Table 8. Prostate dataset: number of data (3D image) in each client.

<table>
<thead>
<tr>
<th>Client</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>10</td>
<td>16</td>
<td>18</td>
<td>18</td>
<td>25</td>
<td>50</td>
<td>137</td>
</tr>
<tr>
<td>Val</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>12</td>
<td>25</td>
<td>68</td>
</tr>
<tr>
<td>Test</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>24</td>
<td>65</td>
</tr>
</tbody>
</table>

Figure 5. Representative original 2D image in retinal dataset (low data similarity). First row: client 1 to 3. Second row: client 4 to 6.

Figure 6. Representative original 2D image slices in prostate dataset (high data similarity). First row: client 1 to 3. Second row: client 4 to 6. E.g., the first slice comes from a 3D image in client 1.

B. FedSM-extra Algorithm

For the training of FedSM-extra, we train the global model and personalized models first, and then train the model selector, which incurs extra $\Delta R$ training rounds. In each training round of FedSM-extra, the communication cost is $2w_g$ in the previous $R$ rounds (the global and personalized models have the same model architecture). It becomes $w_s$ in the extra $\Delta R$ training rounds.

For the inference of FedSM-extra, both the global model and personalized models can be selected. Therefore $k \in$...
{0, 1, · · · , K} (For FedSM, k ∈ {1, 2, · · · , K}).

C. Proof of SoftPull Convergence

Let the current/local training rounds be \( r/R \), current/local local training steps be \( m/M \), the current/global training step be \( 1/T \). We denote the personalized model during training as \( w_{r,p,k}^{r,m} \). For simplicity we use \( f_k \) to denote loss \( L_{D_k} \).

After the local training in the last training round \( r - 1 \) finishes, we get model \( w_{r,p,k}^{r-1,M} \) and want to

\[
\min_{k=1}^{K} \frac{1}{\lambda} w_{r,p,k}^{r-1,M} - \frac{1}{\lambda} \left( \frac{1}{K-1} \sum_{k'=1, k' \neq k}^{K} w_{r,p,k'}^{r-1,M} \right)
\]

In the beginning of the current training round \( r \), from Eq. (11), we will have

\[
w_{r,p,k}^{r,0} = \lambda w_{r,p,k}^{r-1,M} + (1 - \lambda) \left( \frac{1}{K-1} \sum_{k'=1, k' \neq k}^{K} w_{r,p,k'}^{r-1,M} \right)
\]

In the current training round \( r \), we consider two stages. The first stage is a transition from then end of training round \( r - 1 \) to the start of the current training round \( r \), while the second stage is the start to the end of current training round \( r \). Let \( \lambda' = \frac{K-1}{K-1} \), \( 1 - \lambda' = \frac{K}{K-1} \), then

\[
w_{r,p,k}^{r,0} = \lambda' w_{r,p,k}^{r-1,M} + (1 - \lambda') \bar{w}_{r-1,M}^{r-1,M}
\]

where the bar denotes an average over all clients \( k \in \{1, 2, · · · , K\} \). It can be clearly seen that when the data distributions of clients are very similar, we set \( \lambda = \frac{1}{K}, \lambda' = 0 \), i.e., the “hard averaging” in FedAvg. When the data distributions are not similar at all, we set \( \lambda = 1, \lambda' = 1 \) to do local training. In other circumstances, theoretically we should set \( \lambda \in [\frac{1}{K}, 1], \lambda' \in [0, 1] \) according to the data similarity.

C.1. Difference

Suppose the stochastic gradient at iteration \((r, m)\) is \( \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \) and the expected gradient is \( \nabla f_k(w_{r,p,k}^{r,m}) = \mathbb{E}_{x_{r,p,k}^{r,m} \in D_k} \nabla f_k(w_{r,p,k}^{r,m}) = \mathbb{E} \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \). We need to bound

\[
\| (w_{r,p,k}^{r+1,0} - w_{r,p,k}^{r,M}) \|_2^2 = \| (1 - \lambda)(w_{r,p,k}^{r,M} - \frac{1}{K-1} \sum_{k'=1, k' \neq k}^{K} w_{r,p,k'}^{r,M}) \|_2^2
\]

\[
= (1 - \lambda)^2 \| w_{r,p,k}^{r,M} - \frac{1}{K-1} (K w_{r,p,k}^{r,M} - w_{r,p,k}^{r,M}) \|_2^2
\]

\[
= (1 - \lambda)^2 \frac{K^2}{(K-1)^2} \| w_{r,p,k}^{r,M} - \bar{w}_{r,p,k}^{r,M} \|_2^2
\]

where

\[
\mathbb{E}\| w_{r,p,k}^{r,M} - \bar{w}_{r,p,k}^{r,M} \|_2^2 = \eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r'-r} \sum_{m=0}^{M-1} (\mathbb{E} \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) - \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m})) \|_2^2
\]

\[
\leq \eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r'-r} \|_2^2 \mathbb{E}\| \sum_{m=0}^{M-1} (\mathbb{E} \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) - \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m})) \|_2^2
\]

\[
\leq M\eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r'-r} \|_2^2 \mathbb{E}\| \sum_{m=0}^{M-1} (\mathbb{E} \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) - \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m})) \|_2^2
\]

\[
\leq 2M^2(G^2 + \sigma^2) \eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r'-r} \|_2^2
\]

\[
\leq 2M^2(G^2 + \sigma^2) \eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r'-r} \|_2^2
\]

\[
\mathbb{E}\| (w_{r,p,k}^{r+1,0} - w_{r,p,k}^{r,M}) \|_2^2 \leq 2(G^2 + \sigma^2) \text{ based on Assumptions 3 and 2. Then}
\]

\[
\mathbb{E}\| (w_{r,p,k}^{r+1,0} - w_{r,p,k}^{r,M}) \|_2^2 \leq \frac{[1 - (\lambda')^{r+1,0} + (1 - \lambda)^2K^2}{(1 - \lambda')^2(K-1)^2} - 2M^2(G^2 + \sigma^2) \eta^2
\]

\[
= [1 - (\lambda')^{r+1,0} + 2M^2(G^2 + \sigma^2) \eta^2
\]

C.2. Local Objective

Here we consider the local objective function to optimize. From \((r, 0)\) to \((r, M)\), i.e. \( m \in \{0, 1, · · · , M - 1\} \), due to the Lipschitz smooth assumption we have

\[
f_k(w_{r,p,k}^{r,m+1}) - f_k(w_{r,p,k}^{r,m})
\]

\[
\leq \langle \nabla f_k(w_{r,p,k}^{r,m}), w_{r,p,k}^{r,m+1} - w_{r,p,k}^{r,m} \rangle + \frac{L}{2} \| w_{r,p,k}^{r,m+1} - w_{r,p,k}^{r,m} \|_2^2
\]

\[
= -\eta \langle \nabla f_k(w_{r,p,k}^{r,m}), \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \rangle
\]

\[
+ \frac{\eta^2 L}{2} \| \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \|_2^2
\]

\[
= -\eta \langle \nabla f_k(w_{r,p,k}^{r,m}), \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \rangle
\]

\[
+ \frac{\eta^2 L}{2} \| \nabla f_k(w_{r,p,k}^{r,m}, x_{r,p,k}^{r,m}) \|_2^2
\]

\[
(21)
\]
Take the expectation and suppose $\eta \leq \frac{1}{2}$.

\[
\begin{align*}
\mathbb{E}[f_k(w_{k,p}^{r,m+1}) - f_k(w_{k,p}^{r,m})] & \leq -\eta(1 - \frac{\eta L}{2})\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 + \frac{\eta^2 L \sigma^2}{2} \\
& \leq -\frac{\eta}{2}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 + \frac{\eta^2 L \sigma^2}{2} \\
\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 & \leq \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,m}) - f_k(w_{k,p}^{r,m+1})] + \eta L \sigma^2 \\
\sum_{m=0}^{M-1} \mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 & \leq \frac{2}{\eta} \mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r,M})] + M \eta L \sigma^2
\end{align*}
\]

While from $(r, M)$ to $(r + 1, 0)$, we have

\[
\begin{align*}
f_k(w_{k,p}^{r+1,0}) - f_k(w_{k,p}^{r,M}) & \leq \langle \nabla f_k(w_{k,p}^{r,M}), w_{k,p}^{r+1,0} - w_{k,p}^{r,M} \rangle + \frac{L}{2}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
& \leq \frac{\eta}{8}\|\nabla f_k(w_{k,p}^{r,M})\|_2^2 + \frac{2}{\eta} + \frac{L}{2}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
& \leq \frac{\eta}{4}\|\nabla f_k(w_{k,p}^{r,M-1})\|_2^2 + \frac{\eta L^2}{4}\|w_{k,p}^{r,M} - w_{k,p}^{r,M-1}\|_2^2 \\
& \quad + \frac{2}{\eta} + \frac{L}{2}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
& = \frac{\eta}{4}\|\nabla f_k(w_{k,p}^{r,M-1})\|_2^2 + \frac{\eta L^2}{4}\|\nabla f_k(w_{k,p}^{r,M-1}, x_{k,p}^{r,M-1})\|_2^2 \\
& \quad + \frac{2}{\eta} + \frac{L}{2}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2
\end{align*}
\]

Therefore, from $(r, 0)$ to $(r + 1, 0)$, we have

\[
\sum_{m=0}^{M-1} \mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
\leq \frac{2}{\eta} \mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r+1,0})] + M \eta L \sigma^2 \\
+ \frac{2}{\eta} \mathbb{E}[f_k(w_{k,p}^{r+1,0}) - f_k(w_{k,p}^{r,M})] \\
\leq \frac{2}{\eta} \mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r+1,0})] + M \eta L \sigma^2 \\
+ \frac{1}{2}\|\nabla f_k(w_{k,p}^{r,M-1})\|_2^2 + \frac{\eta L^2}{4}(G^2 + \sigma^2) \\
+ \frac{4}{\eta^2} + \frac{L}{\eta}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2
\]
Now we bound the gradient of the proposed objective.

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \nabla f_k^r(u_{k,p}^r) \|_2^2 \\
= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \frac{1}{\lambda} \nabla f_k(u_{k,p}^r) - \frac{1}{\lambda} \frac{1}{K-1} \nabla f_k(u_{k,p}^r) \|_2^2 \\
\leq \frac{2}{K} \sum_{k=1}^{K} \left( \frac{1}{\lambda^2} + \frac{(1-\lambda)^2}{\lambda^2(K-1)} \right) \mathbb{E} \| \nabla f_k(u_{k,p}^r) \|_2^2 \\
+ L^2 \mathbb{E} \| u_{k,p}^r - w_{k,p}^r \|_2^2 \\
= \left( \frac{1}{\lambda^2} + \frac{(1-\lambda)^2}{\lambda^2(K-1)} \right) \frac{2}{K} \sum_{k=1}^{K} \mathbb{E} \| \nabla f_k(u_{k,p}^r) \|_2^2 \\
+ L^2(1-\lambda)^2 K^2 \frac{2}{\lambda^2(K-1)^2} \mathbb{E} \| u_{k,p}^r - w_{k,p}^r \|_2^2 \\
\leq \frac{1}{\lambda^2} + \frac{(1-\lambda)^2}{\lambda^2(K-1)} \frac{4}{K} \sum_{k=1}^{K} (f_k^o - f_k^s) \\
+ 2\eta L^2 \frac{2\eta L^2(G^2 + \sigma^2)}{\eta RM} + \frac{1}{RM} \frac{8}{\eta^2} \frac{2L}{\eta} \frac{(1-\lambda)^2 K^2}{(K-1)^2} \\
\cdot \frac{K}{(K-1)^2} \sum_{k=1}^{K} \sum_{r=0}^{R-1} \sum_{m=0}^{M-1} \mathbb{E} \| u_{k,p}^r - w_{k,p}^r \|_2^2 \\
+ \frac{1}{RM} \frac{L^2(1-\lambda)^2 K^2}{\lambda^2(K-1)^2} \sum_{k=1}^{K} \sum_{r=0}^{R-1} \sum_{m=0}^{M-1} \mathbb{E} \| u_{k,p}^r - w_{k,p}^r \|_2^2
\]

(31)

which converges to

\[
O\left( \frac{1}{\eta RM} \right) + \frac{(1-\lambda)^2}{\lambda^2} \sum_{k=1}^{K} \sum_{r=0}^{R-1} \sum_{m=0}^{M-1} \mathbb{E} \| u_{k,p}^r - w_{k,p}^r \|_2^2 \\
= O\left( \frac{1}{\eta RM} \right) + \frac{M}{R^2} \sum_{r=0}^{R-1} (1-\lambda)^2 \\
+ M^2 \eta^2 \sum_{r=0}^{R-1} (1-\lambda)^2
\]

(33)

Suppose \( \eta = O\left( \frac{1}{\sqrt{RM}} \right) \) and \( M = O(R^4) \), the convergence rate is \( O\left( \frac{1}{\lambda^4 \sqrt{RM}} \right) \) with an error \( O\left( \frac{M \sum_{r=0}^{R-1} (1-\lambda)^2}{R^2 \lambda^2} \right) \).

D. Additional Experimental Results
Figure 7. Visual comparison of retinal disc (green) and cup (blue) segmentation. Dice denotes the retinal disc and cup Dice coefficient.

Figure 8. Visual comparison of prostate (green) segmentation. Dice denotes the Dice coefficient.
### Table 9. Test Dice coefficient comparison of retinal disc segmentation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Client 1</th>
<th>Client 2</th>
<th>Client 3</th>
<th>Client 4</th>
<th>Client 5</th>
<th>Client 6</th>
<th>Client Avg Dice</th>
<th>Global Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized</td>
<td>0.9628</td>
<td>0.9486</td>
<td>0.9489</td>
<td>0.9539</td>
<td>0.9242</td>
<td>0.9565</td>
<td>0.9492</td>
<td>0.9522</td>
</tr>
<tr>
<td>Client 1 Local</td>
<td>0.9454</td>
<td>0.4357</td>
<td>0.8956</td>
<td>0.6073</td>
<td>0.4464</td>
<td>0.8409</td>
<td>0.6952</td>
<td>0.7129</td>
</tr>
<tr>
<td>Client 2 Local</td>
<td>0.2936</td>
<td>0.9431</td>
<td>0.1099</td>
<td>0.8371</td>
<td>0.2439</td>
<td>0.4575</td>
<td>0.4809</td>
<td>0.5589</td>
</tr>
<tr>
<td>Client 3 Local</td>
<td>0.2940</td>
<td>0.3998</td>
<td>0.9468</td>
<td>0.6399</td>
<td>0.4349</td>
<td>0.8256</td>
<td>0.6982</td>
<td>0.7120</td>
</tr>
<tr>
<td>Client 4 Local</td>
<td>0.6830</td>
<td>0.9400</td>
<td>0.4805</td>
<td>0.9526</td>
<td>0.3088</td>
<td>0.7803</td>
<td>0.6909</td>
<td>0.7796</td>
</tr>
<tr>
<td>Client 5 Local</td>
<td>0.6102</td>
<td>0.2169</td>
<td>0.4601</td>
<td>0.2518</td>
<td>0.9033</td>
<td>0.7064</td>
<td>0.5248</td>
<td>0.5373</td>
</tr>
<tr>
<td>Client 6 Local</td>
<td>0.8806</td>
<td>0.7937</td>
<td>0.8354</td>
<td>0.8475</td>
<td>0.4413</td>
<td>0.9547</td>
<td>0.7922</td>
<td>0.8555</td>
</tr>
<tr>
<td>FedAvg</td>
<td>0.9554</td>
<td>0.9410</td>
<td>0.9372</td>
<td>0.9535</td>
<td>0.8653</td>
<td>0.9549</td>
<td>0.9346</td>
<td>0.9444</td>
</tr>
<tr>
<td>FedProx</td>
<td>0.9447</td>
<td>0.9343</td>
<td>0.9229</td>
<td>0.9469</td>
<td>0.7573</td>
<td>0.9480</td>
<td>0.9090</td>
<td>0.9283</td>
</tr>
<tr>
<td>Scaffold</td>
<td>0.9207</td>
<td>0.9297</td>
<td>0.9026</td>
<td>0.9474</td>
<td>0.6347</td>
<td>0.9528</td>
<td>0.8813</td>
<td>0.9170</td>
</tr>
</tbody>
</table>

### Table 10. Test Dice coefficient comparison of retinal cup segmentation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Unseen Client k = 6</th>
<th>Unseen Client k = 5</th>
<th>Unseen Client k = 4</th>
<th>Unseen Client k = 3</th>
<th>Unseen Client k = 2</th>
<th>Unseen Client k = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GM</td>
<td>PM1</td>
<td>PM2</td>
<td>PM3</td>
<td>PM4</td>
<td>PM5 PM6</td>
</tr>
<tr>
<td>Centralized</td>
<td>0.8118</td>
<td>0.8510</td>
<td>0.8302</td>
<td>0.8714</td>
<td>0.8503</td>
<td>0.8322</td>
</tr>
<tr>
<td>FedAvg</td>
<td>0.8484</td>
<td>0.8654</td>
<td>0.8214</td>
<td>0.8866</td>
<td>0.4064</td>
<td>0.8811</td>
</tr>
<tr>
<td>FedProx</td>
<td>0.8589</td>
<td>0.8313</td>
<td>0.8224</td>
<td>0.8551</td>
<td>0.4064</td>
<td>0.8887</td>
</tr>
<tr>
<td>Scaffold</td>
<td>0.8380</td>
<td>0.7856</td>
<td>0.8267</td>
<td>0.8746</td>
<td>0.4171</td>
<td>0.8784</td>
</tr>
<tr>
<td>FedSM</td>
<td>0.8818</td>
<td>0.8619</td>
<td>0.8498</td>
<td>0.8901</td>
<td>0.4118</td>
<td>0.8646</td>
</tr>
<tr>
<td>FedSM-extra</td>
<td>0.8747</td>
<td>0.8685</td>
<td>0.8467</td>
<td>0.8794</td>
<td>0.4265</td>
<td>0.8809</td>
</tr>
</tbody>
</table>

Table 11. (retinal segmentation, Dice = average of disc and cup Dice coefficients) Model selection frequency from the model selector when FL train with clients \( \{1, 2, \cdots, 6\}/\{k\} \) and test on the unseen client \( k \in \{1, 2, \cdots, 6\} \). From left to right, GM denotes the global model and PM denotes the personalized model \( \{1, 2, \cdots, 6\}/\{k\} \). Here we choose the best \( \gamma \).