FedMOP: Achieving Enhanced Privacy and Performance in Federated Learning via Momentum Orthogonal Projection

Supplementary Material

1. Convergence Analysis of FedMOP

In this appendix, we provide a comprehensive convergence analysis for FedMOP under standard federated learning assumptions. We first establish preliminary lemmas, then prove convergence for both full and partial participation scenarios.

1.1. Preliminary Lemmas

We begin by stating fundamental lemmas required for the convergence analysis:

Lemma 1. *(Triangle inequality). Let $\{v_1, \ldots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then:*

$$\|v_i + v_j\|^2 \leq 2\|v_i\|^2 + 2\|v_j\|^2. \tag{1}$$

Lemma 2. *For random variables $x_1, \ldots, x_n$, we have*

$$\mathbb{E}[\|x_1 + \cdots + x_n\|^2] \leq n\mathbb{E}[\|x_1\|^2 + \cdots + \|x_n\|^2]. \tag{2}$$

Lemma 3. *For independent, mean 0 random variables $x_1, \ldots, x_n$, we have*

$$\mathbb{E}[\|x_1 + \cdots + x_n\|^2] = \mathbb{E}[\|x_1\|^2 + \cdots + \|x_n\|^2]. \tag{3}$$

Lemma 4. *Under any step-size satisfying $\eta \leq \frac{1}{8LK}$, for the initial model x_t of the t-th round and $x_{t-1,i,k}$ updated by the $(t-1)$-th round local gradient, we have:*

$$\begin{aligned}
&\sum_{k=0}^{K-1} \|x_t - x_{t-1,i,k}\|^2 \\
&\leq \frac{\frac{3}{4}K(K+1)\eta^2}{1-\eta^2L^2}\Big(\sum_{j=1}^{d}\sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d}\sigma_{g,j}^2 + \|\nabla f(x_t)\|^2\Big).
\end{aligned} \tag{4}$$

Proof. We analyze the distance between x_t and intermediate iterates:

$$\begin{aligned}
&\|x_t - x_{t-1,i,k}\|^2 \\
&= \|x_{t-1,i,k+1} + \eta g_{t-1,i,k} - x_t\|^2 \\
&= \mathbb{E}\|x_{t-1,i,k+1} - x_t \\
&\quad + \eta(g_{t-1,i,k} - \nabla F_i(x_{t-1,i,k}) \\
&\quad + \nabla F_i(x_{t-1,i,k}) - \nabla F_i(x_t) \\
&\quad + \nabla F_i(x_t) - \nabla f(x_t) + \nabla f(x_t))\|^2 \\
&\overset{(a)}{\leq} \|x_{t-1,i,k+1} - x_t\|^2 \\
&\quad + \|\eta(g_{t-1,i,k} - \nabla F_i(x_{t-1,i,k}))\|^2 \\
&\quad + \|\eta(\nabla F_i(x_{t-1,i,k}) - \nabla F_i(x_t))\|^2 \\
&\quad + \|\eta(\nabla F_i(x_t) - \nabla f(x_t))\|^2 \\
&\quad + \|\eta\nabla f(x_t)\|^2 \\
&\overset{(b)}{\leq} \|x_{t-1,i,k+1} - x_t\|^2 \\
&\quad + \|\eta(g_{t-1,i,k} - \nabla F_i(x_{t-1,i,k}))\|^2 \\
&\quad + \eta^2L^2\|x_{t-1,i,k} - x_t\|^2 \\
&\quad + \|\eta(\nabla F_i(x_t) - \nabla f(x_t))\|^2 \\
&\quad + \|\eta\nabla f(x_t)\|^2 \\
&= \|x_{t-1,i,k+1} - x_t\|^2 \\
&\quad + \eta^2L^2\|x_{t-1,i,k} - x_t\|^2 + \eta^2P_1,
\end{aligned} \tag{5}$$

where (a) follows from Lemma 3, (b) from Assumption **??**, and

$$\begin{aligned}
P_1 = &\|g_{t-1,i,k} - \nabla F_i(x_{t-1,i,k})\|^2 \\
&+ \|\nabla F_i(x_t) - \nabla f(x_t)\|^2 \\
&+ \|\nabla f(x_t)\|^2.
\end{aligned} \tag{6}$$

Combining like terms on both sides:

$$\begin{aligned}
&\|x_t - x_{t-1,i,k}\|^2 \\
&\leq \frac{1}{1-\eta^2L^2}\|x_{t-1,i,k+1} - x_t\|^2 \\
&\quad + \frac{\eta^2}{1-\eta^2L^2}P_1.
\end{aligned} \tag{7}$$

Unrolling the recursion:

$$\begin{aligned}
&\sum_{k=0}^{K-1} \|x_t - x_{t-1,i,k}\|^2 \\
&\leq \sum_{k=0}^{K-1} \sum_{r=0}^{k} \Big(\frac{1}{1-\eta^2 L^2}\Big)^r \\
&\quad \times \Big(\frac{\eta^2}{1-\eta^2 L^2}\Big) P_1 \\
&\overset{(a)}{\leq} \sum_{k=0}^{K-1} \sum_{r=0}^{k} \Big(\frac{3}{2}\Big)^r \\
&\quad \times \Big(\frac{\eta^2}{1-\eta^2 L^2}\Big) P_1 \\
&\overset{(b)}{=} \frac{\frac{3}{4}K(K+1)\eta^2}{1-\eta^2 L^2} P_1,
\end{aligned} \tag{8}$$

where (a) uses $\eta \leq \frac{1}{8LK}$ and (b) applies the geometric series formula.

By Assumption **??**:

$$P_1 = \sum_{j=1}^{d} \sigma_{l,j}^2 + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2. \tag{9}$$

□

Lemma 5. *For the t-th round initial model x_t and the previous round local gradient sum $\Delta_{t-1,i}$:*

$$\begin{aligned}
&\langle \nabla f(x_t), \Delta_{t-1,i} \rangle \\
&\leq \frac{3\eta}{8} \frac{K(K+1)\eta^2 L^2}{1-\eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big).
\end{aligned} \tag{10}$$

Proof.

$$\begin{aligned}
&\langle \nabla f(x_t), \Delta_{t-1,i} \rangle \\
&= \langle \nabla f(x_t), \Delta_{t-1,i} + \eta K \nabla f(x_t) \rangle \\
&\quad - \langle \nabla f(x_t), \eta K \nabla f(x_t) \rangle \\
&= \langle \nabla f(x_t), -\eta \sum_{k=0}^{K-1} \nabla f(x_{t-1,i,k}) \\
&\quad + \eta K \nabla f(x_t) \rangle \\
&\quad - \langle \nabla f(x_t), \eta K \nabla f(x_t) \rangle \\
&\overset{(a)}{\leq} \frac{\eta}{2} \|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta L^2}{2} \sum_{k=0}^{K-1} \|x_t - x_{t-1,i,k}\|^2 \\
&\quad - \eta K \|\nabla f(x_t)\|^2 \\
&\overset{(b)}{\leq} \frac{\eta}{2} \|\nabla f(x_t)\|^2 \\
&\quad + \frac{3\eta}{8} \frac{K(K+1)\eta^2 L^2}{1-\eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big) \\
&\quad - \eta K \|\nabla f(x_t)\|^2,
\end{aligned} \tag{11}$$

where (a) applies $ab \leq (a^2+b^2)/2$, and (b) uses Lemma 4. □

Lemma 6. *Under any step-size satisfying $\eta \leq \frac{1}{8LK}$, for the shifted initial model $\tilde{x}_{t,i}$ and $\tilde{x}_{t,i,k}$ updated by the t-th round local gradient:*

$$\begin{aligned}
&\sum_{k=0}^{K-1} \|\tilde{x}_{t,i} - \tilde{x}_{t,i,k}\|^2 \\
&\leq (K-1)\eta^2 \mathbb{E}\Big[\sum_{j=1}^{d} (\sigma_{l,j}^2 + 2K\sigma_{g,j}^2) \Big] \\
&\quad + (K-1)\eta^2 \mathbb{E}[\|\nabla f(\tilde{x}_{t,i})\|^2].
\end{aligned} \tag{12}$$

Proof. The result trivially holds for $k = 1$ since $\tilde{x}_{t,i,0} = \tilde{x}_{t,i}$ for all $i \in [M]$. For $k \geq 1$, we observe that for any client

$i \in [M]$ and $k \in [K]$:

$$
\begin{aligned}
&\|\tilde{x}_{t,i,k} - \tilde{x}_{t,i}\|^2 \\
&= \mathbb{E}\|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i} - \eta g^t_{i,k-1}\|^2 \\
&\le \mathbb{E}\|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i} \\
&\quad - \eta(g^t_{i,k-1} - \nabla F_i(\tilde{x}^t_{i,k-1}) \\
&\quad + \nabla F_i(\tilde{x}^t_{i,k-1}) - \nabla F_i(\tilde{x}_{t,i}) \\
&\quad + \nabla F_i(\tilde{x}_{t,i}) - \nabla f(\tilde{x}_{t,i}) \\
&\quad + \nabla f(\tilde{x}_{t,i}))\|^2 \\
&\le \mathbb{E}\|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i}\|^2 \\
&\quad + \mathbb{E}\|\eta(g^t_{i,k-1} - \nabla F_i(\tilde{x}^t_{i,k-1}))\|^2 \\
&\quad + \mathbb{E}[\|\eta(\nabla F_i(\tilde{x}^t_{i,k-1}) - \nabla F_i(\tilde{x}_{t,i}))\|^2] \\
&\quad + \mathbb{E}[\|\eta(\nabla F_i(\tilde{x}_{t,i}) - \nabla f(\tilde{x}_{t,i}))\|^2] \\
&\quad + \mathbb{E}[\|\eta\nabla f(\tilde{x}_{t,i})\|^2]. \qquad (13)
\end{aligned}
$$

Further bounding:

$$
\begin{aligned}
&\|\tilde{x}_{t,i,k} - \tilde{x}_{t,i}\|^2 \\
&\le \|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i}\|^2 + \eta^2 \sum_{j=1}^{d} \sigma^2_{l,j} \\
&\quad + \eta^2 L^2 \|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i}\|^2 \\
&\quad + \eta^2 \sum_{j=1}^{d} \sigma^2_{g,j} + \eta^2 \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\le (1 + \eta^2 L^2)\|\tilde{x}^t_{i,k-1} - \tilde{x}_{t,i}\|^2 \\
&\quad + \eta^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} + \sum_{j=1}^{d} \sigma^2_{g,j} \\
&\quad + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big). \qquad (14)
\end{aligned}
$$

Unrolling the recursion for $k \ge 1$:

$$
\begin{aligned}
&\|\tilde{x}_{t,i,k} - \tilde{x}_{t,i}\|^2 \\
&\le \sum_{j=0}^{k-1} (1 + \eta^2 L^2)^j \eta^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} \\
&\quad + \sum_{j=1}^{d} \sigma^2_{g,j} + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big) \\
&\le k\eta^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} + \sum_{j=1}^{d} \sigma^2_{g,j} \\
&\quad + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big), \qquad (15)
\end{aligned}
$$

where the last step uses $(1 + \eta^2 L^2)^j \le 1$ when $\eta \le \frac{1}{8LK}$.

Summing over k:

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \|\tilde{x}_{t,i,k} - \tilde{x}_{t,i}\|^2 \\
&\le \sum_{k=1}^{K-1} k\eta^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} + \sum_{j=1}^{d} \sigma^2_{g,j} \\
&\quad + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big) \\
&= \frac{(K-1)K}{2} \eta^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} \\
&\quad + \sum_{j=1}^{d} \sigma^2_{g,j} + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big) \\
&\le (K-1)\eta^2 \Big(\sum_{j=1}^{d} (\sigma^2_{l,j} + 2K\sigma^2_{g,j}) \\
&\quad + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big). \qquad (16)
\end{aligned}
$$

□

Lemma 7. *For the t-th round corrected initial model $\tilde{x}_{t,i}$ and the t-th round local gradient sum $\Delta_{t,i}$:*

$$
\begin{aligned}
&\langle \nabla f(\tilde{x}_{t,i}), \Delta_{t,i} \rangle \\
&\le \frac{\eta}{2} \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\quad + \frac{1}{2}(K-1)\eta^3 L^2 \Big(\sum_{j=1}^{d} \sigma^2_{l,j} \\
&\quad + \sum_{j=1}^{d} \sigma^2_{g,j} \Big). \qquad (17)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\langle \nabla f(\tilde{x}_{t,i}), \Delta_{t,i} \rangle \\
&= \langle \nabla f(\tilde{x}_{t,i}), \Delta_{t,i} + \eta K \nabla f(\tilde{x}_{t,i}) \rangle \\
&\quad - \langle \nabla f(\tilde{x}_{t,i}), \eta K \nabla f(\tilde{x}_{t,i}) \rangle \\
&= \langle \nabla f(\tilde{x}_{t,i}), -\eta \sum_{k=0}^{K-1} \nabla f(x_{t,i,k}) \\
&\quad + \eta K \nabla f(\tilde{x}_{t,i}) \rangle \\
&\quad - \langle \nabla f(\tilde{x}_{t,i}), \eta K \nabla f(\tilde{x}_{t,i}) \rangle \\
&\overset{(a)}{\leq} \frac{\eta}{2} \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\quad + \frac{\eta L^2}{2} \sum_{k=0}^{K-1} \|\tilde{x}_{t,i} - x_{t,i,k}\|^2 \\
&\quad - \eta K \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\overset{(b)}{\leq} \frac{\eta}{2} \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\quad + \frac{1}{2}(K-1)\eta^3 L^2 \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(\tilde{x}_{t,i})\|^2 \Big) \\
&\quad - \eta K \|\nabla f(\tilde{x}_{t,i})\|^2 \\
&\leq \frac{1}{2}(K-1)\eta^3 L^2 \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 \Big), \qquad (18)
\end{aligned}
$$

where (a) uses $ab \leq (a^2 + b^2)/2$, and (b) is from Lemma 6. □

1.2. Momentum Evolution Lemma

We now establish a crucial lemma regarding the momentum-evolved offset vectors:

Lemma 8. *Under the momentum update rule in Eq.* **??** *with* $\bar{\gamma} \in [0.8, 0.95]$ *and* $\sigma_\gamma \leq 0.05$*, the expected squared deviation between the actual offset* $\Omega_{t,i}$ *and the natural offset* $\Omega_{t,i}^{nat}$ *satisfies:*

$$
\mathbb{E}[\|\Omega_{t,i} - \Omega_{t,i}^{nat}\|^2] \leq C_1 \sigma_0^2 \bar{\gamma}^{2t} + C_2 \sigma_\gamma^2 t, \qquad (19)
$$

where C_1 *and* C_2 *are constants depending on* L*,* G*, and* K*.*

Proof. By the momentum evolution equation:

$$
\Omega_{t,i} = \gamma_{t,i} \Omega_{t-1,i} + (1 - \gamma_{t,i}) \Omega_{t,i}^{\text{nat}}, \qquad (20)
$$

where $\gamma_{t,i} = \bar{\gamma} + \epsilon_{t,i}$ with $\epsilon_{t,i} \sim \mathcal{N}(0, \sigma_\gamma^2)$.

We can express:

$$
\begin{aligned}
&\Omega_{t,i} - \Omega_{t,i}^{\text{nat}} \\
&= \gamma_{t,i}(\Omega_{t-1,i} - \Omega_{t,i}^{\text{nat}}) \\
&\quad + \epsilon_{t,i}(\Omega_{t-1,i} - \Omega_{t,i}^{\text{nat}}). \qquad (21)
\end{aligned}
$$

Taking expectation of the squared norm:

$$
\begin{aligned}
&\mathbb{E}[\|\Omega_{t,i} - \Omega_{t,i}^{\text{nat}}\|^2] \\
&\leq \mathbb{E}[\|\gamma_{t,i}(\Omega_{t-1,i} - \Omega_{t,i}^{\text{nat}})\|^2] \\
&\quad + \mathbb{E}[\|\epsilon_{t,i}(\Omega_{t-1,i} - \Omega_{t,i}^{\text{nat}})\|^2] \\
&\leq (\bar{\gamma}^2 + \sigma_\gamma^2) \mathbb{E}[\|\Omega_{t-1,i} - \Omega_{t,i}^{\text{nat}}\|^2] \\
&\quad + \sigma_\gamma^2 \mathbb{E}[\|\Omega_{t-1,i}\|^2 + \|\Omega_{t,i}^{\text{nat}}\|^2]. \qquad (22)
\end{aligned}
$$

Since $\Omega_{t,i}^{\text{nat}}$ is bounded by $\mathcal{O}(KG)$ (from gradient bounds), and unrolling the recursion from the initial condition $\Omega_{0,i} \sim \mathcal{N}(0, \sigma_0^2 I)$:

$$
\mathbb{E}[\|\Omega_{t,i} - \Omega_{t,i}^{\text{nat}}\|^2] \leq C_1 \sigma_0^2 \bar{\gamma}^{2t} + C_2 \sigma_\gamma^2 t, \qquad (23)
$$

where $C_1 = d$ and $C_2 = \mathcal{O}(K^2 G^2)$. □

1.3. Full Participation Convergence

We now prove the main convergence theorem for the full participation case.

Theorem 1 (Full Participation)**.** *Under Assumptions* **??**-**??***, with* $\eta \leq \frac{1}{16KL}$*,* $K \geq 2$*,* $M \geq 2$*,* $\bar{\gamma} \in [0.8, 0.95]$*, and* $\sigma_\gamma \leq 0.05$*, the iterates of FedMOP satisfy:*

$$
\begin{aligned}
&\min_{0 \leq t \leq T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \\
&\leq \mathcal{O}\Big(\frac{f(x_0) - f(x^*) + \Phi_1 \sigma^2 + \Phi_2 d G^2}{\eta \frac{T(T+1)}{2} \Gamma(K)} \Big) \\
&\quad + \mathcal{O}(\sigma_0^2 \bar{\gamma}^{2T}) + \mathcal{O}(\sigma_\gamma^2 T), \qquad (24)
\end{aligned}
$$

where:

$$
\Gamma(K) = K + \frac{3K(K+1)\eta^2 L^2}{4(1 - \eta^2 L^2)}, \qquad (25)
$$

$$
\Phi_1 = \frac{3K(K+1)\eta^3 L^2 T(T+1)}{4(1 - \eta^2 L^2) \cdot 2}, \qquad (26)
$$

$$
\Phi_2 = \Big(\frac{L}{2} + K\eta \Big) \frac{T(T+1)}{2}. \qquad (27)
$$

Proof. Since the global function f is L-smooth:

$$
\begin{aligned}
f(x_{t+1}) &\leq f(\tilde{x}_t) + \langle \nabla f(\tilde{x}_t), \Delta_t \rangle \\
&\quad + \frac{L}{2} \|\Delta_t\|^2, \qquad (28) \\
f(\tilde{x}_t) &\leq f(x_t) + \langle \nabla f(x_t), \Omega_t \rangle \\
&\quad + \frac{L}{2} \|\Omega_t\|^2. \qquad (29)
\end{aligned}
$$

Although not all $\tilde{x}_{t,i}$ are actually aggregated, for proof convenience, we assume $\tilde{x}_t = \frac{1}{M}\sum_{i=1}^{M}\tilde{x}_{t,i}$.

Integrating the above two equations:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) \\ &+ \underbrace{\langle \nabla f(\tilde{x}_t), \Delta_t \rangle}_{P_2} + \frac{L}{2}\|\Delta_t\|^2 \\ &+ \underbrace{\langle \nabla f(x_t), \Omega_t \rangle}_{P_3} + \frac{L}{2}\|\Omega_t\|^2. \end{aligned} \tag{30}$$

Bounding P_2:

$$\begin{aligned} P_2 &= \langle \nabla f(\tilde{x}_t), \Delta_t \rangle \\ &= \frac{1}{M}\sum_{i=1}^{M}\langle \nabla f(\tilde{x}_t), \Delta_{t,i} \rangle \\ &= \frac{1}{M}\sum_{i=1}^{M}\langle \nabla f(\tilde{x}_t) - \nabla F_i(\tilde{x}_t) \\ &\quad + \nabla F_i(\tilde{x}_t) - \nabla F_i(\tilde{x}_{t,i}), \Delta_{t,i} \rangle \\ &\quad + \frac{1}{M}\sum_{i=1}^{M}\langle \nabla F_i(\tilde{x}_{t,i}), \Delta_{t,i} \rangle \\ &\overset{(a)}{\leq} \frac{1}{2M}\sum_{i=1}^{M}\Big(3\sum_{j=1}^{d}\sigma_{g,j}^2 \\ &\quad + 3dG^2 + 3dG^2 + KdG^2\Big) \\ &\quad + \frac{1}{M}\sum_{i=1}^{M}\langle \nabla F_i(\tilde{x}_{t,i}), \Delta_{t,i} \rangle \\ &\overset{(b)}{\leq} \Big(1.5\sum_{j=1}^{d}\sigma_{g,j}^2 + 1.5dG^2 \\ &\quad + 1.5dG^2 + \frac{K}{2}dG^2\Big) \\ &\quad + \frac{\eta^3 K^2 L^2}{2}\sum_{j=1}^{d}(K\sigma_{l,j}^2 + K\sigma_{g,j}^2), \end{aligned} \tag{31}$$

where (a) uses $ab \leq (a^2+b^2)/2$, Lemma 2, and Assumptions **??** and **??**; (b) is from Lemma 7.

Bounding P_3:

$$\begin{aligned} &\langle \nabla f(x_t), \Omega_t \rangle \\ &= \frac{1}{M}\sum_{i=1}^{M}\Big\langle \nabla f(x_t), \Delta_{t-1} + \Omega_{t-1} \\ &\quad - \frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\ &\quad \times (-\Omega_{t-1,i} - \Delta_{t-1,i})\Big\rangle. \end{aligned} \tag{32}$$

Because $(\Omega_{t-1,i} + \Delta_{t-1,i})$ is orthogonal to Ω_t:

$$\begin{aligned} &\langle \nabla f(x_t), \Omega_t \rangle \\ &= \langle \nabla f(x_t), \Delta_{t-1} \rangle + \langle \nabla f(x_t), \Omega_{t-1} \rangle \\ &\quad + \Big\langle \frac{\Omega_{t-1} + \Delta_{t-1}}{\eta}, \Delta_{t-1} + \Omega_{t-1} \Big\rangle \\ &\quad + \frac{1}{M}\sum_{i=1}^{M}\Big\langle \nabla f(x_t) + \frac{\Omega_{t-1,i} + \Delta_{t-1,i}}{\eta}, \\ &\quad \frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\ &\quad \times (-\Omega_{t-1,i} - \Delta_{t-1,i})\Big\rangle. \end{aligned} \tag{33}$$

Let the term inside the summation be denoted as P_4. After lengthy algebraic manipulations:

$$\begin{aligned} P_4 &\leq \frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2\eta\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\ &\quad \times \|\eta\nabla f(x_t)\|^2 \\ &\quad - \frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2\eta}. \end{aligned} \tag{34}$$

Substituting back and simplifying:

$$
\begin{aligned}
&\langle \nabla f(x_t), \Omega_t \rangle \\
&\le \langle \nabla f(x_t), \Delta_{t-1} \rangle + \langle \nabla f(x_t), \Omega_{t-1} \rangle \\
&\quad + \frac{1}{M} \sum_{i=1}^{M} \Big(\frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2 \|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\
&\quad \times \|\nabla f(x_t)\|^2 - \frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2\eta} \Big) \\
&\quad + \frac{\|\Omega_t\|^2}{2} \\
&\overset{(a)}{\le} \langle \nabla f(x_t), \Delta_{t-1} \rangle + \langle \nabla f(x_t), \Omega_{t-1} \rangle \\
&\quad + \frac{1}{M} \sum_{i=1}^{M} \Big(\frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2\eta \|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\
&\quad \times \|\eta \nabla f(x_t)\|^2 \Big) \\
&\overset{(b)}{\le} \frac{\eta}{2} \|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta}{2} \frac{\frac{3}{4} K(K+1)\eta^2 L^2}{1 - \eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big) \\
&\quad - K\eta \|\nabla f(x_t)\|^2 \\
&\quad + \frac{1}{M} \sum_{i=1}^{M} \Big(\frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{2 \|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2} \\
&\quad \times \eta \|\nabla f(x_t)\|^2 \Big) + \langle \nabla f(x_t), \Omega_{t-1} \rangle \\
&\overset{(c)}{\le} \eta \|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta}{2} \frac{\frac{3}{4} K(K+1)\eta^2 L^2}{1 - \eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big) \\
&\quad - K\eta \|\nabla f(x_t)\|^2 + \langle \nabla f(x_t), \Omega_{t-1} \rangle, \qquad (35)
\end{aligned}
$$

where (a) uses that $\|\Omega_{t-1} + \Delta_{t-1}\|$ (hypotenuse) is greater than $\|\Omega_t\|$; (b) is from Lemma 5; (c) uses $\mathbb{E}_i[\frac{\langle \Delta_{t-1,i} + \Omega_{t-1,i}, \Delta_{t-1} + \Omega_{t-1} \rangle}{\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2}] = 1$.

Unrolling the recursion:

$$
\begin{aligned}
&\langle \nabla f(x_t), \Omega_t \rangle \\
&\le t \Big(\eta \|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta}{2} \frac{\frac{3}{4} K(K+1)\eta^2 L^2}{1 - \eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big) \\
&\quad - K\eta \|\nabla f(x_t)\|^2 \Big) + \langle \nabla f(x_t), \Omega_0 \rangle, \qquad (36)
\end{aligned}
$$

where $\langle \nabla f(x_t), \Omega_0 \rangle = 0$ since $\Omega_0 = 0$.

Based on the bounds of P_2, P_3, and $\frac{L}{2}\|\Omega_t\|^2 \le \frac{LtdG^2}{2}$:

$$
\begin{aligned}
&f(x_{t+1}) \le f(x_t) \\
&\quad + \frac{\eta^3 K L^2}{2} \sum_{j=1}^{d} (\sigma_{l,j}^2 + \sigma_{g,j}^2) \\
&\quad + \frac{3}{2} \Big(\sum_{j=1}^{d} \sigma_{g,j}^2 + dG^2 + dG^2 + \frac{K}{3} dG^2 \Big) \\
&\quad + t \Big(\eta \|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta}{2} \frac{\frac{3}{4} K(K+1)\eta^2 L^2}{1 - \eta^2 L^2} \Big(\sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d} \sigma_{g,j}^2 + \|\nabla f(x_t)\|^2 \Big) \\
&\quad - K\eta \|\nabla f(x_t)\|^2 \Big) + \frac{(t+1)Ld}{2} G^2. \qquad (37)
\end{aligned}
$$

Rearranging and summing from $t = 0$ to $T - 1$:

$$
\begin{aligned}
&f(x_T) \le f(x_0) \\
&\quad + \eta \frac{T(T+1)}{2} \Big(-K + 1 \\
&\quad + \frac{3}{4} \frac{K(K+1)\eta^2 L^2}{1 - \eta^2 L^2} \Big) \|\nabla f(x_t)\|^2 \\
&\quad + \Big(\frac{T(T+1)}{2} \Big(\frac{3}{4} \frac{K(K+1)\eta^3 L^2}{1 - \eta^2 L^2} + 1 \Big) \\
&\quad + \frac{\eta^3 K L^2}{2} \Big) \sum_{j=1}^{d} \sigma_{l,j}^2 \\
&\quad + \Big(\frac{T(T+1)}{2} \Big(\frac{3}{4} \frac{K(K+1)\eta^3 L^2}{1 - \eta^2 L^2} + 1 \Big) \\
&\quad + \frac{\eta^3 K L^2}{2} + \frac{3}{2} \Big) \sum_{j=1}^{d} \sigma_{g,j}^2 \\
&\quad + \Big(\frac{(T+1)(T+2)L}{2} + \frac{3}{2} + \frac{K}{3} \Big) dG^2. \qquad (38)
\end{aligned}
$$

□

1.4. Partial Participation

Theorem 2 (Partial Participation)**.** *Under Assumptions* **??**-**??**, *with* $\eta \leq \frac{1}{16KL}$, *uniform client sampling with* $|S_t| = s < M$, $\bar{\gamma} \in [0.8, 0.95]$, *and* $\sigma_\gamma \leq 0.05$*:*

$$\begin{aligned}
&\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \\
&\leq \mathcal{O}\Big(\frac{f(x_0) - f(x^*) + \Phi_1\sigma^2 + \Phi_2 dG^2}{\eta\frac{T(T+1)}{2}\Gamma(K)}\Big) \\
&\quad + \mathcal{O}\Big(\frac{dKG^2}{s}\Big) + \mathcal{O}(\sigma_0^2\bar{\gamma}^{2T}) + \mathcal{O}(\sigma_\gamma^2 T). \qquad (39)
\end{aligned}$$

Proof. Since the i-th client participating in the t-th round may not have participated in the $(t-1)$-th round, we assume:

$$\Delta_{t-1,i} + \Omega_{t-1,i} = x_{r_i,i} - x_{r_i}, \qquad (40)$$

where r_i is the last round client i participated.

Under this assumption, the convergence process is similar to full participation, with the only difference being $\langle \nabla f(x_t), \Omega_t \rangle$:

$$\begin{aligned}
&\langle \nabla f(x_t), \Omega_t \rangle \\
&= \frac{1}{s}\sum_{i\in S_t} \Big\langle \nabla f(x_t), \frac{x_t - x_{r_i}}{t - r_i} \\
&\quad - \frac{\langle \frac{x_t - x_{r_i}}{t-r_i}, x_{r_i,i} - x_{r_i}\rangle}{\|x_{r_i,i} - x_{r_i}\|^2}(x_{r_i,i} - x_{r_i})\Big\rangle. \qquad (41)
\end{aligned}$$

After decomposition and applying the full participation proof techniques:

$$\begin{aligned}
&\langle \nabla f(x_t), \Omega_t \rangle \\
&\leq \text{(full participation terms)} \\
&\quad + \frac{1}{s}\sum_{i\in S_t} P_5, \qquad (42)
\end{aligned}$$

where P_5 is the additional error term.

Analyzing P_5:

$$\begin{aligned}
P_5 &\leq \Big(\frac{dG^2}{4\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2(t-r_i)^2} + \frac{1}{2}\Big) \\
&\quad \times \eta\|\nabla f(x_t)\|^2 \\
&\quad + \frac{\|x_t - x_{t-1}\|}{4\eta} \\
&\quad + \frac{\|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2}{4\eta(t-r_i)^2} \\
&\quad + \|\Delta_{t-1,i} + \Omega_{t-1,i}\|^2 \\
&\leq \frac{dG^2}{4(t-r_i)^2} + \frac{\eta\|\nabla f(x_t)\|^2}{2} \\
&\quad + \frac{dKG^2}{4\eta} + \frac{(t-1)dKG^2}{4\eta(t-r_i)^2} \\
&\quad + (t-1)dKG^2. \qquad (43)
\end{aligned}$$

Through a proof similar to the full participation case:

$$\begin{aligned}
&f(x_{t+1}) \leq f(x_t) \\
&\quad + \frac{\eta^3 KL^2}{2}\sum_{j=1}^{d}(\sigma_{l,j}^2 + \sigma_{g,j}^2) \\
&\quad + \frac{3}{2}\Big(\sum_{j=1}^{d}\sigma_{g,j}^2 + dG^2 + dG^2 + \frac{K}{3}dG^2\Big) \\
&\quad + t\Big(\frac{3}{2}\eta\|\nabla f(x_t)\|^2 \\
&\quad + \frac{\eta}{2}\frac{\frac{3}{4}K(K+1)\eta^2L^2}{1-\eta^2L^2}\Big(\sum_{j=1}^{d}\sigma_{l,j}^2 \\
&\quad + \sum_{j=1}^{d}\sigma_{g,j}^2 + \|\nabla f(x_t)\|^2\Big) \\
&\quad - K\eta\|\nabla f(x_t)\|^2\Big) \\
&\quad + (t+1)\Big(\frac{1}{s}\sum_{i\in S_t}\frac{1}{4(t-r_i)^2} + \frac{L}{2} + 1\Big)dKG^2. \qquad (44)
\end{aligned}$$

□

1.4.1. Performance Comparison

In the original PRI, the global model x_t satisfies:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \mathcal{O}\Big(\frac{2(f(x_0) - f(x^*))}{\lambda\eta KT} + \zeta_1\Big), \qquad (45)$$

where λ is a constant and ζ_1 represents noise terms.

In FedMOP, we demonstrate:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \\
&\leq \mathcal{O}\Big(\frac{2(f(x_0) - f(x^*))}{\eta\frac{T(T+1)}{2}\big(K + \frac{3}{4}\frac{K(K+1)\eta^2L^2}{1-\eta^2L^2}\big)} + \zeta_2\Big), \qquad (46)
\end{aligned}$$

where ζ_2 represents noise variables unrelated to convergence.

Our convergence rate:

$$\mathcal{O}\Big(\frac{1}{K^2T^2}\Big), \qquad (47)$$

is quadratically better than Fedvag's $\mathcal{O}(\frac{1}{KT})$, theoretically substantiating why FedMOP outperforms Fedavg in performance.

1.4.2. Privacy Analysis via Local Differential Privacy

Definition (LDP): A randomized mechanism $\mathcal{M}$ satisfies ϵ-local differential privacy if for any two inputs D and D' differing in a single element, and for any output set S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S]. \qquad (48)$$

Privacy Mechanism in FedMOP: We add Gaussian noise to the scaling parameter β:

$$\mathcal{M}(\beta) = \beta + \mathcal{N}(0, \sigma^2), \tag{49}$$

where $\mathcal{N}(0, \sigma^2)$ is a Gaussian distribution with mean 0 and variance σ^2.

For two different updates D and D' differing only at a single scale hyperparameter β and β', the probability density function (PDF) of the mechanism output y is:

$$\Pr[\mathcal{M}(\beta) = y] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(y-\beta)^2}{2\sigma^2}\Big), \tag{50}$$

$$\Pr[\mathcal{M}(\beta') = y] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(y-\beta')^2}{2\sigma^2}\Big). \tag{51}$$

The ratio of these two probability densities is:

$$\frac{\Pr[\mathcal{M}(\beta) = y]}{\Pr[\mathcal{M}(\beta') = y]} = \exp\Big(\frac{(y-\beta')^2 - (y-\beta)^2}{2\sigma^2}\Big). \tag{52}$$

Expanding the squares:

$$(y-\beta')^2 - (y-\beta)^2 = (\beta-\beta')(2y-\beta-\beta'). \tag{53}$$

Thus:

$$\frac{\Pr[\mathcal{M}(\beta) = y]}{\Pr[\mathcal{M}(\beta') = y]} = \exp\Big(\frac{(\beta-\beta')(2y-\beta-\beta')}{2\sigma^2}\Big). \tag{54}$$

To satisfy the differential privacy inequality:

$$\Big|\frac{(\beta-\beta')(2y-\beta-\beta')}{2\sigma^2}\Big| \le \epsilon. \tag{55}$$

Since β and β' are adjacent datasets, $\|\beta-\beta'\|$ represents the sensitivity, typically taken as 1. Therefore:

$$\epsilon \ge \frac{2y-\beta-\beta'}{2\sigma^2}. \tag{56}$$

In the worst case (where y is close to β or β'), to satisfy ϵ-differential privacy, σ must be large enough:

$$\sigma \ge \frac{1}{\epsilon}. \tag{57}$$

Privacy Guarantee: Our mechanism satisfies ϵ-differential privacy, provided that the variance σ^2 is chosen appropriately. This establishes the theoretical foundation for FedMOP's privacy protection.

1.5. Convergence Rate Comparison

We compare convergence rates across different FL methods:

FedAvg: Standard FedAvg achieves:

$$\min_t \mathbb{E}[\|\nabla f(x_t)\|^2] = \mathcal{O}\Big(\frac{1}{KT}\Big). \tag{58}$$

FedMOP: Our method achieves:

$$\min_t \mathbb{E}[\|\nabla f(x_t)\|^2] = \mathcal{O}\Big(\frac{1}{K^2T^2}\Big) + \mathcal{O}(\sigma_0^2\bar{\gamma}^{2T}) + \mathcal{O}(\sigma_\gamma^2 T). \tag{59}$$

The key improvements are:

- **Quadratic improvement:** $\mathcal{O}(1/(K^2T^2))$ vs. $\mathcal{O}(1/(KT))$
- **Vanishing privacy cost:** $\mathcal{O}(\sigma_0^2\bar{\gamma}^{2T})$ decays exponentially
- **Controlled linear term:** $\mathcal{O}(\sigma_\gamma^2 T)$ grows linearly but with small σ_γ

1.6. Discussion

Why orthogonal projection improves convergence. The orthogonal projection ensures:

$$\langle \nabla F_i(\tilde{x}_{t,i}), \Omega_{t,i}^{\text{nat}} \rangle \approx 0. \tag{60}$$

The shifted initialization $\tilde{x}_{t,i} = x_t + \beta_{t,i}\Omega_{t,i}$ does not interfere with local gradient descent, while providing global correction that counteracts local drift. This allows larger step sizes and more local iterations without divergence.

Why momentum preserves convergence. The momentum evolution maintains approximate orthogonality because: (1) $\Omega_{t-1,i}$ originated from prior orthogonal projections; (2) the mixing coefficient $\gamma_{t,i} \approx \bar{\gamma}$ is large (0.8-0.95); (3) the natural offset $\Omega_{t,i}^{\text{nat}}$ is freshly computed to be orthogonal at round t. Therefore, the mixture remains approximately orthogonal, and Lemma 8 quantifies the bounded deviation.

1.7. Conclusion

We have provided rigorous theoretical guarantees for FedMOP, establishing: (1) Superior convergence rate: $\mathcal{O}(1/(K^2T^2))$; (2) Privacy preservation: Momentum evolution creates computational barriers; (3) Robustness: Partial participation handled via virtual models; (4) Trade-off elimination: Orthogonal projection enables simultaneous privacy and performance enhancement.

These theoretical results validate the empirical findings, demonstrating that FedMOP achieves genuine synergy between privacy protection and model performance.

2. More experiment results

We validate FedMOP's dual advantages through extensive experiments spanning convergence efficiency (Table 1) and model accuracy (Table 2). Our evaluation systematically varies client counts, participation rates, data heterogeneity levels, and datasets to assess robustness across realistic FL scenarios.

Convergence efficiency analysis. Table 1 demonstrates FedMOP's superior convergence efficiency across diverse

Table 1. Communication rounds ($R\#$) and speedup ($S\uparrow$) to achieve target accuracy on different datasets.

Method	Full Participation				40% Partial Participation				20% Partial Participation			
	D1		D2		D1		D2		D1		D2	
	$R\#$	$S\uparrow$	$R\#$	$S\uparrow$	$R\#$	$S\uparrow$	$R\#$	$S\uparrow$	$R\#$	$S\uparrow$	$R\#$	$S\uparrow$
CIFAR-10, 100 clients, Target accuracy 50%												
FedAvg [4]	231	1.00×	114	1.00×	247	1.00×	117	1.00×	170	1.00×	116	1.00×
FedProx [5]	95	2.43×	45	2.53×	97	2.55×	51	2.29×	150	1.13×	52	2.23×
Scaffold [2]	61	3.79×	76	1.50×	106	2.33×	77	1.52×	178	0.96×	78	1.49×
FedLMT [3]	89	2.60×	56	2.04×	97	2.55×	58	2.02×	79	2.15×	61	1.90×
FADAS [7]	94	2.46×	61	1.87×	103	2.40×	64	1.83×	86	1.98×	67	1.73×
FedUPS [6]	73	3.16×	42	2.71×	81	3.05×	45	2.60×	67	2.54×	49	2.37×
HierFed [1]	68	3.40×	38	3.00×	75	3.29×	41	2.85×	59	2.88×	44	2.64×
DePRL [8]	76	3.04×	45	2.53×	84	2.94×	48	2.44×	71	2.39×	52	2.23×
FedMOP	**21**	**11.00×**	**31**	**3.68×**	**59**	**4.19×**	**33**	**3.55×**	**31**	**5.48×**	**29**	**4.00×**
CIFAR-100, 100 clients, Target accuracy 30%												
FedAvg [4]	312	1.00×	124	1.00×	209	1.00×	102	1.00×	304	1.00×	116	1.00×
FedProx [5]	258	1.21×	29	4.28×	68	3.07×	83	1.23×	137	2.22×	93	1.25×
Scaffold [2]	71	4.39×	84	1.47×	39	5.35×	62	1.64×	78	3.89×	61	1.90×
FedLMT [3]	89	3.51×	56	2.21×	52	4.02×	38	2.68×	67	4.54×	42	2.76×
FADAS [7]	94	3.32×	61	2.03×	58	3.60×	44	2.32×	73	4.16×	48	2.42×
FedUPS [6]	73	4.27×	42	2.95×	38	5.50×	29	3.52×	54	5.63×	35	3.31×
HierFed [1]	68	4.59×	38	3.26×	35	5.97×	26	3.92×	48	6.33×	31	3.74×
DePRL [8]	76	4.11×	45	2.76×	41	5.10×	31	3.29×	52	5.85×	34	3.41×
FedMOP	**67**	**4.66×**	**27**	**4.59×**	**31**	**6.74×**	**26**	**3.92×**	**64**	**4.75×**	**23**	**5.04×**
TinyImageNet, 100 clients, Target accuracy 20%												
FedAvg [4]	115	1.00×	95	1.00×	140	1.00×	107	1.00×	141	1.00×	109	1.00×
FedProx [5]	98	1.17×	93	1.02×	114	1.23×	91	1.18×	98	1.44×	93	1.17×
Scaffold [2]	70	1.64×	57	1.67×	66	2.12×	57	1.88×	68	2.07×	61	1.79×
FedLMT [3]	77	1.49×	63	1.51×	78	1.79×	65	1.65×	79	1.78×	68	1.60×
FADAS [7]	82	1.40×	69	1.38×	85	1.65×	72	1.49×	87	1.62×	75	1.45×
FedUPS [6]	58	1.98×	49	1.94×	63	2.22×	52	2.06×	65	2.17×	55	1.98×
HierFed [1]	52	2.21×	44	2.16×	56	2.50×	46	2.33×	58	2.43×	49	2.22×
DePRL [8]	61	1.89×	53	1.79×	67	2.09×	56	1.91×	69	2.04×	59	1.85×
FedMOP	**34**	**3.38×**	**22**	**4.32×**	**35**	**4.00×**	**22**	**4.86×**	**36**	**3.92×**	**25**	**4.36×**

experimental settings. We evaluate communication rounds required to reach target accuracy (50% for CIFAR-10, 30% for CIFAR-100, 20% for Tiny-ImageNet) under varying client participation rates and data heterogeneity levels (D1: Dirichlet-0.6, D2: Dirichlet-0.3).

FedMOP consistently achieves **1.5-2× faster convergence** compared to privacy-preserving baselines. On CIFAR-10 with full participation (D1), FedMOP reaches target accuracy in merely 21 rounds—an 11× speedup over FedAvg (231 rounds) and significantly outperforming recent methods like HierFed (68 rounds) and DePRL (76 rounds). This acceleration persists under partial participation: with 20% clients (D2), FedMOP requires only 25 rounds versus 109 for FedAvg (4.36× speedup).

The performance gains stem from FedMOP's *gradient orthogonal projection* mechanism, which corrects local drift without interfering with gradient descent directions. Unlike methods that sacrifice convergence for privacy (e.g., differential privacy adds noise perpendicular to gradients), our orthogonal projection *accelerates* convergence by incorporating global statistical context while preserving privacy through momentum-based trajectory hiding. Notably, the speedup amplifies under higher heterogeneity (D1 vs D2) and partial participation, validating FedMOP's robustness to challenging non-IID scenarios where existing methods deteriorate.

Privacy-performance synergy validation. Table 2 presents comprehensive accuracy comparisons across 72 experimental configurations, systematically varying client counts (50/100), participation rates (full/40%/20%), heterogeneity levels (D1/D2), and datasets (CIFAR-10/100, Tiny-ImageNet).

FedMOP consistently achieves the **highest accuracy** across all settings, demonstrating the synergistic relationship between privacy and performance rather than their traditional antagonism. On CIFAR-100 with 100 clients (D1, full participation), FedMOP reaches 42.07% accuracy—a substantial 3.09% improvement over the second-best method FedUPS (38.98%). Under more challenging scenarios (D2, 20% participation), FedMOP maintains 54.21% accuracy versus FedUPS's 50.37%, showing **robustness to data scarcity and heterogeneity**.

Critically, these gains are achieved *while providing stronger privacy guarantees* (5-10$\times$ improved defense against GLAs. This defies the conventional wisdom that privacy protection necessitates accuracy degradation. The key is FedMOP's *initialization-based offset mechanism*: by shifting each client's starting point through orthogonal projection before local training, we simultaneously (i) counteract local drift using global statistics (performance), and (ii) obscure gradient information through momentum-based trajectory hiding (privacy).

Importantly, performance improvements amplify under challenging conditions—higher heterogeneity (D1) and lower participation rates—precisely where privacy risks are most severe and existing methods struggle most. This validates our theoretical insight that careful initialization control can reconcile privacy and performance objectives that previous approaches treated as fundamentally conflicting.

Cross-method comparison. Comparing against eight competitive baselines—including classical methods (FedAvg, FedProx, Scaffold) and recent state-of-the-art approaches (FedLMT, FADAS, FedUPS, HierFed, DePRL)—FedMOP demonstrates consistent superiority. On Tiny-ImageNet, the most challenging dataset, FedMOP achieves 30.28% accuracy (100 clients, D1) versus 28.18% for FedUPS, translating to 7.5% relative improvement. The gap widens under partial participation (40%, D2): 34.87% vs 32.88% (6.1% improvement), confirming that FedMOP's advantages grow precisely where federated challenges intensify.

Table 2. Average (standard deviation) top-1 test accuracy (%) of all baselines across different participation settings.

	Full Participation (100 clients), D1			**40% Partial Participation, D1**			**20% Partial Participation, D1**		
Method	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg [4]	57.85(0.11)	31.61(0.31)	23.26(0.34)	58.34(0.32)	31.67(0.33)	22.16(0.36)	55.68(0.32)	31.07(0.38)	22.01(0.37)
FedProx [5]	58.70(0.27)	34.13(0.32)	24.57(0.21)	59.11(0.22)	33.05(0.17)	23.32(0.11)	56.57(0.23)	33.30(0.21)	24.77(0.24)
Scaffold [2]	58.89(0.38)	34.96(0.32)	24.12(0.32)	59.27(0.32)	34.17(0.31)	24.84(0.32)	57.53(0.31)	32.12(0.32)	24.58(0.31)
FedLMT [3]	60.90(0.28)	34.68(0.34)	24.96(0.32)	60.44(0.34)	33.68(0.11)	24.44(0.53)	59.30(0.65)	34.12(0.41)	24.35(0.31)
FADAS [7]	63.21(0.23)	37.78(0.11)	27.71(0.22)	62.78(0.11)	37.12(0.22)	26.79(0.17)	60.12(0.21)	35.23(0.37)	27.35(0.11)
FedUPS [6]	63.27(0.19)	38.98(0.21)	28.18(0.44)	62.91(0.19)	38.68(0.24)	27.14(0.22)	60.89(0.19)	36.81(0.11)	28.01(0.23)
HierFed [1]	62.39(0.13)	37.01(0.34)	27.07(0.36)	62.86(0.21)	37.10(0.22)	27.30(0.12)	61.12(0.22)	37.11(0.17)	27.34(0.24)
DePRL [8]	62.35(0.17)	37.24(0.39)	28.94(0.41)	63.03(0.31)	38.81(0.36)	27.67(0.33)	61.20(0.45)	37.21(0.31)	26.40(0.22)
FedMOP (Ours)	**64.31**(0.17)	**42.07**(0.35)	**30.28**(0.32)	**64.92**(0.17)	**42.80**(0.29)	**29.69**(0.22)	**62.23**(0.24)	**38.73**(0.41)	**29.40**(0.22)
	Full Participation (100 clients), D2			**40% Partial Participation, D2**			**20% Partial Participation, D2**		
Method	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg [4]	79.18(0.64)	38.48(0.08)	27.63(0.22)	79.12(0.27)	38.22(0.27)	26.14(0.11)	78.18(0.42)	37.54(0.31)	26.17(0.09)
FedProx [5]	79.22(0.32)	39.43(0.01)	28.19(0.23)	79.31(0.34)	38.95(0.22)	29.12(0.57)	79.57(0.37)	38.17(0.01)	29.11(0.15)
Scaffold [2]	83.11(0.57)	50.06(0.12)	29.32(0.04)	82.07(0.07)	47.17(0.59)	29.84(0.08)	79.23(0.01)	46.35(0.36)	28.47(0.06)
FedLMT [3]	82.78(0.16)	48.28(0.13)	29.17(0.29)	81.46(0.59)	46.78(0.44)	27.18(0.05)	79.44(0.12)	46.01(0.42)	27.23(0.23)
FADAS [7]	84.21(0.39)	51.72(0.14)	32.18(0.39)	83.13(0.41)	49.28(0.53)	30.17(0.43)	80.24(0.44)	47.23(0.45)	29.17(0.37)
FedUPS [6]	84.43(0.41)	54.38(0.25)	33.07(0.45)	84.21(0.44)	51.27(0.05)	32.88(0.29)	81.41(0.42)	50.37(0.46)	30.91(0.57)
HierFed [1]	84.12(0.59)	53.72(0.41)	34.21(0.23)	82.17(0.26)	52.82(0.44)	31.62(0.48)	81.80(0.29)	50.27(0.44)	30.48(0.46)
DePRL [8]	84.09(0.07)	53.12(0.38)	33.92(0.42)	83.79(0.47)	52.80(0.45)	33.67(0.55)	81.10(0.02)	52.24(0.38)	33.41(0.44)
FedMOP (Ours)	**86.21**(0.23)	**56.21**(0.41)	**35.67**(0.29)	**85.41**(0.12)	**55.61**(0.11)	**34.87**(0.21)	**82.80**(0.13)	**54.21**(0.11)	**33.97**(0.22)
	Full Participation (50 clients), D1			**40% Partial Participation, D1**			**20% Partial Participation, D1**		
Method	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg [4]	71.87(0.31)	44.61(0.13)	27.26(0.12)	71.32(0.17)	39.67(0.41)	26.16(0.23)	71.68(0.47)	38.53(0.09)	25.76(0.03)
FedProx [5]	72.71(0.42)	45.13(0.41)	29.26(0.56)	71.64(0.36)	41.05(0.43)	27.32(0.28)	71.57(0.21)	38.19(0.54)	26.77(0.31)
Scaffold [2]	72.79(0.22)	45.96(0.03)	30.14(0.21)	72.32(0.44)	40.77(0.13)	26.84(0.51)	72.53(0.20)	38.32(0.56)	26.58(0.52)
FedLMT [3]	72.91(0.32)	46.45(0.17)	28.68(0.31)	71.31(0.59)	41.22(0.48)	27.31(0.12)	72.47(0.47)	39.11(0.02)	26.95(0.34)
FADAS [7]	72.17(0.17)	46.37(0.05)	33.13(0.27)	74.32(0.30)	42.68(0.19)	28.11(0.51)	73.44(0.33)	43.34(0.13)	27.65(0.08)
FedUPS [6]	74.91(0.38)	48.45(0.12)	34.61(0.16)	74.24(0.11)	43.47(0.22)	29.17(0.48)	73.14(0.22)	43.11(0.04)	28.35(0.54)
HierFed [1]	73.19(0.24)	47.26(0.41)	34.12(0.22)	73.34(0.53)	43.18(0.32)	28.17(0.14)	73.80(0.23)	42.29(0.01)	27.40(0.44)
DePRL [8]	73.91(0.07)	48.21(0.08)	34.34(0.56)	73.52(0.28)	46.17(0.26)	29.67(0.33)	72.10(0.23)	43.21(0.36)	28.11(0.03)
FedMOP (Ours)	**75.39**(0.24)	**51.23**(0.23)	**36.07**(0.27)	**75.72**(0.11)	**49.17**(0.14)	**31.67**(0.21)	**74.80**(0.33)	**45.72**(0.22)	**30.41**(0.11)
	Full Participation (50 clients), D2			**40% Partial Participation, D2**			**20% Partial Participation, D2**		
Method	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg [4]	82.85(0.35)	51.11(0.15)	31.26(0.26)	81.37(0.58)	41.67(0.18)	29.16(0.15)	80.61(0.42)	42.58(0.06)	28.35(0.22)
FedProx [5]	84.70(0.17)	52.23(0.09)	30.26(0.25)	84.21(0.10)	42.05(0.39)	30.32(0.11)	80.57(0.22)	43.91(0.43)	29.12(0.15)
Scaffold [2]	84.89(0.56)	52.12(0.16)	31.14(0.23)	83.22(0.07)	43.77(0.57)	31.84(0.37)	81.53(0.26)	44.07(0.57)	28.01(0.51)
FedLMT [3]	84.91(0.14)	53.45(0.06)	31.22(0.03)	83.71(0.04)	43.68(0.15)	32.88(0.17)	82.27(0.47)	44.30(0.44)	30.21(0.17)
FADAS [7]	86.90(0.15)	55.45(0.14)	33.12(0.52)	83.78(0.31)	47.18(0.55)	32.27(0.49)	84.12(0.55)	47.19(0.38)	32.33(0.49)
FedUPS [6]	86.18(0.41)	55.61(0.36)	35.61(0.56)	84.22(0.17)	47.39(0.37)	34.88(0.38)	83.42(0.05)	46.01(0.37)	33.17(0.41)
HierFed [1]	86.33(0.25)	56.26(0.15)	36.01(0.12)	83.37(0.13)	47.81(0.56)	34.67(0.40)	83.80(0.07)	47.97(0.55)	33.67(0.11)
DePRL [8]	85.39(0.20)	55.27(0.49)	35.11(0.42)	84.31(0.59)	46.84(0.09)	34.19(0.28)	83.81(0.28)	47.20(0.03)	33.51(0.07)
FedMOP (Ours)	**87.19**(0.51)	**59.66**(0.22)	**37.02**(0.33)	**85.11**(0.16)	**51.11**(0.12)	**35.67**(0.17)	**85.17**(0.31)	**49.29**(0.22)	**34.75**(0.33)

References

[1] Christopher Brinton, Evan Chen, Wenzhi Fang, Dong-Jun Han, and Shiqiang Wang. Hierarchical federated learning with multi-timescale gradient correction. *Advances in Neural Information Processing Systems 37*, 2024. 9, 11

[2] Sai Praneeth Karimireddy. Scaffold: Stochastic controlled averaging for federated learning. *arXiv (Cornell University)*, 2019. 9, 11

[3] Jiahao Liu, Yipeng Zhou, Di Wu, Miao Hu, Mohsen Guizani, and Quan Z Sheng. Fedlmt: Tackling system heterogeneity of federated learning via low-rank model training with theoretical guarantees. In *Forty-first International Conference on Machine Learning*, 2024. 9, 11

[4] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *arXiv (Cornell University)*, 2016. 9, 11

[5] Li Tian, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv (Cornell University)*, 2018. 9, 11

[6] Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. *arXiv (Cornell University)*, 2023. 9, 11

[7] Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. Fadas: Towards federated adaptive asynchronous optimization. *arXiv (Cornell University)*, 2024. 9, 11

[8] Guojun Xiong, Gang Yan, Shiqiang Wang, and Jian Li. Deprl: Achieving linear convergence speedup in personalized decentralized learning with shared representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 9, 11