

A APPENDIX

A.1 PRIVACY ANALYSIS: PROOF FOR THEOREM 1 AND 2

In the proofs of Theorem 1 and 2 we follow the line of reasoning in Abadi et al. (2016), which analyzes the privacy of DPSGD. We first state the following lemma from Abadi et al. (2016).

Lemma 2. (Abadi et al., 2016, Theorem 1) *There exists constants c_1 and c_2 such that given the sampling probability $p = \frac{q}{m}$ and the number of steps T , for any $\epsilon < c_1 p^2 T$, DPSGD is (ϵ, δ) -differentially private for any $\delta > 0$ if we choose $\sigma \geq c_2 \frac{p\sqrt{T \log(1/\delta)}}{\epsilon}$.*

To prove Theorem 1, we also need the following definitions and lemmas.

Definition 3 (ℓ_2 -sensitivity). *Let $f : \mathcal{U} \rightarrow \mathbb{R}^d$ be some arbitrary function, the ℓ_2 -sensitivity of f is defined as*

$$\Delta_2 f = \max_{\text{adjacent } D, D' \in \mathcal{U}} \|f(D) - f(D')\|_2 \quad (15)$$

Definition 4 (Rényi Divergence). (Mironov, 2017, Definition 3) *Let P, Q be two probability distribution over the same probability space, and let p, q be the respective probability density function. The Rényi Divergence with finite order $\alpha \neq 1$ is:*

$$D_\alpha(P \| Q) = \frac{1}{\alpha - 1} \ln \int_{\mathcal{X}} q(x) \left(\frac{p(x)}{q(x)} \right)^\alpha dx \quad (16)$$

Definition 5 ((α, ϵ) -Rényi Differential Privacy). (Mironov, 2017, Definition 4) *A randomized mechanism $f : \mathcal{D} \rightarrow \mathcal{R}$ is said to have (α, ϵ) -Rényi Differential Privacy if for all adjacent $D, D' \in \mathcal{D}$ it holds that:*

$$D_\alpha(f(D) \| f(D')) \leq \epsilon. \quad (17)$$

Lemma 3. (Mironov, 2017, Corollary 3) *The Gaussian mechanism is $(\alpha, \alpha(2(\Delta_2 f)^2 / \sigma^2))$ -Rényi Differentially Private.*

Lemma 4. (Mironov, 2017, Proposition 3) *If f is (α, ϵ) -RDP, then it is $(\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta)$ -DP for all $\delta > 0$.*

We begin by proving the first part of Theorem 1, where $q \neq m$.

Proof for Theorem 1: $q \neq m$. Note that aggregation step in line 8 of Algorithm 1 can be rewritten as

$$\tilde{w}^{t+1} = \tilde{w}^t + \frac{1}{|S_t|} \sum_{k \in S_t} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) + \mathcal{N}(0, \sigma^2 \mathbf{I}_{\mathbf{d} \times \mathbf{d}}) \quad (18)$$

$$= \tilde{w}^t + \frac{1}{q} \sum_{k \in S_t} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) + \mathcal{N} \left(0, \left(\frac{\sigma}{\gamma} \right)^2 \gamma^2 \mathbf{I}_{\mathbf{d} \times \mathbf{d}} \right) \quad (19)$$

$$= \tilde{w}^t + \frac{1}{q} \left(\sum_{k \in S_t} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) + \mathcal{N} \left(0, \left(\frac{q\sigma}{\gamma} \right)^2 \gamma^2 \mathbf{I}_{\mathbf{d} \times \mathbf{d}} \right) \right). \quad (20)$$

From here, we can directly apply Lemma 2 with σ set to be $\frac{q\sigma}{\gamma}$. Hence, we conclude that when $q \neq m$, there exists constants c_1 and c_2 such that given the number of steps T , for any $\epsilon < c_1 \frac{q^2}{m^2} T$, $\mathcal{M}^{1:T}$ is (ϵ, δ) -differentially private for any $\delta > 0$ if we choose $\sigma \geq c_2 \frac{\gamma\sqrt{T \log(1/\delta)}}{m\epsilon}$. \square

This proof can extend to the case where $q = m$. In the remainder of this section, we provide a proof that gives a more specific bound on the variance σ^2 in the case where $q = m$.

Proof for Theorem 1: $q = m$. Define $H^t : \prod_{i=1}^m \mathcal{D}_i \times \mathcal{W} \rightarrow \mathcal{W}$ as

$$H^t(\{D_i\}, \{h_i(\cdot)\}, \tilde{w}^t) = \tilde{w}^t + \frac{1}{m} \sum_{i=1}^m h_i^t(D_i, \tilde{w}^t). \quad (21)$$

As a result, we have $\mathcal{M}^t(\{D_i\}, \{h_i(\cdot)\}, \tilde{w}^t, \sigma) = H^t(\{D_i\}, \{h_i(\cdot)\}, \tilde{w}^t) + \beta^t$.

By Lemma 3, \mathcal{M}^t is $(\alpha, 2\alpha(\Delta_2 H^t)^2/d\sigma^2)$ -Renyi Differentially Private. Note that

$$(\Delta_2 H^t)^2 = \max_j \max_{\text{adjacent } D_j, D'_j \in \mathcal{D}_j} \|H^t(\{D_1, \dots, D_j, \dots, D_m\}) - H^t(\{D_1, \dots, D'_j, \dots, D_m\})\|^2 \quad (22)$$

$$= \max_j \max_{\text{adjacent } D_j, D'_j \in \mathcal{D}_j} \left\| \frac{1}{m} h_j^t(D_j, \tilde{w}^t) - \frac{1}{m} h_j^t(D'_j, \tilde{w}^t) \right\|^2 \quad (23)$$

$$= \frac{1}{m^2} \max_j \max_{\text{adjacent } D_j, D'_j \in \mathcal{D}_j} \|h_j^t(D_j, \tilde{w}^t) - h_j^t(D'_j, \tilde{w}^t)\|^2 \quad (24)$$

$$= \frac{1}{m^2} \max_j (\Delta_2 h_j^t)^2. \quad (25)$$

Hence, by sequential composition of Rényi Differential Privacy (Mironov, 2017, Proposition 1), $\mathcal{M}^{1:T}$ is $(\alpha, \sum_{i=1}^T 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2)$ -RDP.

By Lemma 4, we know that $\mathcal{M}^{1:T}$ is $(\sum_{i=1}^T 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha-1}, \delta)$ -DP.

Plugging in $\alpha = \frac{4\log(1/\delta)}{\epsilon}$, $\sigma = \frac{4\gamma\sqrt{T\log(1/\delta)}}{\epsilon m}$, we have

$$\sum_{i=1}^T 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha-1} \leq \sum_{i=1}^T 2\alpha\gamma^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha-1} \quad (26)$$

$$= \frac{2\frac{4\log(1/\delta)}{\epsilon}\gamma^2}{m^2\left(\frac{4\gamma\sqrt{T\log(1/\delta)}}{\epsilon m}\right)^2} + \frac{\log(1/\delta)}{\frac{4\log(1/\delta)}{\epsilon} - 1} \quad (27)$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \quad (28)$$

$$= \epsilon. \quad (29)$$

Hence, $\mathcal{M}^{1:T}$ is (ϵ, δ) -JDP if we choose $\sigma = \frac{4\gamma\sqrt{T\log(1/\delta)}}{\epsilon m}$. \square

By Theorem 1 and *Billboard Lemma*, it directly follows that Algorithm 1 is (ϵ, δ) -JDP.

Proof for Theorem 2. Theorem 1 shows that Algorithm 1 consists of a (ϵ, δ) -DP process to produce global model. After that each task learner trains local model with the DP global model and its private data. By Lemma 1, it directly follows that Algorithm 1 is (ϵ, δ) -JDP. \square

A.2 CONVERGENCE ANALYSIS(NONCONVEX):

We first present the formal statement of Theorem 3.

Theorem 7 (Convergence under nonconvex loss). *Let f_k be $(L + \lambda)$ -smooth. Assume γ is sufficiently large such that $\gamma \geq \max_{k,t} \|\nabla_{w_k^t} f_k(w_k^t; \tilde{w}^t)\|_2$. Further let $f_k^* = \min_{w, \bar{w}} f_k(w; \bar{w})$ and $p = \frac{q}{m}$. If we use a fixed learning rate $\eta_t = \eta = \frac{1}{pL + (p - \frac{1}{p})\lambda}$, Algorithm 1 satisfies:*

$$\begin{aligned} \frac{1}{mT} \sum_{t=0}^{T-1} \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 &\leq \frac{4\left(L + \lambda + \frac{\lambda}{p^2}\right) \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{mT} \\ &\quad + \frac{\mathcal{O}\left(L + \lambda + \frac{\lambda}{p^2}\right) \sum_{t=0}^{T-1} B_{t+1}}{T} + \mathcal{O}\left(Ld\lambda + d\lambda^2 + \frac{d\lambda^2}{p^2}\right) \sigma^2. \end{aligned} \quad (30)$$

where

$$B_t = \max_k f_k(w_k^t; \tilde{w}^t). \quad (31)$$

Let σ chosen as we set in Theorem 2. Take $T = \mathcal{O}\left(\frac{m}{\lambda d \gamma^2}\right)$, the right hand side is bounded by

$$\begin{aligned} \frac{\lambda d \gamma^2}{m^2} \sum_{t=0}^{T-1} \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 &\leq \frac{4 \left(L + \lambda + \frac{\lambda}{p^2}\right) \lambda d \gamma^2 \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{m^2} \\ &\quad + \mathcal{O}\left(\frac{L + \lambda + \frac{\lambda}{p^2}}{m}\right) \lambda d \gamma^2 \sum_{t=0}^{T-1} B_{t+1} + \mathcal{O}\left(\frac{L + \lambda + \frac{\lambda}{p^2}}{m}\right) \frac{\log(1/\delta)}{\epsilon^2}. \end{aligned} \quad (32)$$

Proof for Theorem 7. Let $w_k^* = \arg \min_w f_k(w; \bar{w}^*)$. Let I_k^t be the random variable indicating whether task k is selected in communication round t . Note that the probability task learner k is

selected in any arbitrary communication round $p_k = \frac{\binom{m-1}{q-1}}{\binom{m}{q}} = \frac{q}{m}$. Thus $\mathbb{E}[I_k^t] = p_k = \frac{q}{m}$. By

L -smoothness of f_k , we have

$$\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^t) - f_k(w_k^t; \tilde{w}^t)] \leq \mathbb{E}\left[\langle \nabla f_k(w_k^t; \tilde{w}^t), w_k^{t+1} - w_k^t \rangle + \frac{L}{2} \|w_k^{t+1} - w_k^t\|^2\right] \quad (33)$$

$$= \mathbb{E}\left[\langle \nabla f_k(w_k^t; \tilde{w}^t), \eta_t I_k^t \nabla f_k(w_k^t; \tilde{w}^t) \rangle + \frac{L}{2} \|\eta_t I_k^t \nabla f_k(w_k^t; \tilde{w}^t)\|^2\right] \quad (34)$$

$$= \left(\frac{L + \lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k\right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2. \quad (35)$$

Hence, we have

$$\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^t; \tilde{w}^t)] \leq \underbrace{\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^{t+1}; \tilde{w}^t)]}_{\mathbf{B}} + \left(\frac{L}{2} \eta_t^2 p_k^2 - \eta_t p_k\right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2. \quad (36)$$

It suffices to bound \mathbf{B} :

$$\mathbf{B} = \mathbb{E}\left[\frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^{t+1}\|^2 - \frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^t\|^2\right] \quad (37)$$

$$= \frac{\lambda}{2} \mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\| \|2w_k^{t+1} - \tilde{w}^t - \tilde{w}^{t+1}\|] \quad (38)$$

$$\leq \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2] \mathbb{E}[\|2w_k^{t+1} - \tilde{w}^t - \tilde{w}^{t+1}\|^2]} \quad (39)$$

$$= \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2]} \sqrt{\mathbb{E}[\|(\tilde{w}^{t+1} - \tilde{w}^t) + 2(w_k^{t+1} - \tilde{w}^{t+1})\|^2]} \quad (40)$$

$$\leq \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2]} \sqrt{\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] + 4\|w_k^{t+1} - \tilde{w}^{t+1}\|^2 + 4\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\| \|w_k^{t+1} - \tilde{w}^{t+1}\|]} \quad (41)$$

$$\leq \frac{\lambda}{2} \sqrt{\underbrace{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2]}_{C_1}} \sqrt{\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] + 4\|w_k^{t+1} - \tilde{w}^{t+1}\|^2 + 4\sqrt{\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] \|w_k^{t+1} - \tilde{w}^{t+1}\|^2}} \quad (42)$$

where the first and third inequality follows from Cauchy-Schwartz Inequality: $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. We can then upper bound C_1 and C_2 .

$$C_1 = \mathbb{E} \left[\left\| \frac{1}{q} \sum_{k \in S_t} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) + \beta^t \right\|^2 \right] \quad (43)$$

$$\leq \left(\sqrt{\mathbb{E} \left[\left\| \frac{1}{q} \sum_{k=1}^m I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{\mathbb{E}[\|\beta^t\|^2]} \right)^2 \quad (44)$$

$$= \left(\sqrt{\mathbb{E} \left[\left\| \frac{1}{q} \sum_{k=1}^m I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{d}\sigma \right)^2 \quad (45)$$

$$\leq \left(\sqrt{\frac{m}{q^2} \sum_{k=1}^m \mathbb{E} \left[\left\| I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{d}\sigma \right)^2 \quad (46)$$

$$\leq \left(\sqrt{\frac{m}{q^2} \sum_{k=1}^m \mathbb{E} \left[\left\| I_k^{t+1} \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2 \right]} + \sqrt{d}\sigma \right)^2 \quad (47)$$

$$\leq \left(\sqrt{\frac{1}{m} \sum_{k=1}^m \left\| \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2} + \sqrt{d}\sigma \right)^2. \quad (48)$$

Denote $h(t) = \sqrt{\frac{1}{m} \sum_{k=1}^m \left\| \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2}$. We have:

$$C_2 \leq \frac{2}{\lambda} \frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^{t+1}\|^2 \quad (49)$$

$$\leq \frac{2}{\lambda} f_k(w_k^{t+1}; \tilde{w}^{t+1}) \quad (50)$$

$$= \frac{2}{\lambda} B_{t+1} \quad (51)$$

Plugging the bounds for C_1 and C_2 into B yields:

$$B \leq \frac{\lambda}{2} (h(t) + \sqrt{d}\sigma) \left(h(t) + \sqrt{d}\sigma + 2\sqrt{\frac{2}{\lambda} B_{t+1}} \right). \quad (52)$$

Denote the right hand side as $\beta(t)$, we have

$$\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^t; \tilde{w}^t)] \leq \beta(t) + \left(\frac{L+\lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k \right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2. \quad (53)$$

Let $\delta_t = \mathbb{E}[f_k(w_k^t; \tilde{w}^t) - f_k(w_k^*; \tilde{w}^*)]$, we have

$$\delta_{t+1} \leq \delta_t + \beta(t) + \left(\frac{L+\lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k \right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2. \quad (54)$$

In the nonconvex case, we have

$$\sum_{t=0}^{T-1} \left(\eta_t p_k - \frac{L+\lambda}{2} \eta_t^2 p_k^2 \right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 - \beta(t) \leq f_k(w_k^0; \tilde{w}^0) - f_k^* \quad (55)$$

Summing over k on the left handed side, when γ is large enough so that no clipping happens we have

$$\sum_{k=1}^m \sum_{t=0}^{T-1} \left(\eta_t p_k - \frac{L+\lambda}{2} \eta_t^2 p_k^2 \right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 - \beta(t) \quad (56)$$

$$= \sum_{t=0}^{T-1} \left(\eta_t \frac{q}{m} - \frac{L+\lambda}{2} \eta_t^2 \frac{q^2}{m^2} \right) \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 - m\beta(t) \quad (57)$$

$$= \sum_{t=0}^{T-1} \left(\eta_t \frac{q}{m} - \frac{L+\lambda}{2} \eta_t^2 \frac{q^2}{m^2} \right) \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 - \frac{\lambda}{2} \left(mh^2(t) + \left(2\sqrt{d}\sigma + 2\sqrt{\frac{2}{\lambda} B_{t+1}} \right) mh(t) + m \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) \right) \quad (58)$$

$$= \sum_{t=0}^{T-1} \left(\eta_t \frac{q}{m} - \frac{L+\lambda}{2} \eta_t^2 \frac{q^2}{m^2} - \frac{\lambda}{2} \eta_t^2 \right) G_t^2 - \lambda\sqrt{m} \left(\sqrt{d}\sigma + \sqrt{\frac{2}{\lambda} B_{t+1}} \right) \eta_t G_t - \frac{\lambda m}{2} \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) \quad (59)$$

$$\leq \sum_{k=1}^m f_k(w_k^0; \tilde{w}^0) - f_k^*, \quad (60)$$

where $G_t = \sqrt{\sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2}$. Picking $\eta_t = \frac{mq}{qL - (m-q)\lambda}$ yields

$$\sum_{t=0}^{T-1} \frac{q^2}{2(q^2 L - (m^2 - q^2)\lambda)} G_t^2 - \frac{\lambda\sqrt{m} \left(\sqrt{d}\sigma + \sqrt{\frac{2}{\lambda} B_{t+1}} \right) mq}{qL - (m-q)\lambda} G_t - \frac{\lambda m}{2} \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) \quad (61)$$

$$\leq \sum_{k=1}^m f_k(w_k^0; \tilde{w}^0) - f_k^*. \quad (62)$$

This is equivalent to

$$\sum_{t=0}^{T-1} G_t^2 - 2\lambda\sqrt{m} \left(\sqrt{d}\sigma + \sqrt{\frac{2}{\lambda} B_{t+1}} \right) \frac{m}{q} G_t - \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \lambda m \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) \quad (63)$$

$$\leq 2 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m f_k(w_k^0; \tilde{w}^0) - f_k^*. \quad (64)$$

Hence, we have

$$\sum_{t=0}^{T-1} \left(G_t - \lambda m \left(\sqrt{d}\sigma + \sqrt{\frac{2}{\lambda} B_{t+1}} \right) \frac{m}{q} \right)^2 \quad (65)$$

$$\leq 2 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m f_k(w_k^0; \tilde{w}^0) - f_k^* + \sum_{t=0}^{T-1} (L\lambda m + m\lambda^2) \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) + 2 \frac{m^3 \lambda B_{t+1}}{q^2}. \quad (66)$$

This implies

$$\sum_{t=0}^{T-1} G_t^2 \quad (67)$$

$$\leq 2 \left(2 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m f_k(w_k^0; \tilde{w}^0) - f_k^* + \sum_{t=0}^{T-1} (L\lambda m + m\lambda^2 + \frac{2\lambda^2 m^3}{q^2}) \left(d\sigma^2 + 2\sigma\sqrt{\frac{2d}{\lambda} B_{t+1}} \right) + 4 \frac{m^3 \lambda B_{t+1}}{q^2} \right). \quad (68)$$

Hence, we conclude that

$$\frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 \quad (69)$$

$$\leq \frac{4 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{T} + \frac{\mathcal{O} \left(L\lambda m + \lambda^2 m + \frac{\lambda^2 m^3}{q^2} \right) \sum_{t=0}^{T-1} \left(d\sigma^2 + 2\sigma \sqrt{\frac{2d}{\lambda} B_{t+1}} + \frac{2B_{t+1}}{\lambda} \right)}{T} \quad (70)$$

$$\leq \frac{4 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{T} + \frac{\mathcal{O} \left(Lm + \lambda m + \frac{\lambda m^3}{q^2} \right) \sum_{t=0}^{T-1} \left(\sqrt{d\lambda} \sigma + \sqrt{2B_{t+1}} \right)^2}{T} \quad (71)$$

$$\leq \frac{4 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{T} + \frac{\mathcal{O} \left(L + \lambda + \frac{\lambda m^2}{q^2} \right) m \sum_{t=0}^{T-1} B_{t+1}}{T} + \mathcal{O} \left(Ld\lambda + d\lambda^2 + \frac{d\lambda^2 m^2}{q^2} \right) m\sigma^2. \quad (72)$$

Taking $\sigma = \frac{c_2 \gamma \sqrt{T \log(1/\delta)}}{m\epsilon}$ and $T = \mathcal{O} \left(\frac{m}{\lambda d \gamma^2} \right)$, we have

$$\frac{1}{mT} \sum_{t=0}^{T-1} \sum_{k=1}^m \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 \quad (73)$$

$$\leq \frac{4 \left(L + \lambda - \frac{m^2}{q^2} \lambda \right) \sum_{k=1}^m (f_k(w_k^0; \tilde{w}^0) - f_k^*)}{mT} + \frac{\mathcal{O} \left(L + \lambda + \frac{m^2}{q^2} \lambda \right) \sum_{t=0}^{T-1} B_{t+1}}{T} + \frac{1}{m} \mathcal{O} \left(L + \lambda + \frac{m^2}{q^2} \lambda \right) \frac{\log(1/\delta)}{\epsilon^2}. \quad (74)$$

□

A.3 CONVERGENCE ANALYSIS (CONVEX):

We first present the formal statement of Theorem 5.

Theorem 8. Let f_k be $(L + \lambda)$ -smooth and $(\mu + \lambda)$ -strongly convex. Assume γ is sufficiently large such that $\gamma \geq \max_{k,t} \|\nabla_{w_k^t} f_k(w_k^t; \tilde{w}^t)\|_2$. Further let $w_k^* = \arg \min_w f_k(w; \bar{w}^*)$, where $\bar{w}^* = \frac{1}{m} \sum_{k=1}^m w_k^*$ and $p = \frac{q}{m}$. If we use a fixed learning rate $\eta_t = \eta = \frac{\frac{c}{L-2} + \lambda p}{p}$ for some constant c such that $0 \leq \eta p(c-2)(\mu + \lambda) \leq 1$, Algorithm 1 satisfies:

$$\begin{aligned} \Delta_T &\leq (1 - \eta p(c-2)(\mu + \lambda))^T \left(\Delta_0 - \frac{m\lambda \left(d\sigma^2 + 2\sqrt{d}\sigma \sqrt{\frac{2}{\lambda} B} + \frac{1}{\lambda} B \right)}{\eta p(c-2)(\mu + \lambda)} \right) \\ &\quad + \frac{m\lambda \left(d\sigma^2 + 2\sqrt{d}\sigma \sqrt{\frac{2}{\lambda} B} + \frac{1}{\lambda} B \right)}{\eta p(c-2)(\mu + \lambda)}, \end{aligned} \quad (75)$$

where $\Delta_t = \sum_{k=1}^m f_k(w_k^t; \tilde{w}^t) - f_k(w_k^*; \tilde{w}^*)$ and $B = \max_t \max_k f_k(w_k^t; \tilde{w}^t)$.

Let σ be chosen as in Theorem 2, then there exists $T = \mathcal{O} \left(\frac{m(c^2-2c)(\mu+\lambda)}{\lambda \left(\frac{L-2}{p^2} + \lambda \right) d\gamma^2} \right)$ such that

$$\begin{aligned} \Delta_T &\leq (1 - \eta p(c-2)(\mu + \lambda))^T \left(\Delta_0 - \frac{\log(1/\delta)}{\epsilon^2} + \mathcal{O} \left(\frac{mB}{\eta p(c-2)(\mu + \lambda)} \right) \right) + \frac{\log(1/\delta)}{\epsilon^2} \\ &\quad + \mathcal{O} \left(\frac{mB}{\eta p(c-2)(\mu + \lambda)} \right). \end{aligned} \quad (76)$$

Proof for Theorem 8. Let $w_k^* = \arg \min_w f_k(w; \bar{w}^*)$. Let I_k^t be the random variable indicating whether task k is selected in communication round t . Thus $\mathbb{E}[I_k^t] = p_k$. By $L + \lambda$ -smoothness and

$\mu + \lambda$ -strong convexity of f_k , we have

$$\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^t) - f_k(w_k^t; \tilde{w}^t)] \leq \mathbb{E} \left[\langle \nabla f_k(w_k^t; \tilde{w}^t), w_k^{t+1} - w_k^t \rangle + \frac{L}{2} \|w_k^{t+1} - w_k^t\|^2 \right] \quad (77)$$

$$= \mathbb{E} \left[\langle \nabla f_k(w_k^t; \tilde{w}^t), \eta_t I_k^t \nabla f_k(w_k^t; \tilde{w}^t) \rangle + \frac{L}{2} \|\eta_t I_k^t \nabla f_k(w_k^t; \tilde{w}^t)\|^2 \right] \quad (78)$$

$$= \left(\frac{L + \lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k \right) \|\nabla f_k(w_k^t; \tilde{w}^t)\|^2 \quad (79)$$

$$\leq \left(\frac{L + \lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k \right) 2(\mu + \lambda)(f(w_k^t; \tilde{w}^t) - f(w_k^*; \tilde{w}^t)) \quad (80)$$

$$\leq \left(\frac{L + \lambda}{2} \eta_t^2 p_k^2 - \eta_t p_k \right) 2(\mu + \lambda)(f(w_k^t; \tilde{w}^t) - f(w_k^*; \tilde{w}^*)). \quad (81)$$

Hence, we have

$$\begin{aligned} \mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^t; \tilde{w}^t)] &\leq \underbrace{\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^{t+1}; \tilde{w}^t)]}_{\mathbf{B}} \\ &\quad + ((L + \lambda)\eta_t^2 p_k^2 - 2\eta_t p_k) (\mu + \lambda)(f(w_k^t; \tilde{w}^t) - f_k^*). \end{aligned} \quad (82)$$

It suffices to bound \mathbf{B} :

$$\mathbf{B} = \mathbb{E} \left[\frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^{t+1}\|^2 - \frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^t\|^2 \right] \quad (83)$$

$$= \frac{\lambda}{2} \mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\| \|2w_k^{t+1} - \tilde{w}^t - \tilde{w}^{t+1}\|] \quad (84)$$

$$\leq \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2] \mathbb{E}[\|2w_k^{t+1} - \tilde{w}^t - \tilde{w}^{t+1}\|^2]} \quad (85)$$

$$= \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2] \mathbb{E}[\|(\tilde{w}^{t+1} - \tilde{w}^t) + 2(w_k^{t+1} - \tilde{w}^{t+1})\|^2]} \quad (86)$$

$$\leq \frac{\lambda}{2} \sqrt{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2] \mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] + 4\|w_k^{t+1} - \tilde{w}^{t+1}\|^2 + 4\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\| \|w_k^{t+1} - \tilde{w}^{t+1}\|]} \quad (87)$$

$$\leq \frac{\lambda}{2} \sqrt{\underbrace{\mathbb{E}[\|\tilde{w}^t - \tilde{w}^{t+1}\|^2]}_{C_1} \left(\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] + 4\|w_k^{t+1} - \tilde{w}^{t+1}\|^2 + 4\sqrt{\mathbb{E}[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] \|w_k^{t+1} - \tilde{w}^{t+1}\|^2} \right)}_{C_2}} \quad (88)$$

where the first and third inequality follows from Cauchy-Schwartz Inequality: $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. It suffices to find the upper bound of C_1 and C_2 .

$$C_1 = \mathbb{E} \left[\left\| \frac{1}{q} \sum_{k \in S_t} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) + \beta^t \right\|^2 \right] \quad (89)$$

$$\leq \left(\sqrt{\mathbb{E} \left[\left\| \frac{1}{q} \sum_{k=1}^m I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{\mathbb{E}[\|\beta^t\|^2]} \right)^2 \quad (90)$$

$$= \left(\sqrt{\mathbb{E} \left[\left\| \frac{1}{q} \sum_{k=1}^m I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{d\sigma} \right)^2 \quad (91)$$

$$\leq \left(\sqrt{\frac{m}{q^2} \sum_{k=1}^m \mathbb{E} \left[\left\| I_k^{t+1} g_k^{t+1} \min \left(1, \frac{\gamma}{\|g_k^{t+1}\|_2} \right) \right\|^2 \right]} + \sqrt{d\sigma} \right)^2 \quad (92)$$

$$\leq \left(\sqrt{\frac{m}{q^2} \sum_{k=1}^m \mathbb{E} \left[\left\| I_k^{t+1} \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2 \right]} + \sqrt{d\sigma} \right)^2 \quad (93)$$

$$\leq \left(\sqrt{\frac{1}{m} \sum_{k=1}^m \left\| \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2} + \sqrt{d\sigma} \right)^2. \quad (94)$$

Denote $h(t) = \sqrt{\frac{1}{m} \sum_{k=1}^m \left\| \eta_t \nabla f_k(w_k^t) \min \left(1, \frac{\gamma}{\eta_t \|\nabla f_k(w_k^t)\|_2} \right) \right\|^2}$. On the other hand,

$$C_2 \leq \frac{2}{\lambda} \frac{\lambda}{2} \|w_k^{t+1} - \tilde{w}^{t+1}\|^2 \quad (95)$$

$$\leq \frac{2}{\lambda} f_k(w_k^{t+1}; \tilde{w}^{t+1}) \quad (96)$$

$$= \frac{2}{\lambda} B_{t+1}. \quad (97)$$

Plug the bounds for C_1 and C_2 into B:

$$B \leq \frac{\lambda}{2} (h(t) + \sqrt{d\sigma}) \left(h(t) + \sqrt{d\sigma} + 2\sqrt{\frac{2}{\lambda} B_{t+1}} \right) \quad (98)$$

$$\leq \lambda \left(h^2(t) + d\sigma^2 + 2\sqrt{d\sigma} \sqrt{\frac{2}{\lambda} B_{t+1}} + \frac{1}{\lambda} B_{t+1} \right) \quad (99)$$

Denoting the right hand side as $\beta(t)$, we have

$$\mathbb{E}[f_k(w_k^{t+1}; \tilde{w}^{t+1}) - f_k(w_k^t; \tilde{w}^t)] \leq \beta(t) + ((L + \lambda)\eta_t^2 p_k^2 - 2\eta_t p_k) (\mu + \lambda)(f(w_k^t; \tilde{w}^t) - f_k^*). \quad (100)$$

Letting $\delta_k^t = \mathbb{E}[f_k(w_k^t; \tilde{w}^t) - f_k(w_k^*; \tilde{w}^*)]$, we have

$$\delta_k^{t+1} \leq (1 - ((L + \lambda)\eta_t^2 p_k^2 - 2\eta_t p_k) (\mu + \lambda)) \delta_k^t + \beta(t). \quad (101)$$

Table 2

| Dataset | Number of tasks | Model | Task Type |
|---------------------------------------------------|-----------------|---------------------|-------------------------------|
| FEMNIST (Cohen et al., 2017; Caldas et al., 2018) | 205 | 4-layer CNN | 62-class image classification |
| StackOverflow (tff) | 400 | Logistic Regression | 500-class tag prediction |
| CelebA (Liu et al., 2015; Caldas et al., 2018) | 515 | 4-layer CNN | Binary image classification |

Summing over k on the left handed side, when γ is large enough so that no clipping happens we have

$$\sum_{k=1}^m \delta_k^{t+1} \quad (102)$$

$$\leq (1 - ((L + \lambda)\eta_t^2 p^2 - 2\eta_t p)(\mu + \lambda)) \sum_{k=1}^m \delta_k^t + m\beta(t) \quad (103)$$

$$= (1 - ((L + \lambda p^2 - 2)\eta_t^2 - 2\eta_t p)(\mu + \lambda)) \sum_{k=1}^m \delta_k^t + m\lambda \left(d\sigma^2 + 2\sqrt{d}\sigma \sqrt{\frac{2}{\lambda} B_{t+1}} + \frac{1}{\lambda} B_{t+1} \right). \quad (104)$$

Let $\Delta_t = \sum_{k=1}^m \delta_k^t$. Assume $\max_{t \leq T} B_t = B$. Pick $C = \frac{m\lambda(d\sigma^2 + 2\sqrt{d}\sigma\sqrt{\frac{2}{\lambda}B} + \frac{1}{\lambda}B)}{((L + \lambda p^2 - 2)\eta_t^2 - 2\eta_t p)(\mu + \lambda)}$, we have

$$\Delta_{t+1} - C \leq (1 - ((L + \lambda p^2 - 2)\eta_t^2 - 2\eta_t p)(\mu + \lambda)) (\Delta_t - C). \quad (105)$$

Choose $\eta_t = \eta = \frac{c}{\frac{L-2}{p} + \lambda p}$ for some constant c such that $0 < (1 - ((L + \lambda p^2 - 2)\eta_t^2 - 2\eta_t p)(\mu + \lambda)) < 1$. Apply recursively to all t , we obtain

$$\Delta_T \leq \left(1 - \frac{(c^2 - 2c)(\mu + \lambda)}{\frac{L-2}{p^2} + \lambda}\right)^T \left(\Delta_0 - \frac{m\lambda(d\sigma^2 + 2\sqrt{d}\sigma\sqrt{\frac{2}{\lambda}B} + \frac{1}{\lambda}B)}{\frac{(c^2 - 2c)(\mu + \lambda)}{\frac{L-2}{p^2} + \lambda}}\right) + \frac{m\lambda(d\sigma^2 + 2\sqrt{d}\sigma\sqrt{\frac{2}{\lambda}B} + \frac{1}{\lambda}B)}{\frac{(c^2 - 2c)(\mu + \lambda)}{\frac{L-2}{p^2} + \lambda}}. \quad (106)$$

Take $\sigma = \frac{c_2 \gamma \sqrt{T \log(1/\delta)}}{m\epsilon}$ and we can find $T = \mathcal{O}\left(\frac{m(c^2 - 2c)(\mu + \lambda)}{\lambda(\frac{L-2}{p^2} + \lambda)d\gamma^2}\right)$ such that,

$$\Delta_T \leq \left(1 - \frac{(c^2 - 2c)(\mu + \lambda)}{\frac{L-2}{p^2} + \lambda}\right)^T \left(\Delta_0 - \frac{\log(1/\delta)}{\epsilon^2} - \mathcal{O}\left(\frac{mB(\frac{L-2}{p^2} + \lambda)}{(c^2 - 2c)(\mu + \lambda)}\right)\right) + \frac{\log(1/\delta)}{\epsilon^2} + \mathcal{O}\left(\frac{mB(\frac{L-2}{p^2} + \lambda)}{(c^2 - 2c)(\mu + \lambda)}\right). \quad (107)$$

Divide both side by m , we have

$$\frac{1}{m} \Delta_T \leq \left(1 - \frac{(c^2 - 2c)(\mu + \lambda)}{\frac{L-2}{p^2} + \lambda}\right)^T \left(\frac{1}{m} \Delta_0 - \frac{\log(1/\delta)}{m\epsilon^2} - \mathcal{O}\left(\frac{B(\frac{L-2}{p^2} + \lambda)}{(c^2 - 2c)(\mu + \lambda)}\right)\right) \quad (108)$$

$$+ \frac{\log(1/\delta)}{m\epsilon^2} + \mathcal{O}\left(\frac{B(\frac{L-2}{p^2} + \lambda)}{(c^2 - 2c)(\mu + \lambda)}\right). \quad (109)$$

□

A.4 DATASETS AND MODELS

We summarize the details of the datasets and models we used in our empirical study in Table 2. Our experiments include both convex (Logistic Regression) and non-convex (CNN) loss objectives on both text (StackOverflow) and image (CelebA and FEMNIST) datasets.

A.5 HYPERPARAMETERS

Each fixed privacy parameter ϵ could be computed by different combinations of noise scale σ , clipping norm γ , number of communication rounds T , and subsampling rate $p = \frac{q}{m}$. In all our experiments, we subsample 100 different tasks for each round, i.e. $q = 100$, to perform local training as well as involved in global aggregation. For FEMNIST and CelebA, we choose $\sigma \in \{0.02, 0.05, 0.1\}$ and $\gamma \in \{0.2, 0.5, 1\}$. For StackOverflow, we choose $\sigma \in \{0.01, 0.05, 0.1\}$ and $\gamma \in \{0.1, 0.5, 1\}$. We summarize both utility and privacy performance for different hyperparameters below.

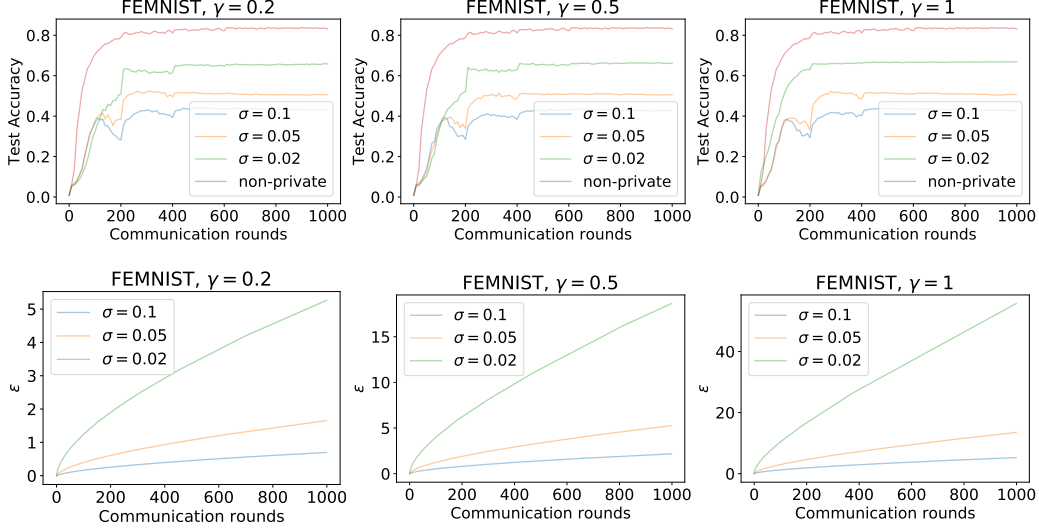


Figure 4: FEMNIST results

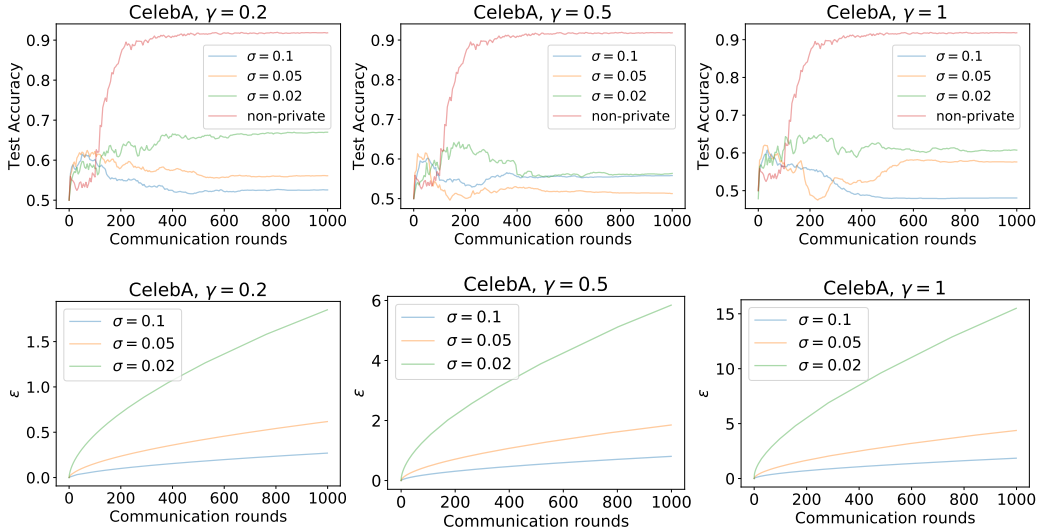


Figure 5: CelebA results

A.6 COMPARISON WITH FEDPROX

Besides FedAvg, we also compared private mean-regularized MTL with other methods that aims to train a global model privately. In particular, we studied private FedProx (Li et al., 2020b) as an alternative global baseline. Note that although the local objective being solved in FedProx is similar to that in mean-regularized MTL, FedProx is a fundamentally different method to handle data heterogeneity in FL from MTL. Specifically, FedProx learns a *global model* where each client

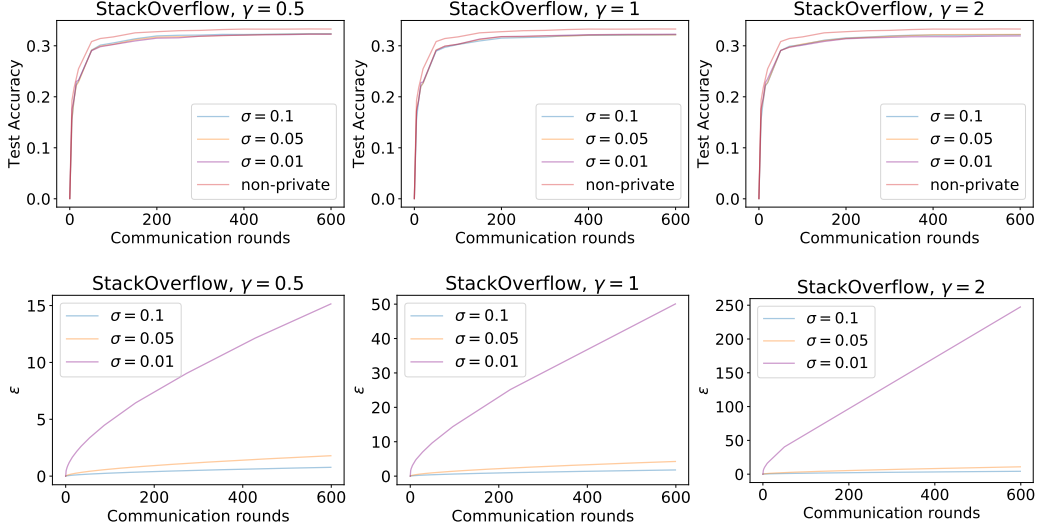


Figure 6: StackOverflow results

solves an inexact minimizer by optimizing local empirical risk with a regularization term. We instead explore learning a *multi-task objective* where each client solves a mean-regularized objective and learns a *separate, client-specific model*. The results are shown in Figure 7. In all three datasets, private FedProx is very similar to private FedAvg under different private parameters ϵ and worse than private MTL. In particular, in FEMNIST and Stackoverflow, private MTL significantly outperforms training a private global model (FedAvg and FedProx), for all ϵ 's.

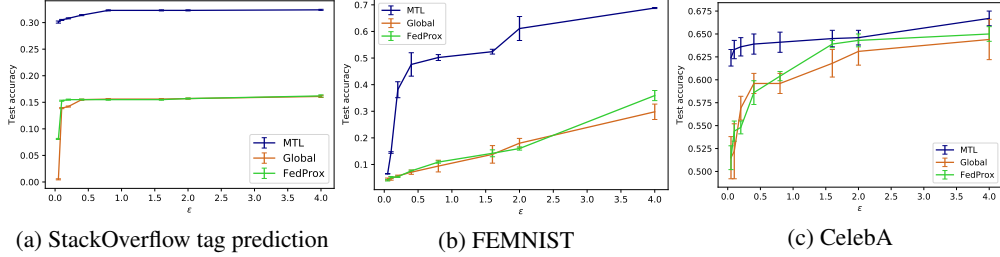


Figure 7: Comparison of PMTL and training a private global model (FedAvg/FedProx).