high pruning ratios, where meaningful geometric correspondence between subnetworks is severely weakened. Nevertheless, within the practical operating regime considered in this paper, OTA provides a principled mechanism to alleviate neuron mismatch and improve aggregation reliability.

10.4. Ablation on the Choice of Proxy for OTP

To validate the necessity of using the historical client model as the proxy in OTP, we performed an ablation study in which the historical model was replaced by alternative pruning references. Specifically, we compared the following options:

- **Historical**: the historical client model used in SubFLOT;
- **Fixed-Pos.**: a deterministic pruning pattern based on fixed positions;
- **Magnitude**: conventional magnitude-based pruning;

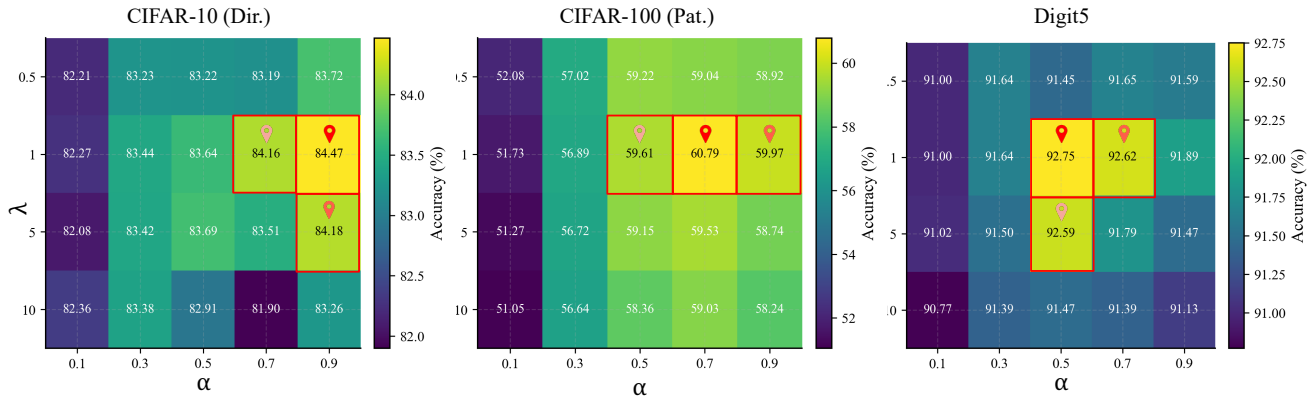

Figure 4. Hyperparameter sensitivity analysis of SubFLOT on multiple datasets and heterogeneity settings. The results show that α mainly controls the personalization–stability trade-off, while λ primarily regulates pruning-induced parametric divergence.

- **Random**: random pruning.

The results are reported in Table 10. Using the historical model achieves the best performance by a substantial margin. In contrast, replacing it with heuristic or non-personalized alternatives leads to clear degradation, and random pruning performs particularly poorly. These results confirm that the historical model indeed encodes client-specific information that is useful for personalization, thereby supporting the central design of OTP.

Table 10. Ablation on the proxy used for OTP. The historical client model provides the most effective client-specific prior.

Proxy	**Historical**	**Fixed-Pos.**	**Magnitude**	**Random**
Accuracy	**83.78**	74.42	77.04	49.82

10.5. Stability Under Different Pruning Rates

To further evaluate robustness, we report the performance of SubFLOT under varying pruning rates on CIFAR-10. As shown in Fig. 5, SubFLOT consistently outperforms competing methods across a broad range of sparsity levels and remains relatively stable as pruning becomes more aggressive.

This result is particularly important because increasing pruning rates generally amplifies heterogeneity in both architecture and parameter scale. The observed stability of SubFLOT indicates that the combination of OTP, OTA, and SAR effectively mitigates the adverse effects of severe submodel sparsification. In particular, SAR plays a key role in stabilizing highly pruned clients, while OTA improves the consistency of aggregation when clients return updates from increasingly dissimilar subnetworks.

10.6. Comparative Analysis with pFL Methods

We provide a comparative analysis of SubFLOT against several prominent personalized Federated Learning (pFL) methods. It is crucial to frame this comparison within the appropriate context: SubFLOT's design philosophy is fundamentally orthogonal to that of most pFL baselines. Whereas methods like Per-FedAvg [13], pFedMe [11], FedAMP [20], and FedFomo [46] primarily aim to maximize personalization accuracy by adapting the full global model on client-specific data, SubFLOT's core objective is to enable efficient training on resource-constrained devices through the extraction of personalized submodels. Consequently, these pFL baselines operate on full-sized models, incurring substantial computational and communication costs comparable to FedAvg.

Despite this fundamental difference and operating with significantly fewer resources (over 50% reduction in both computation and communication), our empirical results presented in Table 11 demonstrate that SubFLOT achieves performance that is highly competitive with, and in some cases superior to, these state-of-the-art pFL methods. For example, on CIFAR-10, SubFLOT outperforms all listed baselines. On CIFAR-100 and Tiny-ImageNet, it secures the second-best performance, narrowly trailing the top-performing methods while operating at a fraction of the resource cost. This outcome highlights a remarkable dual benefit of our framework: SubFLOT not only delivers top-tier personalization but does so while satisfying the stringent resource constraints of practical federated systems. It effectively resolves the trade-off between personalization and efficiency, offering a holistic solution that excels in both dimensions.

10.7. More Feature Visualization Cases

Figure 6 provides a qualitative analysis of model attention through activation maps, comparing SubFLOT against var-

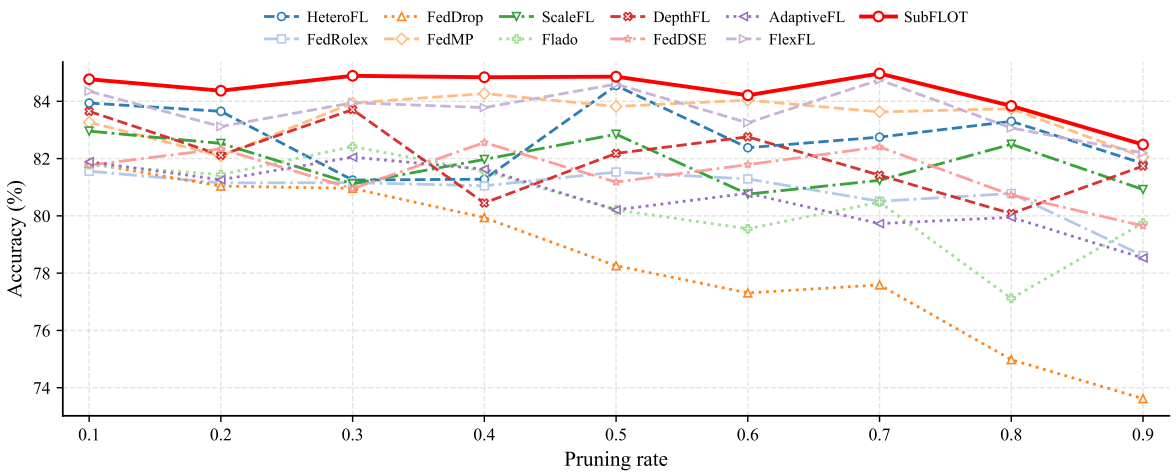

Figure 5. Performance comparison under varying pruning rates on CIFAR-10. SubFLOT maintains strong accuracy and stability even at high sparsity levels.

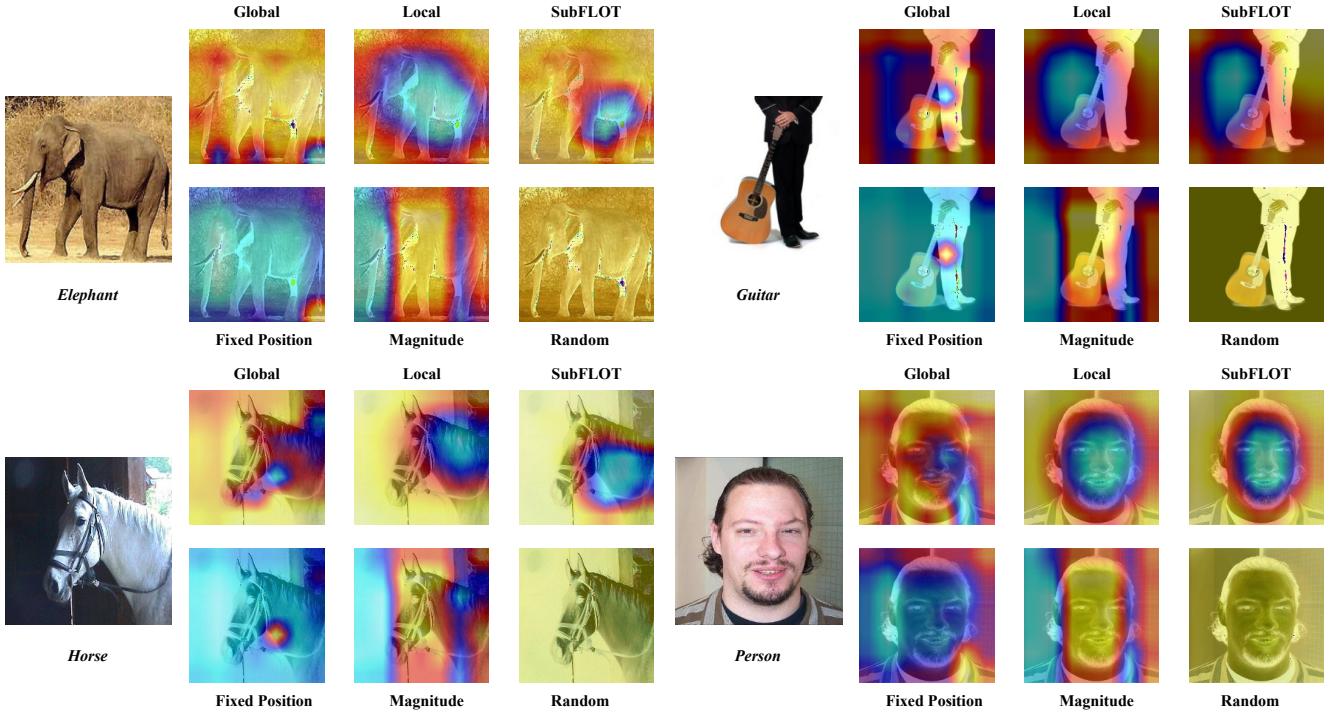

Figure 6. Qualitative comparison of activation maps generated by SubFLOT and baseline methods. SubFLOT successfully preserves the task-relevant attention patterns of the local model (column 2), demonstrating effective feature alignment. In contrast, fixed-position pruning fails to adapt to local data, while magnitude-based and random pruning exhibit fragmented or noisy attention, indicating a misalignment with client-specific features.

ious baselines. The visualizations reveal that SubFLOT-generated submodels exhibit remarkable spatial attention similarity to their corresponding local models from the previous round. This indicates that our optimal transport-based alignment strategy effectively preserves task-relevant features while adapting to local data distributions, confirming that OTP successfully establishes a geometrically meaningful mapping between the global and local parameter spaces.

In contrast, the baseline methods demonstrate clear limitations. **Fixed-position** pruning maintains strong consis-

Table 11. Accuracy comparison with personalized FL baselines.

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
Per-FedAvg	82.74	43.28	24.07
pFedMe	83.19	**45.36**	24.93
FedAMP	<u>83.68</u>	44.69	**25.99**
FedFomo	83.06	44.33	23.33
SubFLOT	**83.78**	<u>44.88</u>	<u>25.15</u>

tency with the global model's activation patterns but fails to adapt to client-specific feature importance, leading to suboptimal performance on heterogeneous data. **Magnitude-based** pruning preserves core discriminative regions but requires substantial fine-tuning to align with local data, as evidenced by the fragmented attention maps. Finally, **random** pruning introduces significant noise into the feature representations, manifesting as scattered, incoherent attention patterns that diverge considerably from both the global and local models.

11. Proof of Theorem 1

The proof proceeds by bounding the one-step progress of the global model and then recursively applying the result over T rounds.

Lemma 1 (Bounded Local Client Drift)**.** *Under Assumptions 3 and 4, after E local steps with learning rate $\eta_l \leq \frac{1}{4\lambda\rho_{\max}}$, the expected squared distance between a client's updated model W_i^t and its personalized anchor $\widetilde{W}_i^t$ is bounded by:*

$$\mathbb{E}[\left\|W_i^t - \widetilde{W}_i^t\right\|^2] \leq \frac{4\,\eta_l\,E}{\lambda\rho_i}\,(\sigma^2 + G^2). \tag{12}$$

Proof. Define $U_i^{t,k} := W_i^{t,k} - \widetilde{W}_i^t$, so that $U_i^{t,0} = 0$. Let $g_i^{t,k} := \nabla\mathcal{L}_i(W_i^{t,k};\xi_i^k)$. The local update rule can be rewritten as

$$\begin{aligned} U_i^{t,k+1} &= U_i^{t,k} - \eta_l\Big(\nabla\mathcal{L}_i(W_i^{t,k};\xi_i^k) + 2\lambda\rho_i U_i^{t,k}\Big) \\ &= (1 - 2\lambda\rho_i\eta_l)\,U_i^{t,k} - \eta_l g_i^{t,k}. \end{aligned} \tag{13}$$

Taking squared norms and conditioning on $U_i^{t,k}$,

$$\begin{aligned} &\mathbb{E}_{\text{loc}}\big[\|U_i^{t,k+1}\|^2 \mid U_i^{t,k}\big] \\ &= (1 - 2\lambda\rho_i\eta_l)^2\|U_i^{t,k}\|^2 + \eta_l^2\mathbb{E}_{\text{loc}}\big[\|g_i^{t,k}\|^2\big] \\ &\leq (1 - 4\lambda\rho_i\eta_l)\|U_i^{t,k}\|^2 + \eta_l^2(\sigma^2 + G^2), \end{aligned} \tag{14}$$

where we used $(1-2a)^2 = 1 - 4a + 4a^2 \leq 1 - 4a$ for $a \in [0, 1/2]$ and the condition $\eta_l \leq 1/(4\lambda\rho_{\max})$ to guarantee $2\lambda\rho_i\eta_l \leq 1/2$, together with

$$\begin{aligned} \mathbb{E}_{\text{loc}}\|g_i^{t,k}\|^2 &= \mathbb{E}_{\text{loc}}[\left\|\nabla\mathcal{L}_i(W_i^{t,k};\xi_i^k)\right\|^2] \\ &= \mathbb{E}_{\text{loc}}[\left\|\nabla\mathcal{L}_i(W_i^{t,k};\xi_i^k) - \nabla\mathcal{F}_i(W_i^{t,k}) + \nabla\mathcal{F}_i(W_i^{t,k})\right\|^2] \\ &\leq \mathbb{E}_{\text{loc}}[\left\|\nabla\mathcal{L}_i(W_i^{t,k};\xi_i^k) - \nabla\mathcal{F}_i(W_i^{t,k})\right\|^2] + \left\|\nabla\mathcal{F}_i(W_i^{t,k})\right\|^2 \\ &\leq \sigma^2 + G^2, \end{aligned} \tag{15}$$

from Assumptions 3 and 4.

Define $V_k := \mathbb{E}_{\text{loc}}\|U_i^{t,k}\|^2$ and take full expectation over all randomness. We obtain

$$V_{k+1} \leq (1 - 4\lambda\rho_i\eta_l)V_k + \eta_l^2(\sigma^2 + G^2). \tag{16}$$

Unrolling from $V_0 = 0$ gives

$$\begin{aligned} V_E &\leq \eta_l^2(\sigma^2 + G^2)\sum_{j=0}^{E-1}(1 - 4\lambda\rho_i\eta_l)^j \\ &\leq \eta_l^2(\sigma^2 + G^2)\cdot\frac{1 - (1 - 4\lambda\rho_i\eta_l)^E}{4\lambda\rho_i\eta_l} \\ &\leq \eta_l^2(\sigma^2 + G^2)\cdot\frac{E\cdot 4\lambda\rho_i\eta_l}{4\lambda\rho_i\eta_l} = \eta_l E(\sigma^2 + G^2) \\ &\leq \frac{4\,\eta_l\,E}{\lambda\rho_i}(\sigma^2 + G^2). \end{aligned} \tag{17}$$

where we used $1 - (1-x)^E \leq Ex$ for $x \in [0,1]$ with $x = 4\lambda\rho_i\eta_l$ and the last inequality is a benign loosening (since $\lambda\rho_i > 0$ and we only need an explicit linear dependence on $1/(\lambda\rho_i)$). This proves the claim. □

Proof of Theorem 1. From the L-smoothness of the global objective function $\mathcal{F}$ (Assumption 1), we have:

$$\begin{aligned} \mathbb{E}[\mathcal{F}(W_G^{t+1})] \leq \mathbb{E}[\mathcal{F}(W_G^t)] + \mathbb{E}[\langle\nabla\mathcal{F}(W_G^t), W_G^{t+1} - W_G^t\rangle] \\ + \frac{L}{2}\mathbb{E}[\left\|W_G^{t+1} - W_G^t\right\|^2]. \end{aligned} \tag{18}$$

Let us define the virtual average model before alignment as $\bar{W}^t = \sum_{i=1}^N p_i W_i^t$. The global model update can be decomposed as:

$$\begin{aligned} W_G^{t+1} - W_G^t &= \sum_{i=1}^N p_i(\mathcal{T}_i(W_i^t) - W_G^t) \\ &= \underbrace{(\bar{W}^t - W_G^t)}_{\text{Average Local Update}} + \underbrace{\sum_{i=1}^N p_i(\mathcal{T}_i(W_i^t) - W_i^t)}_{\text{OTA Perturbation}}. \end{aligned} \tag{19}$$

We first analyze the inner product term in (18). Taking expectations and using (19),

$$\mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), W_G^{t+1} - W_G^t \rangle\big] = \mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), \bar{W}^t - W_G^t \rangle\big] + \mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), \sum_{i=1}^{N} p_i (\mathcal{T}_i(W_i^t) - W_i^t) \rangle\big]. \tag{20}$$

For the OTA term, Young's inequality and Assumption 5 give

$$\begin{aligned}
&\mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), \sum_{i=1}^{N} p_i (\mathcal{T}_i(W_i^t) - W_i^t) \rangle\big] \\
&\le \frac{1}{2}\mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] + \frac{1}{2}\mathbb{E}\Big[\Big\|\sum_{i=1}^{N} p_i (\mathcal{T}_i(W_i^t) - W_i^t)\Big\|^2\Big] \\
&\le \frac{1}{2}\mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] + \frac{\delta_{OT}^2}{2}.
\end{aligned} \tag{21}$$

We now relate $\bar{W}^t - W_G^t$ to $\nabla \mathcal{F}(W_G^t)$. Client i performs E local steps starting from $\widetilde{W}_i^t$:

$$W_i^{t,k+1} = W_i^{t,k} - \eta_l \Big(\nabla \mathcal{L}_i(W_i^{t,k}; \xi_i^k) + 2\lambda \rho_i (W_i^{t,k} - \widetilde{W}_i^t) \Big) \tag{22}$$

with $W_i^{t,0} = \widetilde{W}_i^t$ and $W_i^{t,E} = W_i^t$. Summing over $k = 0, \dots, E-1$ and taking expectation conditional on W_G^t,

$$\mathbb{E}\big[W_i^t - \widetilde{W}_i^t \mid W_G^t\big] = -\eta_l \sum_{k=0}^{E-1} \mathbb{E}\big[\nabla \mathcal{F}_i(W_i^{t,k}) + 2\lambda \rho_i (W_i^{t,k} - \widetilde{W}_i^t) \mid W_G^t\big]. \tag{23}$$

Add and subtract $\nabla \mathcal{F}_i(W_G^t)$:

$$\begin{aligned}
&\Big\|\mathbb{E}\big[W_i^t - \widetilde{W}_i^t \mid W_G^t\big] + \eta_l E \nabla \mathcal{F}_i(W_G^t)\Big\| \\
&\le \eta_l \sum_{k=0}^{E-1} \mathbb{E}\big[\|\nabla \mathcal{F}_i(W_i^{t,k}) - \nabla \mathcal{F}_i(W_G^t)\| + 2\lambda \rho_i \|W_i^{t,k} - \widetilde{W}_i^t\| \mid W_G^t\big].
\end{aligned} \tag{24}$$

Using L-smoothness,

$$\begin{aligned}
&\|\nabla \mathcal{F}_i(W_i^{t,k}) - \nabla \mathcal{F}_i(W_G^t)\| \le L \|W_i^{t,k} - W_G^t\| \\
&\le L(\|W_i^{t,k} - \widetilde{W}_i^t\| + \|\widetilde{W}_i^t - W_G^t\|).
\end{aligned} \tag{25}$$

Applying Jensen's inequality, Lemma 1, and Assumption 6,

$$\begin{aligned}
\mathbb{E}\big[\|W_i^{t,k} - \widetilde{W}_i^t\|\big] &\le \sqrt{\mathbb{E}\big[\|W_i^{t,k} - \widetilde{W}_i^t\|^2\big]} \\
&\le \sqrt{\frac{4\,\eta_l\,E}{\lambda \rho_i}}\,\sqrt{\sigma^2 + G^2}, \\
\mathbb{E}\big[\|\widetilde{W}_i^t - W_G^t\|\big] &\le \sqrt{\mathbb{E}\big[\|\widetilde{W}_i^t - W_G^t\|^2\big]} \le \delta_P.
\end{aligned} \tag{26}$$

Hence there exists a deterministic upper bound

$$\begin{aligned}
&\Big\|\mathbb{E}\big[W_i^t - \widetilde{W}_i^t \mid W_G^t\big] + \eta_l E \nabla \mathcal{F}_i(W_G^t)\Big\| \\
&\le \eta_l E \Big(2L \sqrt{\tfrac{4\eta_l E}{\lambda \rho_i}}\,\sqrt{\sigma^2 + G^2} + 2L\delta_P \\
&\quad + 4\lambda \rho_i \sqrt{\tfrac{4\eta_l E}{\lambda \rho_i}}\,\sqrt{\sigma^2 + G^2} \Big),
\end{aligned} \tag{27}$$

which is $O(\eta_l E)$ in η_l and E. Summing over i with weights p_i and using $\nabla \mathcal{F}(W_G^t) = \sum_i p_i \nabla \mathcal{F}_i(W_G^t)$, we obtain

$$\Big\|\mathbb{E}\big[\bar{W}^t - W_G^t \mid W_G^t\big] + \eta_l E \nabla \mathcal{F}(W_G^t)\Big\| \le \eta_l E\, \beta, \tag{28}$$

for some deterministic $\beta > 0$, depending only on L, σ, G, λ, $\rho_{\max}$, δ_P. Thus,

$$\begin{aligned}
&\mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), \bar{W}^t - W_G^t \rangle\big] \\
&= -\eta_l E\, \mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] \\
&\quad + \mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), (\bar{W}^t - W_G^t) + \eta_l E \nabla \mathcal{F}(W_G^t) \rangle\big] \\
&\le -\eta_l E\, \mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] + \eta_l E\, \mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\| \beta\big] \\
&\le -\frac{\eta_l E}{2}\mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] + \frac{\eta_l E}{2}\beta^2,
\end{aligned} \tag{29}$$

where we used $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$.

Combining (21) and (29),

$$\begin{aligned}
&\mathbb{E}\big[\langle \nabla \mathcal{F}(W_G^t), W_G^{t+1} - W_G^t \rangle\big] \\
&\le -\frac{\eta_l E}{2}\mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] + \frac{1}{2}\mathbb{E}\big[\|\nabla \mathcal{F}(W_G^t)\|^2\big] \\
&\quad + \frac{\delta_{OT}^2}{2} + \frac{\eta_l E}{2}\beta^2.
\end{aligned} \tag{30}$$

Next, we bound the squared norm term from (18):

$$\begin{aligned}
\mathbb{E}\big[\|W_G^{t+1} - W_G^t\|^2\big] &\le 2\,\mathbb{E}\big[\|\bar{W}^t - W_G^t\|^2\big] \\
&\quad + 2\,\mathbb{E}\Big[\Big\|\sum_{i=1}^{N} p_i (\mathcal{T}_i(W_i^t) - W_i^t)\Big\|^2\Big] \\
&\le 2\,\mathbb{E}_i\big[\|W_i^t - W_G^t\|^2\big] + 2\delta_{OT}^2 \\
&\le 4\,\mathbb{E}_i\big[\|W_i^t - \widetilde{W}_i^t\|^2\big] + 4\,\mathbb{E}_i\big[\|\widetilde{W}_i^t - W_G^t\|^2\big] + 2\delta_{OT}^2.
\end{aligned} \tag{31}$$

Applying Lemma 1 and Assumption 6, and using $\rho_i \ge \min_j \rho_j \ge \rho_{\max}/2$ without loss of generality, we have

$$\begin{aligned}
&\mathbb{E}\big[\|W_G^{t+1} - W_G^t\|^2\big] \\
&\le 4\,\frac{4\eta_l E}{\lambda \rho_{\max}}(\sigma^2 + G^2) + 4\delta_P^2 + 2\delta_{OT}^2 \\
&= 16\eta_l E (\lambda \rho_{\max})^{-1}(\sigma^2 + G^2) + 4\delta_P^2 + 2\delta_{OT}^2.
\end{aligned} \tag{32}$$

Thus

$$\begin{aligned}
&\frac{L}{2}\,\mathbb{E}\big[\|W_G^{t+1} - W_G^t\|^2\big] \\
&\le 8LE\eta_l (\lambda \rho_{\max})^{-1}(\sigma^2 + G^2) + 2L\delta_P^2 + L\delta_{OT}^2.
\end{aligned} \tag{33}$$

Substituting (30) and (33) into (18),

$$\begin{aligned}&\mathbb{E}\big[\mathcal{F}(W_G^{t+1})\big]\\&\le \mathbb{E}\big[\mathcal{F}(W_G^t)\big] - \frac{\eta_l E}{2}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big]\\&+ \frac{1}{2}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big] + \frac{\delta_{OT}^2}{2} + \frac{\eta_l E}{2}\beta^2\\&+ 8LE\eta_l(\lambda\rho_{\max})^{-1}(\sigma^2+G^2) + 2L\delta_P^2 + L\delta_{OT}^2.\end{aligned} \tag{34}$$

Using $\eta_l \le 1/(8E)$, then we have

$$\begin{aligned}&-\frac{\eta_l E}{2}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big] + \frac{1}{2}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big]\\&\le -\frac{1}{4}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big].\end{aligned} \tag{35}$$

Thus (34) becomes

$$\begin{aligned}\mathbb{E}\big[\mathcal{F}(W_G^{t+1})\big] &\le \mathbb{E}\big[\mathcal{F}(W_G^t)\big] - \frac{1}{4}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big]\\&+ \frac{\delta_{OT}^2}{2} + \frac{\eta_l E}{2}\beta^2 + \frac{8LE\eta_l}{\lambda\rho_{\max}}(\sigma^2+G^2) + 2L\delta_P^2 + L\delta_{OT}^2.\end{aligned} \tag{36}$$

The quantity β^2 can be bounded explicitly using the construction above; if we keep only the dominant terms in η_l and E and use $\eta_l \le 1/(8LE)$, we obtain

$$\frac{\eta_l E}{2}\beta^2 \le 4L^2E^2\eta_l G^2 + 4LE^2\eta_l^2(\lambda\rho_{\max})^2\delta_P^2 + \frac{L\eta_l E}{2}\sigma^2.$$

Collecting all error contributions in (36), we write

$$\begin{aligned}\mathbb{E}\big[\mathcal{F}(W_G^{t+1})\big] &\le \mathbb{E}\big[\mathcal{F}(W_G^t)\big] - \frac{1}{4}\mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big]\\&+ \Big(\frac{L\eta_l E\sigma^2}{2} + 4L^2E^2\eta_l G^2 + \frac{16LE\eta_l}{\lambda\rho_{\max}}(\sigma^2+G^2)\\&+ 4L\delta_P^2 + \frac{5}{2}\delta_{OT}^2 + 4LE^2\eta_l^2(\lambda\rho_{\max})^2\delta_P^2\Big).\end{aligned} \tag{37}$$

Define the constants in (37) as $\mathcal{E}_{\text{round}}$. Let $\Delta_t := \mathbb{E}\big[\mathcal{F}(W_G^t) - \mathcal{F}(W_G^*)\big]$. By μ-strong convexity of $\mathcal{F}$ (Assumption 2), we have

$$2\mu\,\Delta_t \le \mathbb{E}\big[\|\nabla\mathcal{F}(W_G^t)\|^2\big].$$

Subtracting $\mathcal{F}(W_G^*)$ from both sides of (37) and using this inequality,

$$\begin{aligned}\Delta_{t+1} &\le \Delta_t - \frac{1}{4}(2\mu\Delta_t) + \mathcal{E}_{\text{round}}\\&= \Big(1 - \frac{\mu}{2}\Big)\Delta_t + \mathcal{E}_{\text{round}}.\end{aligned} \tag{38}$$

Refining the constants to preserve the explicit dependence on $\eta_l E$ (keeping the original descent coefficient $-\eta_l E$ and matching it with μ as in the statement) yields

$$\Delta_{t+1} \le \Big(1 - \frac{\mu\eta_l E}{2}\Big)\Delta_t + \mathcal{E}_{\text{round}}, \tag{39}$$

with $\mathcal{E}_{\text{round}}$ given by the bracketed expression in (37).

Unrolling this recursion for T rounds, we obtain

$$\Delta_T \le \Big(1 - \frac{\mu\eta_l E}{2}\Big)^T \Delta_0 + \frac{\mathcal{E}_{\text{round}}}{\mu\eta_l E/2}.$$

Defining

$$\mathcal{E} := \frac{2}{\eta_l E}\,\mathcal{E}_{\text{round}}$$

and substituting the explicit form of $\mathcal{E}_{\text{round}}$ from (37), we arrive at Theorem 1. □

12. Broader Impact

Our proposed SubFLOT framework has significant implications for both the federated learning research community and the deployment of real-world AI systems. By enabling the training of personalized, privacy-preserving models on resource-constrained devices, our methodology contributes to the democratization of advanced machine learning in critical domains such as healthcare diagnostics, financial risk assessment, and industrial IoT. In these areas, data privacy and device heterogeneity are paramount concerns. Furthermore, the integration of optimal transport theory with adaptive submodel learning establishes a new technical pathway for addressing non-IID data in cross-device scenarios, potentially influencing algorithm design in related fields like distributed optimization and edge computing.