A Primal-Dual Algorithm for Hybrid Federated Learning: Supplementary Materials

1 Convergence Proofs

Lemma 1. *Let $l_i : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz continuous function. Then for any real value α with $|\alpha| > L$ we have that $l_i^*(\alpha) = \infty$*

Proof. Proof provided in Lemma 21 in (Shalev-Shwartz and Zhang, 2013). □

1.1 Proof of Theorem 1

Proof. For simplicity, the global dual and primal variables are denoted as: $w = w_0$ and $\alpha_i = \alpha_{0,i}$. We also frequently use the following relationships.

1. The map from dual to primal is $w = \frac{1}{\lambda N}\sum_{i=1}^N \alpha_i x_i = \mathbf{A}\alpha$.
2. $\alpha_i^t = \alpha_i^{t-1} + \gamma_t \Delta\alpha_i^t$
3. Equalities $\alpha_i^{t-1} = \alpha_{b,i}^{t-1} = \frac{1}{Q}\sum_{b\in\mathcal{B}_i} \alpha_{b,i}^{t-1}$ hold because the aggregated dual variables on the server are sent back to the clients every iteration.
4. Following from Cauchy-Schwarz, we have $||\sum_{i=1}^N z_i||^2 \leq N\sum_{i=1}^N ||z_i||^2$.

Starting with the difference in dual objective after one outer iteration, we have

$$N[D(\alpha^t) - D(\alpha^{t-1})] = \underbrace{[-\sum_{i=1}^N l_i^*(-\alpha_i^t) - \frac{\lambda N}{2}||\mathbf{A}\alpha^t||^2]}_{A} - \underbrace{[-\sum_{i=1}^N l_i^*(-\alpha_i^{t-1}) - \frac{\lambda N}{2}||\mathbf{A}\alpha^{t-1}||^2]}_{B}.$$

We rewrite A as

$$A = \underbrace{-\sum_{i=1}^N l_i^*(-\alpha_i^t)}_{A_1} \underbrace{-\frac{\lambda N}{2}||\mathbf{A}\alpha^t||^2}_{A_2}$$

and then bound A_1 as

$$\begin{aligned} A_1 &= -\sum_{i=1}^N l_i^*(-\alpha_i^{t-1} - \gamma_t\Delta\alpha_i^t) \\ &= -\sum_{i=1}^N l_i^*(-(1-\gamma_t)\alpha_i^{t-1} - \gamma_t(\alpha_i^{t-1} + \Delta\alpha_i^t)) \\ &\geq -\sum_{i=1}^N [\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) + (1-\gamma_t) l_i^*(-\alpha_i^{t-1})] \end{aligned}$$

and rewrite A_2 as

$$\begin{aligned}
A_2 &= -\frac{\lambda N}{2}||\mathbf{A}\alpha^t||^2 \\
&= -\frac{\lambda N}{2}||\mathbf{A}[\alpha^{t-1} + \gamma_t \Delta\alpha^t]||^2 \\
&= -\frac{\lambda N}{2}||w^{t-1} + \gamma_t \mathbf{A}\Delta\alpha^t||^2 \\
&= -\frac{\lambda N}{2}[||w^{t-1}||^2 + 2\gamma_t (w^{t-1})^T \mathbf{A}\Delta\alpha^t + \gamma_t^2 ||\mathbf{A}\Delta\alpha^t||^2].
\end{aligned}$$

Our bound on A is

$$A \geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) + (1-\gamma_t) l_i^*(-\alpha_i^{t-1})] - \frac{\lambda N}{2}[||w^{t-1}||^2 + 2\gamma_t (w^{t-1})^T \mathbf{A}\Delta\alpha^t + \gamma_t^2 ||\mathbf{A}\Delta\alpha^t||^2].$$

Expression B appears in the RHS, so we can simplify

$$A - B \geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1})] - \frac{\lambda N}{2}[2\gamma_t (w^{t-1})^T \mathbf{A}\Delta\alpha^t + \gamma_t^2 ||\mathbf{A}\Delta\alpha^t||^2].$$

Next, we re-write the linear operator $\mathbf{A}$ as a summation over samples

$$\begin{aligned}
A - B &\geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1})] - \frac{\lambda N}{2}[2\gamma_t (w^{t-1})^T \sum_{i=1}^{N}\frac{1}{\lambda N}x_i \Delta\alpha_i^t + \gamma_t^2 ||\sum_{i=1}^{N}\frac{1}{\lambda N}x_i\Delta\alpha_i^t||^2] \\
&= -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i \Delta\alpha_i^t] - \frac{\lambda N}{2}[\gamma_t^2 ||\sum_{i=1}^{N}\frac{1}{\lambda N}x_i\Delta\alpha_i^t||^2] \\
&\geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i \Delta\alpha_i^t] - \frac{\lambda N}{2}[\frac{\gamma_t^2}{\lambda^2 N}\sum_{i=1}^{N}||x_i\Delta\alpha_i^t||^2].
\end{aligned}$$

Without loss of generality, we assume our data is normalized such that $||x_i|| \leq 1$, which yields

$$A - B \geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i \Delta\alpha_i^t + \frac{\gamma_t^2}{2\lambda}(\Delta\alpha_i^t)^2].$$

Our dual updates are defined as $\Delta\alpha_i^t = \frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t$, which is just the mean of the dual variable updates on each client that contains sample i. In turn we have

$$\begin{aligned}
A - B &\geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i (\frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t) + \frac{\gamma_t^2}{2\lambda}(\frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t)^2] \\
&= -\sum_{i=1}^{N}[\gamma_t l_i^*(-\frac{1}{Q}\sum_{b\in\mathcal{B}_i}(\alpha_i^{t-1} + \Delta\alpha_{b,i}^t)) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i (\frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t) + \frac{\gamma_t^2}{2\lambda}(\frac{1}{Q}\sum_{b\in\mathcal{B}_i}\Delta\alpha_{b,i}^t)^2].
\end{aligned}$$

By convexity of l_i^* we get

$$\begin{aligned}
A - B &\geq -\sum_{i=1}^{N}[\frac{\gamma_t}{Q}\sum_{b\in\mathcal{B}_i} l_i^*(-(\alpha_i^{t-1} + \Delta\alpha_{b,i}^t)) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \frac{\gamma_t}{Q}\sum_{b\in\mathcal{B}_i}(w^{t-1})^T x_i \Delta\alpha_{b,i}^t + \frac{\gamma_t^2}{2\lambda Q}\sum_{b\in\mathcal{B}_i}(\Delta\alpha_{b,i}^t)^2] \\
&= -\frac{\gamma_t}{Q}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_i}[l_i^*(-(\alpha_i^{t-1} + \Delta\alpha_{b,i}^t)) - l_i^*(-\alpha_i^{t-1}) + (w^{t-1})^T x_i \Delta\alpha_{b,i}^t + \frac{\gamma_t}{2\lambda}(\Delta\alpha_{b,i}^t)^2].
\end{aligned}$$

The stochastic element of this algorithm is in randomly determining which $\Delta\alpha_{b,i}$ are nonzero. If a particular $\Delta\alpha_{b,i}$ is zero, then that term in the double series in the right hand side of the inequality is zero. If that $\Delta\alpha_{b,i}$ is nonzero, then the local dual method chooses the $\Delta\alpha_{b,i}$ as follows. We choose our local dual method such that $\Delta\alpha^t_{b,i} = s_{b,i}(u_i^{t-1} - \alpha_{b,i}^{t-1})$ where $s_{b,i} \in [0,1]$ and $u_i^{t-1} \in \partial l_i(x_i^T w_0^{t-1})$. We define our local dual method to find $s_{b,i}$ as follows

$$s_{b,i} = \underset{s\in[0,1]}{\arg\max}\{-l_i^*(-(\alpha_{b,i}^{t-1} + s(u_i^{t-1} - \alpha_{b,i}^{t-1}))) - s(w^{t-1})^T x_i(u_i^{t-1} - \alpha_{b,i}^{t-1}) - \frac{\gamma_t}{2\lambda}(s(u_i^{t-1} - \alpha_{b,i}^{t-1}))^2\}.$$

Now we must find the probability that a particular $\alpha_{b,i}$ is updated. This is HK/N, because there are a total of QN pairings of samples and clients and there are HQK total updates per outer iteration. So we take the expectation over these client/sample pairs while conditioning on the previous state, α_i^{t-1}, to obtain

$$\begin{aligned}\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{QN}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_i}[l_i^*(-(\alpha_i^{t-1} + s_{b,i}(u_i^{t-1} - \alpha_{b,i}^{t-1}))) - l_i^*(-\alpha_i^{t-1})\\ + (w^{t-1})^T x_i s_{b,i}(u_i^{t-1} - \alpha_{b,i}^{t-1}) + \frac{\gamma_t}{2\lambda}(s_{b,i}(u_i^{t-1} - \alpha_{b,i}^{t-1}))^2].\end{aligned}$$

Due to the definition of the LocalDualMethod, the inequality holds for any choice of $s_t \in [0,1]$, and thus

$$\begin{aligned}\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{QN}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_i}[l_i^*(-(\alpha_i^{t-1} + s_t(u_i^{t-1} - \alpha_{b,i}^{t-1}))) - l_i^*(-\alpha_i^{t-1})\\ + (w^{t-1})^T x_i s_t(u_i^{t-1} - \alpha_{b,i}^{t-1}) + \frac{\gamma_t}{2\lambda}(s_t(u_i^{t-1} - \alpha_{b,i}^{t-1}))^2].\end{aligned}$$

By convexity of l_i^*, we conclude

$$\begin{aligned}\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{QN}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_i}[(1-s_t)l_i^*(-\alpha_{b,i}^{t-1}) + s_t l_i^*(-u_i^{t-1}) - l_i^*(-\alpha_i^{t-1})\\ + (w^{t-1})^T x_i s_t(u_i^{t-1} - \alpha_{b,i}^{t-1}) + \frac{\gamma_t}{2\lambda}(s_t(u_i^{t-1} - \alpha_{b,i}^{t-1}))^2]\end{aligned}$$

and thus from $\alpha_{b,i}^{t-1} = \alpha_i^{t-1}$, we have that

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{QN}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_i}[-s_t l_i^*(-\alpha_i^{t-1}) + s_t l_i^*(-u_i^{t-1}) + s_t(w^{t-1})^T x_i(u_i^{t-1} - \alpha_i^{t-1}) + \frac{\gamma_t s_t^2}{2\lambda}(u_i^{t-1} - \alpha_i^{t-1})^2].$$

We obtain

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{N}\sum_{i=1}^{N}[-s_t l_i^*(-\alpha_i^{t-1}) + s_t l_i^*(-u_i^{t-1}) + s_t(w^{t-1})^T x_i(u_i^{t-1} - \alpha_i^{t-1}) + \frac{\gamma_t s_t^2}{2\lambda}(u_i^{t-1} - \alpha_i^{t-1})^2].$$

Now from convex conjugates we know that $l_i(x_i^T w^{t-1}) = -l_i^*(-u_i^{t-1}) - u_i^{t-1}x_i^T w^{t-1}$, and thus

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t HK}{N}\sum_{i=1}^{N}[-s_t l_i^*(-\alpha_i^{t-1}) - s_t l_i(x_i^T w^{t-1}) - s_t(w^{t-1})^T x_i\alpha_i^{t-1} + \frac{\gamma_t s_t^2}{2\lambda}(u_i^{t-1} - \alpha_i^{t-1})^2]. \tag{1}$$

The primal-dual gap at iteration $t-1$ is defined as

$$\begin{aligned}P(w^{t-1}) - D(\alpha^{t-1}) &= (\frac{\lambda}{2}||w^{t-1}||^2 + \frac{1}{N}\sum_{i=1}^{N} l_i((w^{t-1})^T x_i)) - (-\frac{\lambda}{2}||\frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{t-1}x_i||^2 - \frac{1}{N}\sum_{i=1}^{N} l_i^*(-\alpha_i^{t-1}))\\ &= \frac{1}{N}\sum_{i=1}^{N} l_i^*(-\alpha_i^{t-1}) + \frac{1}{N}\sum_{i=1}^{N} l_i((w^{t-1})^T x_i) + \lambda||w^{t-1}||^2\\ &= \frac{1}{N}\sum_{i=1}^{N} l_i^*(-\alpha_i^{t-1}) + \frac{1}{N}\sum_{i=1}^{N} l_i((w^{t-1})^T x_i) + \lambda(w^{t-1})^T w^{t-1}\\ &= \frac{1}{N}\sum_{i=1}^{N} l_i^*(-\alpha_i^{t-1}) + \frac{1}{N}\sum_{i=1}^{N} l_i((w^{t-1})^T x_i) + \lambda(w^{t-1})^T(\frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{t-1}x_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}[l_i((w^{t-1})^T x_i) + l_i^*(-\alpha_i^{t-1}) + \alpha_i^{t-1}x_i^T w^{t-1}].\end{aligned}$$

We plug this primal-dual gap into (1) to derive

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq s_t\gamma_t HK[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{\gamma_t^2 s_t^2 HK}{2N\lambda}\sum_{i=1}^{N}(u_i^{t-1} - \alpha_i^{t-1})^2$$

and after dividing by N

$$\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})|\alpha^{t-1}] \geq \frac{s_t\gamma_t HK}{N}[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{\gamma_t^2 s_t^2 HK}{2N^2\lambda}\sum_{i=1}^{N}(u_i^{t-1} - \alpha_i^{t-1})^2.$$

By assumption, $\gamma_t = 1$ and setting $G^{t-1} = \frac{1}{2N\lambda}\sum_{i=1}^{N}(u_i^{t-1} - \alpha_i^{t-1})^2$, we get

$$\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})|\alpha^{t-1}] \geq \frac{s_t HK}{N}[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{s_t^2 HK}{N}G^{t-1}. \tag{2}$$

We know that $\varepsilon_D^{t-1} = D(\alpha^*) - D(\alpha^{t-1}) \leq P(w^{t-1}) - D(\alpha^{t-1})$ and $D(\alpha^t) - D(\alpha^{t-1}) = \varepsilon_D^{t-1} - \varepsilon_D^t$. It must also hold that $\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})|\alpha^{t-1}] = \mathbb{E}[\varepsilon_D^{t-1} - \varepsilon_D^t|\alpha^{t-1}] = \varepsilon_D^{t-1} - \mathbb{E}[\varepsilon_D^t|\alpha^{t-1}]$. Together with (2) this yields

$$\varepsilon_D^{t-1} - \mathbb{E}[\varepsilon_D^t|\alpha^{t-1}] \geq \frac{s_t HK}{N}\varepsilon_D^{t-1} - \frac{s_t^2 HK}{N}G^{t-1}.$$

Let $G^{t-1} \leq G$. Then,

$$\mathbb{E}[\varepsilon_D^t|\alpha^{t-1}] \leq (1 - \frac{s_t HK}{N})\varepsilon_D^{t-1} + \frac{s_t^2 HK}{N}G.$$

Taking the expectation of both sides and using the law of iterated expectation we obtain

$$\mathbb{E}[\varepsilon_D^t] \leq (1 - \frac{s_t HK}{N})\mathbb{E}[\varepsilon_D^{t-1}] + \frac{s_t^2 HK}{N}G.$$

Finally we need to bound G. From Lemma 1 we know that each $|\alpha_{b,i}| < L$ due to the choice of LocalDualMethod. Furthermore, $|u_i| < L$ because l_i are L-Lipschitz. We conclude that

$$G = \frac{2L^2}{\lambda}.$$

This concludes the proof.

□

1.2 Proof of Theorem 2

Proof. This is proved by using induction on t. It also uses the fact that $1 + x \leq e^x$. Let $s_t = 1$ for $t \leq t_0$ and $s_t = \frac{1}{1+\frac{HK}{2N}(t-t_0)}$ for $t > t_0$.

The base case of $t = t_0$ follows as

$$\begin{aligned}
\mathbb{E}[\varepsilon_D^t] &\leq (1 - \frac{HK}{N})^t \mathbb{E}[\varepsilon_D^0] + \frac{HK}{N} G \sum_{\tau=1}^{t-1} (1 - \frac{HK}{N})^\tau \\
&\leq (\exp(\frac{-HK}{N}))^t \mathbb{E}[\varepsilon_D^0] + \frac{HK}{N} G (\frac{1 - (1 - \frac{HK}{N})^{t-1}}{1 - (1 - \frac{HK}{N})}) \\
&\leq (\exp(\frac{-HK}{N} \log(\mathbb{E}[\varepsilon_D^0]/G))) \mathbb{E}[\varepsilon_D^0] + G \underbrace{(1 - (1 - \frac{HK}{N})^{t-1})}_{\leq 1} \\
&\leq (\frac{G}{\varepsilon_D^0})^{\frac{HK}{N}} \mathbb{E}[\varepsilon_D^0] + G \\
&\leq G + G \\
&= 2G.
\end{aligned}$$

We now assume (3) holds for some t, and we need to prove that it also holds for $t + 1$. We start with

$$\begin{aligned}
\mathbb{E}[\varepsilon_D^{t+1}] &\leq (1 - \frac{sHK}{N}) \mathbb{E}[\varepsilon_D^t] + \frac{s^2 HK}{N} G \\
&\leq (1 - \frac{sHK}{N})(\frac{2G}{1 + \frac{HK}{2N}(t - t_0)}) + \frac{s^2 HK}{N} G \\
&= (1 - \frac{HK/N}{1 + \frac{HK}{2N}(t - t_0)})(\frac{2G}{1 + \frac{HK}{2N}(t - t_0)}) + \frac{HKG/N}{(1 + \frac{HK}{2N}(t - t_0))^2} \\
&= 2G \underbrace{\frac{1 + \frac{HK}{2N}(t - t_0) - \frac{HK}{2N}}{(1 + \frac{HK}{2N}(t - t_0))^2}}_{U}.
\end{aligned}$$

We bound U by

$$\begin{aligned}
U &= \frac{1 + \frac{HK}{2N}(t - t_0) - \frac{HK}{2N}}{(1 + \frac{HK}{2N}(t - t_0))^2} \\
&= \frac{1}{1 + \frac{HK}{2N}(t + 1 - t_0)} \frac{(1 + \frac{HK}{2N}(t + 1 - t_0))(1 + \frac{HK}{2N}(t - 1 - t_0))}{(1 + \frac{HK}{2N}(t - t_0))^2} \\
&= \frac{1}{1 + \frac{HK}{2N}(t + 1 - t_0)} \frac{1 + \frac{HK}{N}(t - t_0) + \frac{H^2K^2}{4N^2}((t - t_0)^2 - 1)}{1 + \frac{HK}{N}(t - t_0) + \frac{H^2K^2}{4N^2}(t - t_0)^2} \\
&= \frac{1}{1 + \frac{HK}{2N}(t + 1 - t_0)} (1 - \frac{H^2K^2}{4N^2(1 + \frac{HK}{N}(t - t_0) + \frac{H^2K^2}{4N^2}(t - t_0)^2)}) \\
&= \frac{1}{1 + \frac{HK}{2N}(t + 1 - t_0)} \underbrace{(1 - \frac{H^2K^2}{4N^2(1 + \frac{HK}{2N}(t - t_0))^2})}_{\leq 1} \\
&\leq \frac{1}{1 + \frac{HK}{2N}(t + 1 - t_0)}.
\end{aligned}$$

Therefore,

$$\mathbb{E}[\varepsilon_D^{t+1}] \leq \frac{2G}{1 + \frac{HK}{2N}(t + 1 - t_0)} \tag{3}$$

completing the proof. □

1.3 Proof of Theorem 3

Proof. This proof follows the proof of Theorem 1 closely. We first note some important facts that are key to the proof for the horizontal case.

1. We have $Q = |\mathcal{B}_i| = 1$, therefore each dual variable α_i only belongs to a single client. We simplify the notation by omitting the dual variable subscript pertaining to the client, e.g. we write α_i^t instead of $\alpha_{b,i}^t$.
2. The inner product for sample i, $x_i^T w^t$ can be found completely on the single client and no aggregation by the server is needed.
3. The relationship $w^{t-1} = \frac{1}{\lambda N}\sum_{i=1}^{N} \alpha_i^{t-1} x_i$ still holds. Because some of the α_i have not been updated in iteration $t-1$, the algorithm has stored the contributions of each component in $w_{0,\hat{k},m}$ of the sum so that w^{t-1} can be exactly found.

From the proof of Theorem 1, we have

$$A - B \geq -\sum_{i=1}^{N}[\gamma_t l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t (w^{t-1})^T x_i \Delta\alpha_i^t + \frac{\gamma_t^2}{2\lambda}(\Delta\alpha_i^t)^2].$$

In the hybrid case, we compute $\Delta\alpha_i^t = \frac{1}{Q}\sum_{b\in\mathcal{B}_i} \Delta\alpha_{b,i}^t$. However, in this case, this can be simplified due to $Q = |\mathcal{B}_i| = 1$.

Now we must find the probability that a particular α_i is updated. This is probability PH/N, because there are a total of N samples and there are PH total expected updates per outer iteration. So taking the expectation over these available clients while conditioning on the previous state, α^{t-1}, yields

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{\gamma_t PH}{N}\sum_{i=1}^{N}[l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - l_i^*(-\alpha_i^{t-1}) + (w^{t-1})^T x_i \Delta\alpha_i^t + \frac{\gamma_t}{2\lambda}(\Delta\alpha_i^t)^2].$$

Due to the choice of LocalDualMethod, for any $s_t \in [0, 1]$, we obtain

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{PH\gamma_t}{N}\sum_{i=1}^{N}[-s_t l_i^*(-\alpha_i^{t-1}) + s_t l_i^*(-u_i^{t-1}) + (w^{t-1})^T x_i s_t(u_i^{t-1} - \alpha_i^{t-1}) + \frac{\gamma_t}{2\lambda}s_t^2(u_i^{t-1} - \alpha_i^{t-1})^2].$$

In the same vein as the in the proof of Theorem 1, we derive

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq -\frac{PH\gamma_t}{N}\sum_{i=1}^{N}[-s_t l_i^*(-\alpha_i^{t-1}) - s_t l_i(x_i^T w^{t-1}) - s_t (w^{t-1})^T x_i \alpha_i^{t-1} + \frac{\gamma_t}{2\lambda}s_t^2(u_i^{t-1} - \alpha_i^{t-1})^2]$$

and

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq PH\gamma_t s_t[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{PH\gamma_t^2}{2\lambda N}s_t^2(u_i^{t-1} - \alpha_i^{t-1})^2]$$

which in turn yields

$$\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})|\alpha^{t-1}] \geq \frac{PHs_t}{N}[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{PHs_t^2}{N}G^{t-1}.$$

The rest of the proof is the same as in Theorem 1 and 2. □

1.4 Proof of Theorem 4

Proof. We first note some important points.

1. If set c is selected for updates, then for all $b \in \mathcal{B}_c$, we have $\alpha_{b,i}^{t-1} = \alpha_i^{t-1}$ because all clients in that set are sent the global dual variables from the server before performing local updates.
2. The dot product scalar, $x_i^T w$, that is calculated and passed from the server at the start of iteration t is representative of the true $x_i^T w^{t-1}$. This is because if the feature indices are updated on a particular client, then that component of the dot product is computed and sent to the server and the server stores the components of the dot product from feature indices that were not updated.
3. We frequently use Lemma 1 and the choice of the LocalDualMethod to bound $|\alpha_i| < L$. Furthermore, $|u_i| < L$ because l_i are L-Lipschitz.

We examine the change in the dual objective after one outer iteration. We assume that $t \geq C$. We start with

$$N[D(\alpha^t) - D(\alpha^{t-1})] \geq \underbrace{[-\sum_{i=1}^N l_i^*(-\alpha_i^t) - \frac{\lambda N}{2}||\mathbf{A}\alpha^t||^2]}_{A} - \underbrace{[-\sum_{i=1}^N l_i^*(-\alpha_i^{t-1}) - \frac{\lambda N}{2}||\mathbf{A}\alpha^{t-1}||^2]}_{B}.$$

We first examine A by

$$\begin{aligned}
A &= -\sum_{i=1}^N l_i^*(-\alpha_i^t) - \frac{\lambda N}{2}||\mathbf{A}\alpha^t||^2 \\
&= -\sum_{i=1}^N l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \frac{\lambda N}{2}||\mathbf{A}\alpha^{t-1} + \mathbf{A}\Delta\alpha^t||^2 \\
&= -\sum_{i=1}^N l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - \frac{\lambda N}{2}[||\mathbf{A}\alpha^{t-1}||^2 + 2(\mathbf{A}\alpha^{t-1})^T(\mathbf{A}\Delta\alpha^t) + ||\mathbf{A}\Delta\alpha^t||^2].
\end{aligned}$$

Then we have

$$\begin{aligned}
A - B &\geq -\sum_{i=1}^N [l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - l_i^*(-\alpha_i^{t-1})] - \lambda N(\mathbf{A}\alpha^{t-1})^T(\mathbf{A}\Delta\alpha^t) - \frac{\lambda N}{2}||\mathbf{A}\Delta\alpha^t||^2 \\
&\geq -\sum_{i=1}^N [l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - l_i^*(-\alpha_i^{t-1})] - \sum_{i=1}^N (\mathbf{A}\alpha^{t-1})^T\Delta\alpha_i^t x_i - \frac{1}{2\lambda}\sum_{i=1}^N (\Delta\alpha_i^t)^2 \\
&= -\sum_{i=1}^N [l_i^*(-\alpha_i^{t-1} - \Delta\alpha_i^t) - l_i^*(-\alpha_i^{t-1}) + (\mathbf{A}\alpha^{t-1})^T\Delta\alpha_i^t x_i + \frac{1}{2\lambda}(\Delta\alpha_i^t)^2].
\end{aligned}$$

We assume that at iteration t, a set c is selected and the clients in that set are updated. We have that $\Delta\alpha_i^t = \frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t$ and

$$A - B \geq -\sum_{i=1}^N [l_i^*(-\alpha_i^{t-1} - \frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t) - l_i^*(-\alpha_i^{t-1}) + (\mathbf{A}\alpha^{t-1})^T(\frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t)x_i + \frac{1}{2\lambda}(\frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t)^2].$$

Using convexity we have that

$$\begin{aligned}
l_i^*(-\alpha_i^{t-1} - \frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t) &= l_i^*(-\frac{1}{Q/C}\sum_{b\in\mathcal{B}_c}\alpha_i^{t-1} - \frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^t) \\
&\leq \frac{1}{Q/C}\sum_{b\in\mathcal{B}_c} l_i^*(-\alpha_i^{t-1} - \gamma_t\Delta\alpha_{b,i}^t).
\end{aligned}$$

Returning to $A - B$ we derive

$$\begin{aligned}
A - B &\geq -\sum_{i=1}^{N}[\frac{1}{Q/C}\sum_{b\in\mathcal{B}_c} l_i^*(-\alpha_i^{t-1} - \gamma_t \Delta\alpha_{b,i}^t) - l_i^*(-\alpha_i^{t-1}) + \frac{\gamma_t}{Q/C}\sum_{b\in\mathcal{B}_c}(\mathbf{A}\alpha^{t-1})^T(\Delta\alpha_{b,i}^t)x_i + \frac{1}{2\lambda}\frac{1}{Q/C}\sum_{b\in\mathcal{B}_c}(\gamma_t\Delta\alpha_{b,i}^t)^2] \\
&= -\frac{1}{Q/C}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_c}[l_i^*(-\alpha_i^{t-1} - \gamma_t\Delta\alpha_{b,i}^t) - l_i^*(-\alpha_i^{t-1}) + \gamma_t(\mathbf{A}\alpha^{t-1})^T(\Delta\alpha_{b,i}^t)x_i + \frac{1}{2\lambda}(\gamma_t\Delta\alpha_{b,i}^t)^2].
\end{aligned}$$

The stochastic component of this algorithm is deciding if a particular sample i is selected for the corresponding dual variable to be updated. If the sample-client pair is not selected to be updated, then $\Delta\alpha_{b,i}^t$ is zero, and the entire right hand size of the inequality is zero. There are HQ/C possible updates in a given outer iteration and NQ/C total sample-client combinations. Therefore, the probability of a given sample-client pair to be selected to be updated is H/N. By conditioning on α_i^{t-1}, we get

$$\begin{aligned}
\mathbb{E}[A - B|\alpha^{t-1}] \geq &-\frac{H}{QN/C}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_c}[l_i^*(-\alpha_i^{t-1} - \gamma_t s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1})) - l_i^*(-\alpha_i^{t-1}) + \gamma_t(\mathbf{A}\alpha^{t-1})^T(s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1}))x_i \\
&+ \frac{1}{2\lambda}(\gamma_t s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1}))^2] \\
= &-\frac{H}{QN/C}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_c}[l_i^*(-\alpha_i^{t-1} - \gamma_t s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1})) - l_i^*(-\alpha_i^{t-1}) + \gamma_t s_{b,i}^t(\mathbf{A}\alpha^{t-1})^T(u_i^{t-1} - \alpha_i^{t-1})x_i \\
&+ \frac{1}{2\lambda}(\gamma_t s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1}))^2]. \qquad (4)
\end{aligned}$$

We now express $\mathbf{A}\alpha^{t-1}$ in terms of w^{t-1}. In the vertical case with incomplete client participation, w^{t-1} is different because only a subset of clients are available at each outer iteration. Without loss of generality, let us assume that at iteration $t-1$ set 1 was updated, at iteration $t-2$ set 2 was updated, etc. We define $x_{c,i}$ as the portion of features that are available to clients in set c with padded zeros at indices that are not available to the clients in set c.

Then based on PrimalAggregation we have

$$\begin{aligned}
w^{t-1} &= \frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{t-1}x_{1,i} + \frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{t-2}x_{2,i} + \cdots + \frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{t-C}x_{C,i} \\
&= \mathbf{A}\alpha^{t-1} - \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\alpha_i^{t-1}x_{c,i} + \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\alpha_i^{t-c}x_{c,i}.
\end{aligned}$$

We next manipulate this equation to

$$\begin{aligned}
\mathbf{A}\alpha^{t-1} &= w^{t-1} + \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\alpha_i^{t-1}x_{c,i} - \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\alpha_i^{t-c}x_{c,i} \\
&= w^{t-1} + \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}(\alpha_i^{t-1} - \alpha_i^{t-c})x_{c,i} \\
&= w^{t-1} + \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_i^{t-d}x_{c,i}
\end{aligned}$$

and substitute this into (4) to obtain

$$\begin{aligned}\mathbb{E}[A-B|\alpha^{t-1}] \geq -\frac{H}{QN/C}\sum_{i=1}^{N}\sum_{b\in\mathcal{B}_c}[&l_i^*(-\alpha_i^{t-1}-\gamma_t s_{b,i}^t(u_i^{t-1}-\alpha_i^{t-1}))-l_i^*(-\alpha_i^{t-1})+\gamma_t s_{b,i}^t(w^{t-1})^T(u_i^{t-1}-\alpha_i^{t-1})x_i\\ &+\gamma_t s_{b,i}^t(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j})^T(u_i^{t-1}-\alpha_i^{t-1})x_i+\frac{1}{2\lambda}(\gamma_t s_{b,i}^t(u_i^{t-1}-\alpha_i^{t-1}))^2].\end{aligned}$$

Based on LocalDualMethod

$$s_{b,i}=\operatorname*{arg\,max}_{s\in[0,1]} -l_i^*(-\alpha_i^{t-1}-\gamma_t s(u_i^{t-1}-\alpha_i^{t-1}))-\gamma_t s(u_i^{t-1}-\alpha_i^{t-1})(w^{t-1})^T x_i-\frac{\gamma_t^2 s^2}{2\lambda}(u_i^{t-1}-\alpha_i^{t-1})^2,$$

therefore, we can replace $s_{b,i}$ in the terms included in the maximization with any value in $[0,1]$ and the bound still holds. Furthermore, we know that $s_{b,i}$ are the same for $b\in\mathcal{B}_c$. Therefore, we simply set $s_{b,i}=1$ in the terms in maximization and use convexity of l_i^* to get

$$\begin{aligned}\mathbb{E}[A-B|\alpha^{t-1}] &\geq -\frac{H}{N}\sum_{i=1}^{N}[l_i^*(-\alpha_i^{t-1}-\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))-l_i^*(-\alpha_i^{t-1})+\gamma_t(w^{t-1})^T(u_i^{t-1}-\alpha_i^{t-1})x_i\\ &+\gamma_t s_{b,i}^t(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j})^T(u_i^{t-1}-\alpha_i^{t-1})x_i+\frac{1}{2\lambda}(\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))^2]\\ &\geq -\frac{H}{N}\sum_{i=1}^{N}[-\gamma_t l_i^*(-\alpha_i^{t-1})+\gamma_t l_i^*(-u_i^{t-1})+\gamma_t(w^{t-1})^T((u_i^{t-1}-\alpha_i^{t-1}))x_i\\ &+\gamma_t s_{b,i}^t(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j})^T(u_i^{t-1}-\alpha_i^{t-1})x_i+\frac{1}{2\lambda}(\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))^2].\end{aligned} \tag{5}$$

By using the fact that $s_{b,i}^t\in[0,1]$, we bound the next to last term as

$$\begin{aligned}\gamma_t s_{b,i}^t(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j})^T(u_i^{t-1}-\alpha_i^{t-1})x_i &= \gamma_t s_{b,i}^t(u_i^{t-1}-\alpha_i^{t-1})(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j}^T x_i)\\ &\leq |\gamma_t s_{b,i}^t(u_i^{t-1}-\alpha_i^{t-1})(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j}^T x_i)|\\ &= \frac{\gamma_t s_{b,i}^t}{\lambda N}|(u_i^{t-1}-\alpha_i^{t-1})||(\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_j^{t-d}x_{c,j}^T x_i)|\\ &\leq \frac{\gamma_t|(u_i^{t-1}-\alpha_i^{t-1})|}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}|\Delta\alpha_j^{t-d}x_{c,j}^T x_i|\\ &= \frac{\gamma_t|(u_i^{t-1}-\alpha_i^{t-1})|}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}|\Delta\alpha_j^{t-d}||x_{c,j}^T x_i|\\ &\leq \frac{\gamma_t|(u_i^{t-1}-\alpha_i^{t-1})|}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}|\Delta\alpha_j^{t-d}|\\ &= \frac{\gamma_t|(u_i^{t-1}-\alpha_i^{t-1})|}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}|\frac{\gamma_{t-d}}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,j}^{t-d}|.\end{aligned}$$

By definition, $\Delta\alpha_{b,j}^{t-d} = s_{b,j}^{t}(u_j^{t-d-1} - \alpha_j^{t-d-1})$ and thus

$$
\begin{aligned}
\gamma_t(\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_i^{t-d}x_{c,i})^T(u_i^{t-1}-\alpha_i^{t-1})x_i &\leq \frac{\gamma_t|(u_i^{t-1}-\alpha_i^{t-1})|}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\gamma_{t-d}|u_j^{t-d-1}-\alpha_j^{t-d-1}|\cdot\max_{b\in\mathcal{B}_c}|s_{b,j}^t| \\
&\leq \frac{2L\gamma_t}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}2L\gamma_{t-d} \\
&\leq \frac{4L^2\gamma_t}{\lambda N}\sum_{c=2}^{C}\sum_{j=1}^{N}\sum_{d=1}^{c-1}\gamma_{t-d} \\
&\leq \frac{4L^2\gamma_t}{\lambda}\sum_{c=2}^{C}(c-1)\gamma_{t-c+1} \\
&\leq \frac{4L^2\gamma_t(C-1)^2}{\lambda}\gamma_{t-C+1} \\
&\leq \frac{4L^2\gamma_{t-C+1}^2(C-1)^2}{\lambda}.
\end{aligned}
$$

We continue with (5) to derive

$$
\begin{aligned}
\mathbb{E}[A-B|\alpha^{t-1}] \geq -\frac{H}{N}\sum_{i=1}^{N}[&-\gamma_t l_i^*(-\alpha_i^{t-1}) + \gamma_t l_i^*(-u_i^{t-1}) + \gamma_t(w^{t-1})^T(u_i^{t-1}-\alpha_i^{t-1})x_i \\
&+ \frac{4L^2\gamma_{t-C+1}^2(C-1)^2}{\lambda} + \frac{1}{2\lambda}(\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))^2].
\end{aligned}
$$

Now from convex conjugates we know that $l_i(x_i^T w^{t-1}) = -l_i^*(-u_i^{t-1}) - u_i^{t-1}x_i^T w^{t-1}$ which yields

$$
\begin{aligned}
\mathbb{E}[A-B|\alpha^{t-1}] \geq -\frac{H}{N}\sum_{i=1}^{N}[&-\gamma_t l_i^*(-\alpha_i^{t-1}) - \gamma_t l_i(x_i^T w^{t-1}) - \gamma_t(w^{t-1})^T\alpha_i^{t-1}x_i \\
&+ \frac{4L^2\gamma_{t-C+1}^2(C-1)^2}{\lambda} + \frac{1}{2\lambda}(\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))^2]. && (6) \\
& && (7)
\end{aligned}
$$

We next explore the primal-dual gap. We set $D = \frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_i^{t-d}x_{c,i}$ to obtain

$$
\begin{aligned}
P(w^{t-1}) - D(\alpha^{t-1}) &= (\frac{\lambda}{2}||w^{t-1}||^2 + \frac{1}{N}\sum_{i=1}^{N}l_i(x_i^T w^{t-1})) - (-\frac{\lambda}{2}||\mathbf{A}\alpha^{t-1}||^2 - \frac{1}{N}l_i^*(-\alpha_i^{t-1})) \\
&= (\frac{\lambda}{2}||w^{t-1}||^2 + \frac{1}{N}\sum_{i=1}^{N}l_i(x_i^T w^{t-1})) - (-\frac{\lambda}{2}||w^{t-1}+D||^2 - \frac{1}{N}l_i^*(-\alpha_i^{t-1})).
\end{aligned}
$$

We have that $\lambda||w^{t-1}||^2 = \lambda(w^{t-1})^T w^{t-1} = \lambda(w^{t-1})^T(\mathbf{A}\alpha^{t-1} - D) = \lambda(w^{t-1})^T\mathbf{A}\alpha^{t-1} - \lambda(w^{t-1})^T D$.

This results in

$$\begin{aligned}P(w^{t-1}) - D(\alpha^{t-1}) &= \frac{1}{N}\sum_{i=1}^{N}[l_i(x_i^T w^{t-1}) + l_i^*(-\alpha_i^{t-1})] + \lambda (w^{t-1})^T \mathbf{A}\alpha^{t-1} + \frac{\lambda}{2}||D||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}[l_i(x_i^T w^{t-1}) + l_i^*(-\alpha_i^{t-1}) + (w^{t-1})^T \alpha_i^{t-1} x_i] + \frac{\lambda}{2}||D||^2\end{aligned}$$

and in turn

$$P(w^{t-1}) - D(\alpha^{t-1}) - \frac{\lambda}{2}||D||^2 = \frac{1}{N}\sum_{i=1}^{N}[l_i(x_i^T w^{t-1}) + l_i^*(-\alpha_i^{t-1}) + (w^{t-1})^T \alpha_i^{t-1} x_i].$$

We substitute this into (7) to get

$$\mathbb{E}[A - B|\alpha^{t-1}] \geq H\gamma_t[P(w^{t-1}) - D(\alpha^{t-1})] - \frac{H\gamma_t\lambda}{2}||D||^2 - \frac{H}{N}\sum_{i=1}^{N}[\frac{4L^2\gamma_{t-C+1}^2(C-1)^2}{\lambda} + \frac{1}{2\lambda}(\gamma_t(u_i^{t-1} - \alpha_i^{t-1}))^2]. \tag{8}$$

To bound $||D||^2$ we use

$$\begin{aligned}||D||^2 &= ||\frac{1}{\lambda N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_i^{t-d} x_{c,i}||^2 \\ &= \frac{1}{\lambda^2 N^2}||\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\Delta\alpha_i^{t-d} x_{c,i}||^2 \\ &\leq \frac{(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}(\Delta\alpha_i^{t-d})^2||x_{c,i}||^2 \\ &\leq \frac{(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}(\Delta\alpha_i^{t-d})^2 \\ &= \frac{(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}(\frac{\gamma_{t-d}}{Q/C}\sum_{b\in\mathcal{B}_c}\Delta\alpha_{b,i}^{t-d})^2 \\ &= \frac{(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}(\gamma_{t-d}\frac{1}{Q/C}\sum_{b\in\mathcal{B}_c}s_{b,i}^t(u_i^{t-1} - \alpha_i^{t-1}))^2 \\ &\leq \frac{(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\gamma_{t-d}^2(u_i^{t-1} - \alpha_i^{t-1})^2\cdot(\frac{1}{Q/C}\sum_{b\in\mathcal{B}_c}s_{b,i})^2 \\ &\leq \frac{4L^2(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}\sum_{d=1}^{c-1}\gamma_{t-d}^2 \\ &\leq \frac{4L^2(C-1)^2}{\lambda^2 N}\sum_{c=2}^{C}\sum_{i=1}^{N}(c-1)\gamma_{t-c+1}^2 \\ &\leq \frac{4L^2(C-1)^4}{\lambda^2}\gamma_{t-C+1}^2.\end{aligned}$$

We substitute this into (8) to get

$$\begin{aligned}
\mathbb{E}[A-B|\alpha^{t-1}] &\geq H\gamma_t[P(w^{t-1})-D(\alpha^{t-1})] - \frac{H\gamma_t\lambda}{2}\frac{4L^2(C-1)^4}{\lambda^2}\gamma^2_{t-C+1} - \frac{H}{N}\sum_{i=1}^{N}[\frac{4L^2\gamma^2_{t-C+1}(C-1)^2}{\lambda} \\
&+ \frac{1}{2\lambda}(\gamma_t(u_i^{t-1}-\alpha_i^{t-1}))^2] \\
&\geq H\gamma_t[P(w^{t-1})-D(\alpha^{t-1})] - \frac{2L^2H(C-1)^4}{\lambda}\gamma^2_{t-C+1} - \frac{H}{N}\sum_{i=1}^{N}[\frac{4L^2\gamma^2_{t-C+1}(C-1)^2}{\lambda} + \frac{2L^2\gamma_t^2}{\lambda}] \\
&\geq H\gamma_t[P(w^{t-1})-D(\alpha^{t-1})] - \frac{2L^2H(C-1)^4}{\lambda}\gamma^2_{t-C+1} - H[\frac{4L^2\gamma^2_{t-C+1}(C-1)^2}{\lambda} + \frac{2L^2\gamma^2_{t-C+1}}{\lambda}] \\
&\geq H\gamma_t[P(w^{t-1})-D(\alpha^{t-1})] - \gamma^2_{t-C+1}H\underbrace{[\frac{2L^2(C-1)^4}{\lambda} + \frac{4L^2(C-1)^2}{\lambda} + \frac{2L^2}{\lambda}]}_{G}.
\end{aligned}$$

As in the previous proofs this yields

$$\mathbb{E}[\varepsilon_D^t] \leq (1-\frac{\gamma_t H}{N})\mathbb{E}[\varepsilon_D^{t-1}] + \frac{\gamma^2_{t-C+1}H}{N}G$$

and in turn

$$\begin{aligned}
\mathbb{E}[\varepsilon_D^t] &\leq (1-\frac{\gamma_t H}{N})[(1-\frac{\gamma_{t-1}H}{N})\mathbb{E}[\varepsilon_D^{t-2}] + \frac{\gamma^2_{t-C}H}{N}G] + \frac{\gamma^2_{t-C+1}H}{N}G \\
&\leq (1-\frac{\gamma_t H}{N})[(1-\frac{\gamma_{t-1}H}{N})[(1-\frac{\gamma_{t-2}H}{N})\mathbb{E}[\varepsilon_D^{t-3}] + \frac{\gamma^2_{t-C-1}H}{N}G] + \frac{\gamma^2_{t-C}H}{N}G] + \frac{\gamma^2_{t-C+1}H}{N}G \\
&\leq \underbrace{\prod_{\tau=C}^{t}(1-\frac{H\gamma_\tau}{N})\mathbb{E}[\varepsilon_D^{C-1}]}_{B_1} + \underbrace{\frac{GH}{N}\sum_{i=1}^{t-C+1}\gamma_i^2\prod_{\tau=i}^{t-C}(1-\frac{H\gamma_{\tau+C}}{N})}_{B_2}.
\end{aligned}$$

Note that $\frac{H}{N} \leq 1$ by definition.

First, we bound B_1. Using the fact that $\ln(1-x) \leq -x$ for every $0 \leq x \leq 1$ and that $\frac{1}{\tau}$ monotonically decreases for $\tau > 0$, we have

$$\begin{aligned}
\ln(\prod_{\tau=C}^{t}(1-\frac{H}{N\tau})) &= \sum_{\tau=C}^{t}\ln(1-\frac{H}{\tau N}) \\
&\leq -\frac{H}{N}\sum_{\tau=C}^{t}\frac{1}{\tau} \\
&\leq -\frac{H}{N}\int_{C}^{t+1}\frac{1}{\tau'}d\tau' \\
&= -\frac{H}{N}\ln(\tau')\Big|_{C}^{t+1} \\
&= -\frac{H}{N}\ln(\frac{t+1}{C}) \\
&= \ln((\frac{t+1}{C})^{-\frac{H}{N}}).
\end{aligned}$$

Thus we have

$$B_1 \leq \mathbb{E}[\varepsilon_D^{C-1}]\frac{C^{H/N}}{(t+1)^{H/N}}.$$

Next, we bound B_2. Using the fact that $\ln(1-x) \le -x$ for every $0 \le x \le 1$ and that $\frac{1}{\tau+C}$ monotonically decreases for $\tau > 0$ we have

$$\begin{aligned}
\ln(\prod_{\tau=i}^{t-C}(1-\frac{H\gamma_{\tau+C}}{N})) &= \sum_{\tau=i}^{t-C}\ln(1-\frac{H\gamma_{\tau+C}}{N}) \\
&\le -\frac{H}{N}\sum_{\tau=i}^{t-C}\frac{1}{\tau+C} \\
&\le -\frac{H}{N}\int_{i}^{t-C+1}\frac{1}{\tau'+C}d\tau' \\
&= -\frac{H}{N}\ln(\tau'+C)\Big|_{i}^{t-C+1} \\
&= -\frac{H}{N}\ln(\frac{t+1}{i+C}) \\
&= \ln((\frac{i+C}{t+1})^{\frac{H}{N}}).
\end{aligned}$$

Therefore, we have $\prod_{\tau=i}^{t-C}(1-\frac{H\gamma_{\tau+C}}{N}) \le (\frac{i+C}{t+1})^{\frac{H}{N}}$.

We know that $H/N \le 1$ and thus

$$\begin{aligned}
B_2 &= \frac{GH}{N}\sum_{i=1}^{t-C+1}\gamma_i^2\prod_{\tau=i}^{t-C}(1-\frac{H\gamma_{\tau+C}}{N}) \\
&\le \frac{GH}{N}\sum_{i=1}^{t-C+1}\gamma_i^2(\frac{i+C}{t+1})^{\frac{H}{N}} \\
&\le \frac{GH}{N}\sum_{i=1}^{t-C+1}\gamma_i^2(\frac{i+C}{t})^{\frac{H}{N}} \\
&= \frac{GH}{N}t^{-H/N}\sum_{i=1}^{t-C+1}\frac{(i+C)^{\frac{H}{N}}}{i^2} \\
&\le \frac{GH}{N}t^{-H/N}\sum_{i=1}^{t-C+1}\frac{i+C}{i^2} \\
&= \frac{GH}{N}t^{-H/N}\sum_{i=1}^{t-C+1}[\frac{1}{i}+\frac{C}{i^2}] \\
&\le \frac{GH}{N}t^{-H/N}\sum_{i=1}^{t-C+1}[\frac{1}{i}+\frac{C}{i}] \\
&= \frac{GH}{N}t^{-H/N}(C+1)\sum_{i=1}^{t-C+1}\frac{1}{i} \\
&= \frac{GH}{N}t^{-H/N}(C+1)\mathrm{Har}_{t-C+1},
\end{aligned}$$

where Har_{t-C+2} is the $(t-C+1)$-th harmonic number. Using the well-known bound on the harmonic numbers $\mathrm{Har}_n \le \ln(n)+1$, we have

$$B_2 \le \frac{GH(C+1)(\ln(t-C+1)+1)}{Nt^{H/N}}.$$

All that is left is to bound $\mathbb{E}[\varepsilon_D^{C-1}]$, which follows a similar recurrence relation as before but with slightly different terms. This results in a finite expression that depends on ε_D^0. Since $\varepsilon_D^0 \leq 1$ (proved in Lemma 20 of (Shalev-Shwartz and Zhang, 2013)), it is clear that $\mathbb{E}[\varepsilon_D^{C-1}]$ is bounded by a constant.

Thus we have

$$\mathbb{E}[\varepsilon_D^t] \leq \frac{J_1 + J_2(\ln(t - C + 1) + 1)}{t^{H/N}}$$

where

$$\begin{aligned} J_1 &= C^{H/N}\mathbb{E}[\varepsilon_D^{C-1}] \\ J_2 &= \frac{2HL^2(C+1)}{N\lambda}[(C-1)^4 + 2(C-1)^2 + 1]. \end{aligned}$$

This completes the proof.

□

2 Computational Study

We provide the full details of the experiments here. The details provided are enough to reproduce results, and the full set of experimental results are provided here as well.

2.1 Implementation

In order to investigate the performance of HyFDCA, we implemented both HyFDCA using the practical LocalDualMethod as described in Algorithm 5 and FedAvg using Python 3.10. NumPy was used to handle all matrices and matrix algebra for the MNIST and Covtype datasets, and due to the sparsity of the News20 dataset, we employed SciPy's sparse matrices.

Kubernetes was employed for container management to enable several concurrent experiments. Each dataset's HyFDCA and FedAvg experiments were run on identical pods on the same node (server) to ensure fair comparison of results. The specifications of the node/pod that each experiment was run on are described in Table 1.

Finally, homomorphic encryption was not actually conducted as part of the experiment. Our simulation simply found the number of necessary encryptions/operations in each iteration and used this to compute the estimated encryption time penalty based on published benchmarks for homomorphic encryption algorithms. For the first step of the algorithm, there is a decryption penalty only for the dual variables that were updated and are stale on the clients. This is variable depending on the iteration. For the SecureInnerProduct, only the sample indices that are going to be chosen to be updated on at least one of the clients need to be encrypted and sent to the server for addition and back for decryption. Since there may be significant overlap in the indices chosen between clients, this is also variable depending on the iteration. The dual variable updates found on each client must also be encrypted before sending to the server. The number of dual variables that then need to be decrypted by each client varies because the overlap in which dual variables are selected for updates. In the experiments, the number of necessary encryption/decryption operations were found in each iteration and used to calculate the encryption penalties to be added to the total wall time.

The total number of outer iterations used for each problem setting was 2,500 outer iterations for MNIST, 10,000 outer iterations for News20, and 30,000 outer iterations for Covtype.

Table 1: Computer Hardware Information

	MNIST	**News20**	**Covtype**
CPU	Intel Core i7-6850K CPU @ 3.60GHz	Intel Xeon Silver 4108 CPU @ 1.80GHz	Intel Core i7-6850K CPU @ 3.60GHz
Number of Processors	11	32	12
Maximum Memory Allocated	6 GiB	6 GiB	10 GiB

2.2 Data Partitioning

Table 2 outlines the key characteristics of each dataset. Sparsity is defined as the percent of zero values divided by the total number of values.

Table 2: Dataset Information

	MNIST	**News20**	**Covtype**
Type	Image	Text	Multivariate
Classes	10	2	2
Samples	70,000	19,996	581,012
Features	784	1,355,191	54
Sparsity	80.858%	99.966%	77.878%

Due to the differences in the datasets, specifically the meaning of the features, different methodologies were used to partition the data among a number of clients while remaining IID in nature.

Each MNIST sample is a 28x28 pixel image where values were normalized to $[0, 1]$. Each sample was split into four equal-sized quadrants, and thus $Q = 4$ for all problem settings. A bias feature of value 10 was appended to the fourth quadrant's data; this larger value was chosen to prevent the bias term from being affected as strongly by regularization. Each client was then provided with $\frac{4N}{KQ}$ sample quadrants where KQ is the total number of clients. The first quarter of clients received features from the first quadrant of the images, the second quarter of clients received features from the second quadrant of the images, and so on.

News20 and Covtype used a different assignment procedure as geometrically segmenting features makes no sense for either dataset. First, the samples are evenly divided into K number of sample groups. Within each sample group, Q clients are defined. Examining each sample within a segment separately, each client receives $1/Q$ of the non-zero features. No bias term was employed for either the News20 or Covtype datasets. As such, there were $K \cdot Q$ clients. Similar to MNIST, this process ensures that there is no data overlap between clients and minimizes data imbalances.

2.3 Hyperparameter Tuning

For HyFDCA, the only hyperparameter to tune was the number of inner iterations because no diminishing learning rate is needed for convergence. FedAvg requires tuning of the number of inner iterations on each client and the constants a and b in the learning rate $\gamma_t = \frac{a}{b+\sqrt{t}}$. Inner iterations in FedAvg correspond to the number of times the primal weights are updated in a given iteration. Our implementation used a batch size of one to find stochastic gradients, and made H updates to the primal weights in each iteration.

The number of inner iterations was not directly tuned. Rather, we employed an inner iterations constant, IIC, that defined the number of inner iterations as a function of the number of training samples, total number of clients, and IIC as follows:

$$H = \left\lceil \frac{IIC \cdot N}{KQ} \right\rceil. \tag{9}$$

We employed a random search method in order to collect sufficient data to select hyperparameters (Bergstra and Bengio, 2012). Nine client-fraction combinations were tuned independently for each dataset. For MNIST, this consisted of all combinations of 5, 500, and 5000 clients with 0.1, 0.5, 0.9 fraction of clients available. For News20 and Covtype, all combinations using three sets of (sample groups, feature groups) - $(3, 3)$, $(5, 5)$, $(12, 12)$ - were examined using the same three fractions - 0.1, 0.5, 0.9 - leading to a 9 client-fraction combinations. Based on an understanding of reasonable hyperparameters from preliminary testing, the search area was bounded as follows:

- HyFDCA IIC: $[\frac{\min clients}{samples}, 1.0]$
- FedAvg IIC: $[\frac{\min clients}{samples}, 5.0]$
- FedAvg a: $[10^{-5}, 25.0]$
- FedAvg b: $[10^{-5}, 25.0]$.

We used (9) to calculate appropriate ranges for IIC for each dataset. From preliminary testing, we believed that smaller values of all hyperparameters were more likely to be chosen as optimal leading us to sample values from a logarithmic distribution. We randomly sampled a value, x_i, from the uniform distribution $[log_{10}x_{min}, log_{10}x_{max}]$ where x_{min} and x_{max} are the lower and upper bounds of the given hyperparameter, respectively. The randomly selected hyperparameter is therefore defined as 10^{x_i}.

Hyperparameter tuning in the case of federated learning is more complicated due to the large number of competing metrics that define an algorithm's performance. For example, we may wish to minimize the total number of outer iterations (communication rounds) to reach a satisfactory loss function value but also wish to minimize the total computation time on each client because of computational limits on devices such as smart phones. These two goals are directly conflicting. For this reason, we frame this hyperparameter selection problem with multiple metrics as a multiobjective optimization problem where the optimal solution must be selected from the Pareto-Optimal front. We solve this using Gray Relational Analysis as described in (Wang and Rangaiah, 2017). The metrics we use for the hyperparameter selection are as follows:

1. Average runtime per iteration with 0.000 seconds of round-trip latency
2. Average runtime per iteration with 0.2575 seconds of round-trip latency
3. Average runtime per iteration with 0.8000 seconds of round-trip latency
4. Average of last 5 loss function values
5. Maximum validation accuracy
6. Volatility - standard deviation of differences in consecutive loss function values
7. Number of iterations to reach 90% progress of minimizing the loss function

Here, 0.000 second of round-trip latency is meant to represent the most ideal scenario, 0.2575 seconds represents a long-distance server connection (US-Singapore AWS server representation (Adorjan, 2020)), and 0.8000 represents a GEO satellite connection (Telesat, 2017). There are 4.5 round-trips of information transmissions for HyFDCA and 1.0 round-trips for FedAvg per iteration.

All divergent runs were excluded from GRA. All optimally selected hyperparameters for both HyFDCA and FedAvg are listed in Tables 3-5.

Table 3: MNIST optimal hyperparameters selected and employed

	HyFDCA IIC	**FedAvg IIC**	**FedAvg a**	**FedAvg b**
5, 0.1	0.0001876	0.04872	0.05857	0.1124
5, 0.5	0.001514	0.04872	0.05857	0.1124
5, 0.9	0.002816	0.01718	0.02867	0.0004823
500, 0.1	0.0006865	0.1718	0.02867	0.0004823
500, 0.5	0.001187	0.06561	0.04185	0.00004166
500, 0.9	0.08337	0.04872	0.05857	0.1124
5000, 0.1	0.002816	1.618	0.003995	0.8498
5000, 0.5	0.002816	1.618	0.003995	0.8498
5000, 0.9	0.0006109	1.618	0.003995	0.8498

Table 4: News20 optimal hyperparameters selected and employed

	HyFDCA IIC	**FedAvg IIC**	**FedAvg a**	**FedAvg b**
5, 0.1	0.02256	0.003037	0.8753	0.4597
3, 3, 0.5	0.009124	0.002873	4.891	4.496
3, 3, 0.9	0.007732	0.002873	4.891	4.496
5, 5, 0.1	0.01769	0.002873	4.891	4.496
5, 5, 0.5	0.02173	0.003037	0.8753	0.4597
5, 5, 0.9	0.02577	0.002873	4.891	4.496
12, 12, 0.1	0.1362	4.542	3.792	0.1815
12, 12, 0.5	0.1091	4.542	3.792	0.1815
12, 12, 0.9	0.1091	4.542	3.792	0.1815

2.4 Analysis Methods

Table 6 shows the number of centralized iterations to find the optimal centralized solutions along with the minimum loss value for these centralized runs.

Due to a combination of factors including the large number of outer iterations, characteristics of the datasets, and the nature of the algorithms employed, there was significant volatility in the loss function making plots hard to read. We applied a

Table 5: Covtype optimal hyperparameters selected and employed

	HyFDCA IIC	**FedAvg IIC**	**FedAvg a**	**FedAvg b**
5, 0.1	0.00004962	0.02291	0.4475	0.0003501
3, 3, 0.5	0.0002264	0.02291	0.4475	0.0003501
3, 3, 0.9	0.00002769	0.02291	0.4475	0.0003501
5, 5, 0.1	0.0002049	0.03647	0.3247	4.514
5, 5, 0.5	0.0009082	0.02291	0.4475	0.0003501
5, 5, 0.9	0.00009929	0.03647	0.3247	4.514
12, 12, 0.1	0.0007434	0.0005590	0.05344	0.004065
12, 12, 0.5	0.006039	0.0005590	0.05344	0.004065
12, 12, 0.9	0.00033673	0.2494	11.93	0.05131

Table 6: Centralized Training Information

	MNIST	**News20**	**Covtype**
Iterations	30,000	40,000	1,000,000
Min Loss	0.07242	0.008017	0.5286

moving average in order to smooth the loss function solely for the sake of readability in plotting. The moving average had a window, ω, of 50 outer iteration for MNIST, 150 for News20, and 300 for Covtype.

Additionally, we employed a relative time measure and percent of outer iterations completed for clearer and normalized plotting. These were performed separately for each dataset, and it preserves the relative differences between the two algorithms while allowing the results from all three datasets to be plotted together. Such values are defined as follows:

$$T_R = \frac{T}{\max\{T_1, T_2\}} \tag{10}$$

where T_R is the relative time value, T_1 is the total wall time for HyFDCA, T_2 is the total wall time for FedAvg, and T is the original time quantity for either HyFDCA or FedAvg, and we have

$$t_R = \frac{t}{t_{max} - \omega} \tag{11}$$

where t_R is the percent of outer iterations completed, ω is the width of the moving average window, t_{max} is the maximum number of outer iterations, and t is the number of outer iterations.

2.5 Full Results

Due to the large number of problem settings investigated and the various metrics of interest, only selected plots were included in the main paper. Additional data are included herein to ensure completeness and transparency in reporting.

Accompanying Figure 4 of the main paper, appendix Figures 1 and 2 show the effect of different problem settings on the performance of HyFDCA for the MNIST and Covtype datasets, respectively.

Figure 5 of the main paper displays computational and encryption time data only for three client-fraction settings. The full data are represented in Tables 7, 8, and 9 for MNIST, News20 and Covtype, respectively.

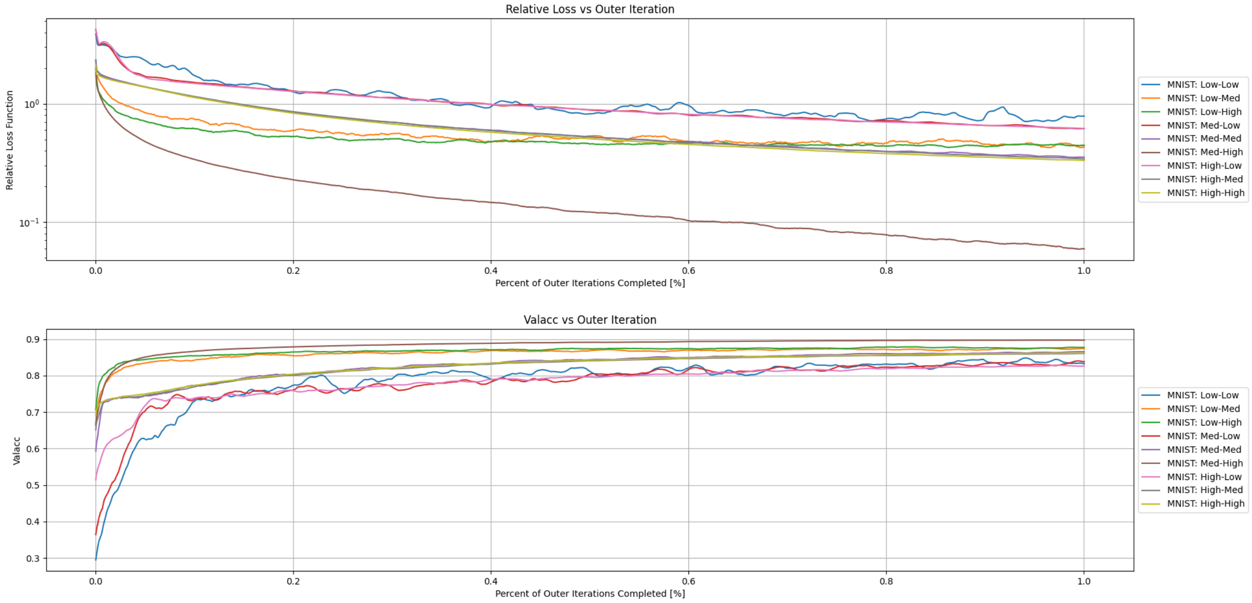

Figure 1: Effect of number of clients and fraction of participating clients on HyFDCA performance on MNIST.

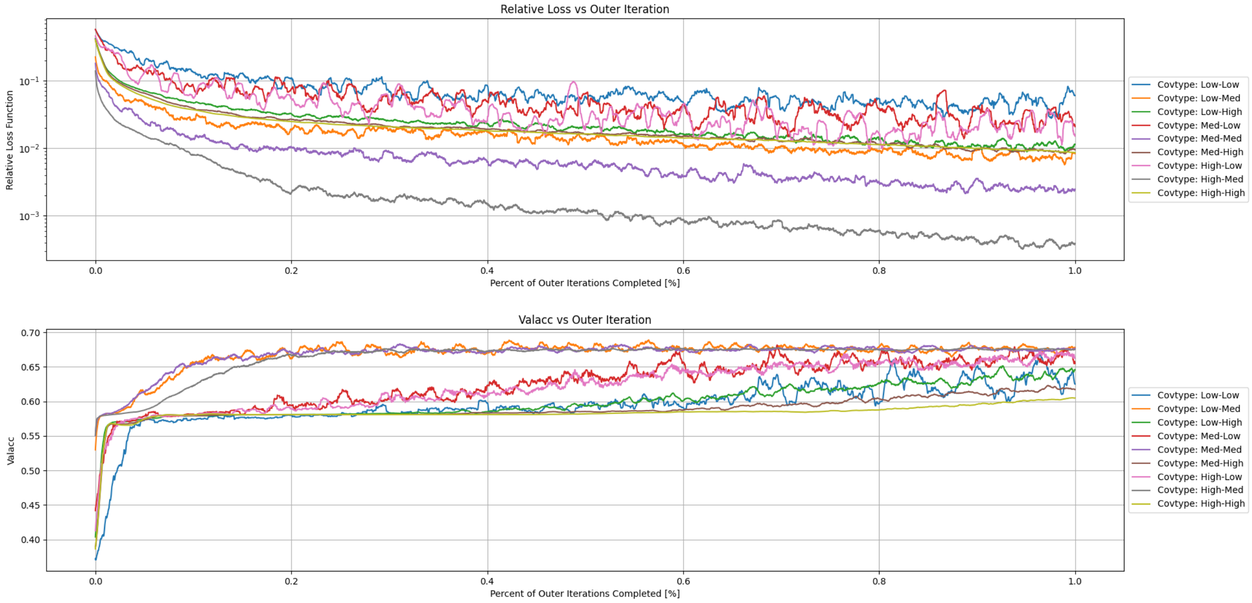

Figure 2: Effect of number of clients and fraction of participating clients on HyFDCA performance on Covtype.

Table 7: Computational and Encryption Times for MNIST (seconds)

	HyFDCA Computational	**HyFDCA Encryption**	**FedAvg Computational**
5, 0.1	0.7228	0.1107	0.2268
5, 0.5	0.7895	3.164	0.2544
5, 0.9	0.5174	11.32	0.1543
500, 0.1	0.3478	2.037	0.1027
500, 0.5	1.104	13.67	0.1922
500, 0.9	1.684	431.0	0.2064
5000, 0.1	1.778	61.69	0.5775
5000, 0.5	7.246	1136	0.6021
5000, 0.9	13.48	3903	0.5455

Table 8: Computational and Encryption Times for News20 (seconds)

	HyFDCA Computational	**HyFDCA Encryption**	**FedAvg Computational**
3, 3, 0.1	0.1554	0.6173	1.074
3, 3, 0.5	0.1760	1.815	1.023
3, 3, 0.9	0.1976	2.661	0.9357
5, 5, 0.1	0.1828	0.6312	0.7311
5, 5, 0.5	0.1965	3.533	0.6438
5, 5, 0.9	0.2283	6.760	0.6572
12, 12, 0.1	0.2309	3.773	1.544
12, 12, 0.5	0.2884	13.26	1.376
12, 12, 0.9	0.3398	24.65	1.409

Table 9: Computational and Encryption Times for Covtype (seconds)

	HyFDCA Computational	**HyFDCA Encryption**	**FedAvg Computational**
3, 3, 0.1	0.02608	0.03922	0.04077
3, 3, 0.5	0.06815	1.144	0.05855
3, 3, 0.9	0.1073	0.3229	0.05638
5, 5, 0.1	0.03468	0.1843	0.02457
5, 5, 0.5	0.1064	3.169	0.02655
5, 5, 0.9	0.1313	0.6798	0.03297
12, 12, 0.1	0.08541	0.5996	0.001403
12, 12, 0.5	0.2455	16.879	0.002198
12, 12, 0.9	0.4046	3.632	0.03100