FLASH: Automating Federated Learning using CASH (Supplementary Material)

Md Ibrahim Ibne Alam[1] **Koushik Kar**[1] **Theodoros Salonidis**[2] **Horst Samulowitz**[2]

[1]Department of ECSE, Rensselaer Polytechnic Institute , Troy, NY, USA - 12180
[2]IBM T.J. Watson Research Center , Yorktown Heights, NY, USA - 10598

In our analysis, for simplicity we assume that the dataset (or equivalently, its distribution) $\mathcal{D}$ has finite discrete support. Our results generalize when this assumption is relaxed, although the analysis in that case becomes more complex. Also, for ease of exposition, we are going to use $A \in (A^{(1)}, \cdots, A^{(J)})$ to denote a generic Algorithm, and $*$ to denote the optimum algorithm A^*.

Proof of Lemma 1: We define the true loss projection for an algorithm A as $\underline{LP}(A, a_n) = \underline{\ell}(A, a_n) + (1 - a_n) \cdot \underline{\ell}'(A, a_n)$, where $\underline{\ell}$ represents the true training loss function (assuming cross-validation), and $\underline{\ell}'$ its derivative. Similarly $LP(A, a_n)$ is defined as the loss projection calculated from $\ell(A, a_n)$ and computed as $\ell(A, a_n) + (1 - a_n) \cdot \ell'(A, a_n)$. where ℓ is the loss function computed by FLASH and ℓ' its discrete derivative (defined later). Hence we can use Taylor series expansion on true loss function $\underline{\ell}$ with $0 < a_n, a_m \leq 1$, to get the following equations;

$$\begin{aligned}
\underline{\ell}(*, 1) &= \underline{\ell}(*, a_m) + (1 - a_m) \cdot \underline{\ell}'(*, a_m) \\
&\quad + \frac{1}{2}(1 - a_m)^2 \cdot \underline{\ell}''(*, \bar{a}_m), \\
\text{or, } \underline{\ell}(*, 1) &= \underline{LP}(*, a_m) + \frac{1}{2}(1 - a_m)^2 \cdot \underline{\ell}''(*, \bar{a}_m) \qquad (3) \\
\text{and, } \quad \underline{\ell}(A, 1) &= \underline{\ell}(A, a_n) + (1 - a_n) \cdot \underline{\ell}'(A, a_n) \\
&\quad + \frac{1}{2}(1 - a_n)^2 \cdot \underline{\ell}''(A, \bar{a}_n), \\
\text{or, } \underline{\ell}(A, 1) &= \underline{LP}(A, a_n) + \frac{1}{2}(1 - a_n)^2 \cdot \underline{\ell}''(A, \bar{a}_n). \qquad (4)
\end{aligned}$$

Where $a_m \leq \bar{a}_m \leq 1$, $a_n \leq \bar{a}_n \leq 1$, $*$ is the optimum algorithm ensuring that minimized (1), and $A \in \mathcal{A}$ is any other algorithm. We know from the definition of $\ell^* = \underline{\ell}(*, 1) \leq \underline{\ell}(A, 1)$. Hence we can write the following using (3) and (4);

$$\begin{aligned}
&\underline{LP}(*, a_m) + \frac{1}{2}(1 - a_m)^2 \cdot \underline{\ell}''(*, \bar{a}_m) \\
&\leq \underline{LP}(A, a_n) + \frac{1}{2}(1 - a_n)^2 \cdot \underline{\ell}''(A, \bar{a}_n) \\
&\text{or, } \underline{LP}(*, a_m) \leq \underline{LP}(A, a_n) + \frac{1}{2}(1 - a_n)^2 \cdot \underline{\ell}''(A, \bar{a}_n) \\
&\text{or, } \underline{LP}(*, a_m) - \underline{LP}(A, a_n) \leq B/2. \qquad (5)
\end{aligned}$$

The last line in (5) comes from the fact that $\underline{\ell}''(A, \bar{a}_n) < B$. We recall that σ is defined in the following way,

$$|\ell(A, a) - \underline{\ell(A, a)}| \leq \sigma \qquad (6)$$

Since, ℓ is calculated at discrete points (i.e., $a_{m-1}, a_m, ..$) hence the discrete derivative of ℓ, ℓ' is defined as

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

$\frac{\ell(A,a_m)-\ell(A,a_{m-1})}{\delta_m}$, for some value of a_m and $\delta_m = a_m - a_{m-1}$. Hence we have the following using (6);

$$\begin{aligned}\ell'(A,a_m) =& \frac{\ell(A,a_m)-\ell(A,a_{m-1})}{\delta_m}\\ \leq& \frac{\underline{\ell}(A,a_m)+\sigma-\underline{\ell}(A,a_{m-1})+\sigma}{\delta_m}\\ =& \frac{\underline{\ell}(A,a_m)-\underline{\ell}(A,a_{m-1})+2\sigma}{\delta_m}.\end{aligned} \tag{7}$$

Using Taylor series expansion we get,

$$\begin{aligned}\underline{\ell}(A,a_{m-1}) &= \underline{\ell}(A,a_m) - \delta_m\underline{\ell}'(A,a_m) + \frac{\delta_m^2}{2}\underline{\ell}''(A,\bar{a})\\ \text{or, } \underline{\ell}(A,a_m) - \underline{\ell}(A,a_{m-1}) &= \delta_m\underline{\ell}'(A,a_m) - \frac{\delta_m^2}{2}\underline{\ell}''(A,\bar{a}),\end{aligned}$$

for some $\bar{a} \in [a_{m-1}, a_m]$. From which we get the following two inequalities;

$$\frac{\underline{\ell}(A,a_m)-\underline{\ell}(A,a_{m-1})}{\delta_m} \leq \underline{\ell}'(A,a_m), \tag{8}$$

$$\text{and, } \frac{\underline{\ell}(A,a_m)-\underline{\ell}(A,a_{m-1})}{\delta_m} \geq \underline{\ell}'(A,a_m) - \frac{\delta_m}{2}\cdot B. \tag{9}$$

Comparing (7) with (8) we get the following:

$$\ell'(A,a_m) = \frac{\ell(A,a_m)-\ell(A,a_{m-1})}{\delta_m} \leq \underline{\ell}'(A,a_m) + \frac{2\sigma}{\delta_m}. \tag{10}$$

Hence, we can write,

$$\begin{aligned}LP(*,a_m) &= \ell(*,a_m) + (1-a_m)\cdot\ell'(*,a_m)\\ &\leq \underline{\ell}(*,a_m) + \sigma + (1-a_m)(\underline{\ell}'(*,a_m) + \frac{2\sigma}{\delta_m})\\ \text{or, } LP(*,a_m) &\leq \underline{LP}(*,a_m) + \sigma + \frac{2\sigma}{\delta_m}.\end{aligned} \tag{11}$$

On the other hand, from the definition of LP and (10) we have,

$$\begin{aligned}&LP(A,a_m) = \ell(A,a_m) + (1-a_m)\ell'(A,a_m)\\ &\geq \underline{\ell}(A,a_m) - \sigma + (1-a_m)\cdot\left(\frac{\ell(A,a_m)-\ell(A,a_m)}{\delta_m}\right)\\ &= \underline{\ell}(A,a_m) - \sigma\\ &\qquad + (1-a_m)\left(\frac{\underline{\ell}(A,a_m)-\underline{\ell}(A,a_m)-2\sigma}{\delta_m}\right)\\ &\geq \underline{\ell}(A,a_m) - \sigma + (1-a_m)\left(\underline{\ell}'(A,a_m) - \frac{\delta_m}{2}B - \frac{2\sigma}{\delta_m}\right)\\ &= \underline{\ell}(A,a_m) + (1-a_m)\underline{\ell}'(A,a_m)\\ &\qquad - \sigma - (1-a_m)\left(\frac{\delta_m}{2}B + \frac{2\sigma}{\delta_m}\right)\\ &\geq \underline{LP}(A,a_m) - \left(\sigma + \frac{\delta_m}{2}B + \frac{2\sigma}{\delta_m}\right).\end{aligned} \tag{12}$$

Now, we are ready to bound the value of $LP(*, a_m) - LP(A, a_m)$ using (11) and (12).

$$\begin{aligned}
&LP(*, a_m) - LP(A, a_m) \\
\leq& \underline{LP}(*, a_m) + \sigma + \frac{2\sigma}{\delta_m} - \\
&\qquad \underline{LP}(A, a_m) + \sigma + \frac{\delta_m B}{2} + \frac{2\sigma}{\delta_m} \\
=& \underline{LP}(*, a_m) - \underline{LP}(A, a_m) + 2\sigma + \frac{4\sigma}{\delta_m} + \frac{\delta_m}{2} B \\
\leq& B/2 + 2\sigma + \frac{4\sigma}{\delta_m} + \frac{\delta_m}{2} B \text{ [using (5)]} \\
\leq& B + 2\sigma + \frac{4\sigma}{\delta}. \qquad (13)
\end{aligned}$$

This implies that for any algorithm (A), the $LP(A, a_m)$ cannot be less $LP(*, a_m)$ by a value greater than $B + 2\sigma + \frac{4\sigma}{\delta}$. Thus, if $B + 2\sigma + \frac{4\sigma}{\delta} \leq \Delta$, it ensures the training of the optimum algorithm $(*)$, which proves Lemma 1. □

Proof of Theorem 2: Let $A^\dagger$ be the Algorithm chosen by the FLASH, and $*$ is the optimum algorithm. We know from Lemma 1 that if $B + 2\sigma + \frac{4\sigma}{\delta}$, then $*$ will be in the final choice of algorithms alongside $A^\dagger$ (when $a = 1$). Since, $A^\dagger$ was chosen by FLASH instead of $*$,

$$\ell(A^\dagger, 1) \leq \ell(*, 1). \qquad (14)$$

Since, $\underline{\ell}(A^\dagger, 1) \geq \underline{\ell}(*, 1)$, hence we need to bound the value of $\underline{\ell}(A^\dagger, 1) - \underline{\ell}(*, 1)$ to prove Theorem 2.

$$\begin{aligned}
&\underline{\ell}(A^\dagger, 1) - \underline{\ell}(*, 1) \\
=& \underline{\ell}(A^\dagger, 1) - \ell(A^\dagger, 1) + \ell(A^\dagger, 1) - \underline{\ell}(*, 1) \\
\leq& \sigma + \ell(A^\dagger, 1) - \underline{\ell}(*, 1) \\
=& \sigma + \ell(*, 1) - \underline{\ell}(*, 1) \text{ [using (14)]} \\
\leq& \sigma + \sigma = 2\sigma. \qquad (15)
\end{aligned}$$

In the calculation above, at the third line from the top, we have bounded $\underline{\ell}(A^\dagger, 1) - \ell(A^\dagger, 1)$ by σ which is true for all cases. However, for the case of RM and $LKBM$, since $A^\dagger$ is chosen over $*$ in the revalidation step, $\ell(A^\dagger, 1) \leq \ell(*, 1) = \underline{\ell}(*, 1)$. Hence, the bound in (15) reduces to σ from 2σ for those two FL-HPO methods. □

Proof of Theorem 3: Consider any round n is which an Algorithm A is allocated additional data for training. Since, $\Delta > B + 2\sigma + \frac{4\sigma}{\delta}$, the optimum algorithm $*$ is included for training in that round as well. We use (11) and (12) to get the following upper bound for $\underline{LP}(A, a_n) - \underline{LP}(*, a_m)$,

$$\begin{aligned}
&\underline{LP}(A, a_n) - \underline{LP}(*, a_m) \\
\leq& LP(A, a_n) + \sigma + \frac{\delta_n B}{2} + \frac{2\sigma}{\delta_n} - LP(*, a_m) + \sigma + \frac{2\sigma}{\delta_m} \\
=& LP(A, a_n) - LP(*, a_m) + 2\sigma + \frac{2\sigma}{\delta_n} + \frac{2\sigma}{\delta_m} + \frac{\delta_n B}{2}. \\
\leq& LP(A, a_n) - LP(*, a_m) + 2\sigma + \frac{4\sigma}{\delta} + \frac{B}{2}. \qquad (16)
\end{aligned}$$

Let O the algorithm with the best LP in that round (n). Let n, m and p be the last round in which the LP values of $A, *$ and O have been updated $(n, m, p \leq M)$. Then we have,

$$\begin{aligned}
0 \leq LP(*, a_m) - LP(O, a_p) \leq \Delta \\
\text{and, } 0 \leq LP(A, a_n) - LP(O, a_p). \leq \Delta
\end{aligned}$$

From above two inequalities we have,

$$LP(A, a_n) - LP(*, a_m) \leq \Delta. \qquad (17)$$

Then using (3) and (4) we get,

$$\begin{aligned}
&\underline{\ell}(A,1) - \underline{\ell}(*,1) = \epsilon_A = \underline{LP}(A,a_n) - \underline{LP}(*,a_m) \\
&\quad + \frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n) - \frac{1}{2}(1-a_m)^2 \cdot \underline{\ell}''(*,\bar{a}_m) \\
\text{or, } &\underline{LP}(A,a_n) - \underline{LP}(*,a_m) \\
&\quad \geq \epsilon_A - \frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n). \qquad (18)
\end{aligned}$$

Now using (16) and (18), we get,

$$\begin{aligned}
&LP(A,a_n) - LP(*,a_m) + 2\sigma + \frac{4\sigma}{\delta} + \frac{B}{2} \\
&\quad \geq \epsilon_A - \frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n) \\
\text{or, } &LP(A,a_n) - LP(*,a_m) \geq \epsilon_A - 2\sigma - \frac{4\sigma}{\delta} \\
&\quad - \frac{B}{2} - \frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n). \qquad (19)
\end{aligned}$$

To finalize the proof, we use (17) and (19) to get,

$$\begin{aligned}
&\Delta \geq \epsilon_A - 2\sigma - \frac{B}{2} + \frac{4\sigma}{\delta} \\
&\quad - \frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n) \\
\text{or, } &\frac{1}{2}(1-a_n)^2 \cdot \underline{\ell}''(A,\bar{a}_n) \\
&\quad \geq \epsilon_A - 2\sigma - \frac{B}{2} - \frac{4\sigma}{\delta} - \Delta \\
\text{or, } &(1-a_n)^2 \geq \frac{\epsilon_A - 2\sigma - \frac{B}{2} - \frac{4\sigma}{\delta} - \Delta}{B/2} \\
\text{or, } &a_n \leq 1 - \sqrt{\frac{\epsilon_A - 2\sigma - \frac{B}{2} - \frac{4\sigma}{\delta} - \Delta}{B/2}}. \qquad (20)
\end{aligned}$$

Therefore, Algorithm A does not get any training data greater than $1 - \sqrt{\frac{\epsilon_A - 2\sigma - \frac{B}{2} - \frac{4\sigma}{\delta} - \Delta}{B/2}}$ in round n. The result in Theorem 3 easily follows from this. □

Preliminaries for Proof of Theorem 4: To prove Theorem 4, we define a new loss function ($\tilde{\ell}(A^{(j)}_{\boldsymbol{\lambda}}, \hat{\mathcal{D}})$) as,

$$\begin{aligned}
&\tilde{\ell}(A^{(j)}_{\boldsymbol{\lambda}}, \hat{\mathcal{D}}) = \sum_i \alpha_i \, \mathcal{L}(\mathcal{F}(A^{(j)}_{\boldsymbol{\lambda}}, \cup_i \hat{\mathcal{D}}_i), \hat{\mathcal{D}}_i). \\
\text{Hence, } &\hat{\ell}(A^{(j)}, \hat{\mathcal{D}}) = \min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}^j} \tilde{\ell}(A^{(j)}_{\boldsymbol{\lambda}}, \hat{\mathcal{D}}),
\end{aligned}$$

where $\hat{\ell}(A^{(j)}_{\boldsymbol{\lambda}}, \hat{\mathcal{D}})$ is as defined in the Theoretical analysis section. We define the loss rate computed by FLASH for algorithm $A^{(j)}$ and HP $\boldsymbol{\lambda}$ on dataset $\mathcal{D}^a$ as $\ell(A^{(j)}_{\boldsymbol{\lambda}}, \mathcal{D}^a)$. Finally, for some Algorithm A and HP $\boldsymbol{\lambda}$ we make the following assumptions to prove Theorem 4,

$$|\tilde{\ell}(A_{\boldsymbol{\lambda}}, \hat{\mathcal{D}}_1) - \tilde{\ell}(A_{\boldsymbol{\lambda}}, \hat{\mathcal{D}}_2)| \leq \beta' \nu(\hat{\mathcal{D}}_1, \hat{\mathcal{D}}_2). \qquad (21)$$

$$|\boldsymbol{\lambda}(\hat{\mathcal{D}}_1) - \boldsymbol{\lambda}(\hat{\mathcal{D}}_2)| \leq \beta'' \nu(\hat{\mathcal{D}}_1, \hat{\mathcal{D}}_2). \qquad (22)$$

$$|\tilde{\ell}(A_{\boldsymbol{\lambda}_1}, \hat{\mathcal{D}}) - \tilde{\ell}(A_{\boldsymbol{\lambda}_2}, \hat{\mathcal{D}})| \leq \beta''' |\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2|. \qquad (23)$$

In (23), $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are two HP settings of Algorithm A. For two different datasets $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$ we use (22) and (23) to derive another inequality which will be helpful in the proof of Theorem 4.

$$\begin{aligned}
&|\hat{\ell}(A,\hat{\mathcal{D}}_1)-\hat{\ell}(A,\hat{\mathcal{D}}_2)|\\
=&|\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_1)},\hat{\mathcal{D}}_1)-\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_2)},\hat{\mathcal{D}}_2)|\\
\leq&|\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_1)},\hat{\mathcal{D}}_1)-\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_1)},\hat{\mathcal{D}}_2)|\\
&\quad+|\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_1)},\hat{\mathcal{D}}_2)-\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_2)},\hat{\mathcal{D}}_2)|\\
\leq&\beta'\nu(\hat{\mathcal{D}}_1,\hat{\mathcal{D}}_2)+\beta'''|\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_1)-\boldsymbol{\lambda}^*(\hat{\mathcal{D}}_2)|\\
\leq&\beta'\nu(\hat{\mathcal{D}}_1,\hat{\mathcal{D}}_2)+\beta'''\beta''\nu(\hat{\mathcal{D}}_1,\hat{\mathcal{D}}_2)\\
=&\beta\nu(\hat{\mathcal{D}}_1,\hat{\mathcal{D}}_2),
\end{aligned}\tag{24}$$

where $\beta=\beta'+\beta''\beta'''$.

Proof of Theorem 4: For any algorithm A and dataset $\mathcal{D}^a$, let $\boldsymbol{\lambda}^\dagger$ as the HP setting chosen by FLASH and $\boldsymbol{\lambda}^*$ as the optimum HP setting that minimizes $\tilde{\ell}$. Then from the definition of σ we have,

$$\begin{aligned}
\sigma&\geq|\ell(A_{\boldsymbol{\lambda}^\dagger},\mathcal{D}^a)-\underline{\ell}(A,a)|\\
&=|\ell(A_{\boldsymbol{\lambda}^\dagger},\mathcal{D}^a)-\mathbb{E}_{\hat{\mathcal{D}}^a\in\hat{\boldsymbol{\mathcal{D}}}^a}\min_{\boldsymbol{\lambda}}\hat{\ell}(A_{\boldsymbol{\lambda}},\hat{\mathcal{D}}^a)|\\
&=|\ell(A_{\boldsymbol{\lambda}^\dagger},\mathcal{D}^a)-\mathbb{E}_{\hat{\mathcal{D}}^a\in\hat{\boldsymbol{\mathcal{D}}}^a}\ \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}^a)},\hat{\mathcal{D}}^a)|\\
&\leq|\ell(A_{\boldsymbol{\lambda}^\dagger},\mathcal{D}^a)-\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)},\mathcal{D}^a)|\\
&\quad+|\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)},\mathcal{D}^a)-\mathbb{E}_{\hat{\mathcal{D}}^a\in\hat{\boldsymbol{\mathcal{D}}}^a}\ \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}^a)},\hat{\mathcal{D}}^a)|\\
&=|X|+|Y|,
\end{aligned}$$

where we denoted the first term in $|\cdot|$ as X and the second term to be Y. Now, we will bound X and Y individually since X depends on the FL-HPO variant used (as we will observe), whereas Y is independent of that. Since, Y is common for all HPO-aggregation variants, so we start by bounding the value of Y.

Bounding $|Y|$: For any specific $\hat{\mathcal{D}}^a$ we define,

$$\begin{aligned}
\hat{y}(\hat{\mathcal{D}^a})&=\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)},\mathcal{D}^a)-\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\hat{\mathcal{D}}^a)},\hat{\mathcal{D}}^a)\\
&=\hat{\ell}(A,\mathcal{D}^a)-\hat{\ell}(A,\hat{\mathcal{D}}^a)
\end{aligned}$$

From (24), $|\hat{\ell}(A,\mathcal{D}^a)-\hat{\ell}(A,\hat{\mathcal{D}}^a)|\leq\beta\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)$, we get,

$$-\beta\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)\leq\hat{y}(\hat{\mathcal{D}^a})\leq\beta\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a).$$

Taking expectation with respect to $\mathcal{D}^a$ on all sides,

$$\begin{aligned}
&-\beta\mathbb{E}_{\hat{\mathcal{D}}^a}\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)\leq\mathbb{E}_{\hat{\mathcal{D}}^a}\hat{y}(\hat{\mathcal{D}^a})\leq\beta\mathbb{E}_{\hat{\mathcal{D}}^a}\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)\\
\therefore\ &|Y|\leq\beta|\mathbb{E}_{\hat{\mathcal{D}}^a}\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)|.
\end{aligned}\tag{25}$$

Note that the distance function $\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)=f(\hat{\mathcal{D}}^a)$ is convex in $\hat{\mathcal{D}}^a$, with bounded convexity in the support of $\mathcal{D}^a$ ($|\nabla^2 f''|\leq\hat{\beta}$, with some constant $\hat{\beta}$). Also, we denote the variance of $\hat{\mathcal{D}}^a$ as $V(\hat{\mathcal{D}}^a)$ with an upper bound of σ^2. Then we have,

$$\begin{aligned}
&\mathbb{E}_{\hat{\mathcal{D}}^a}\nu(\mathcal{D}^a,\hat{\mathcal{D}}^a)=\mathbb{E}_{\hat{\mathcal{D}}^a}f(\hat{\mathcal{D}}^a)\\
&\leq f(\mathbb{E}(\hat{\mathcal{D}}^a))+\hat{\beta}V(\hat{\mathcal{D}}^a)\leq\nu(\mathcal{D}^a,\underline{\mathcal{D}}^a)+\hat{\beta}\sigma^2
\end{aligned}$$

Hence,

$$|Y|\leq\beta|\nu(\mathcal{D}^a,\underline{\mathcal{D}}^a)+\hat{\beta}\sigma^2|=\beta\nu(\mathcal{D}^a,\underline{\mathcal{D}}^a)+\mu,\tag{26}$$

where, $\beta\hat{\beta}\sigma^2 = \mu$. (26) gives us the bound on Y. Now we proceed to bound $|X|$, since this bound depends on FL-HPO variant used, we consider each variant seperately.

Bounding $|X|$ for LBM: For LBM we know that $\boldsymbol{\lambda}^\dagger = \sum_i \alpha_i \boldsymbol{\lambda}_i^\dagger$, where $\sum_i \alpha_i = 1$ and $\boldsymbol{\lambda}_i^\dagger$ is the best HP setting found by HPO on client i's data. Also, $\tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}_i^a) = \tilde{\ell}(A_{\boldsymbol{\lambda}_i^*(\mathcal{D}_i^a)}, \mathcal{D}_i^a)$, where $\boldsymbol{\lambda}_i^*(\mathcal{D}_i^a)$ is the HP setting optimized on $\mathcal{D}_i^a$ data of client i. Then we can write,

$$\begin{aligned}
&\ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) = \sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}_i^a) \\
=&\sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}_i^*(\mathcal{D}_i^a)}, \mathcal{D}_i^a) \leq \sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_i^a) \\
=&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a).
\end{aligned} \tag{27}$$

Hence, we have proved that $\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) \geq \ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a)$. Now, under the reasonable assumption that $\tilde{\ell}(A_{\boldsymbol{\lambda}}, \mathcal{D}^a)$ is convex in $\boldsymbol{\lambda}$, we can write,

$$\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) \leq \tilde{\ell}(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) \leq \sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}^a). \tag{28}$$

Also, since $\ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) = \sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}_i^a)$, using (28), we can write,

$$\begin{aligned}
&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) - \ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) \\
&\leq \sum_i \alpha_i \left(\tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}^a) - \tilde{\ell}(A_{\boldsymbol{\lambda}_i^\dagger}, \mathcal{D}_i^a) \right) \\
&\leq \beta' \sum_i \alpha_i \nu(\mathcal{D}_i^a, \mathcal{D}^a).
\end{aligned} \tag{29}$$

Equation (29) upper bounds the value of $\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) - \ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a)$, which is also shown to have lower bound of 0 (from (27)). Hence, $|X| \leq \beta' \sum_i \alpha_i \nu(\mathcal{D}_i^a, \mathcal{D}^a)$, and combining this with (26), we get the bound for the LBM case in Theorem 4.

Bounding $|X|$ for $LKBM$: For $LKBM$ we know the HP setting chosen by FLASH is $\boldsymbol{\lambda}^\dagger = \boldsymbol{\lambda}_{i'}^*(\mathcal{D}_i^a)$, where $\boldsymbol{\lambda}_{i'}$ is the HP setting found by performing HPO at client i that gives the lowest average loss over all clients. Since there is a re-validation process in $LKBM$, hence we can directly say that $X = \ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) - \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) \geq 0$, which means that at some iteration the loss calculated by $LKBM$ cannot be less than the optimized loss calculation overall (which could have happened in LBM).

Now for any client $i \neq i'$ we have,

$$\begin{aligned}
&\tilde{\ell}(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}_i^a) \\
\leq&\tilde{\ell}(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}_{i'}^a) + \beta' \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a) \ \text{[using (21)]} \\
\leq&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) + \beta' \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a) \\
\leq&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) + \beta' \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a).
\end{aligned}$$

Hence,

$$\begin{aligned}
\ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) &= \sum_i \alpha_i \tilde{\ell}(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) \\
&\leq \sum_i \alpha_i \left(\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) + \beta \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a) \right) \\
&= \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) + \beta \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a).
\end{aligned} \tag{30}$$

The last line comes from the fact that $\sum_i \alpha_i = 1$. On the other hand, using (21) again we have,

$$\begin{aligned}
&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a) \\
\geq&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) - \beta' \nu(\mathcal{D}^a, \mathcal{D}_{i'}^a) \\
\geq&\tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}_{i'}^a) - \beta' \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a).
\end{aligned} \tag{31}$$

Combining (30) and (31) we get,

$$\begin{aligned}&\ell(A_{\boldsymbol{\lambda}^\dagger}, \mathcal{D}^a) - \tilde{\ell}(A_{\boldsymbol{\lambda}^*(\mathcal{D}^a)}, \mathcal{D}^a)\\ &= X \leq 2\beta' \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a).\end{aligned}$$

Combining this with the fact that $X \geq 0$ (which we already argued), we have $|X| \leq 2\beta' \max_{i \neq i'} \nu(\mathcal{D}_i^a, \mathcal{D}_{i'}^a)$. Then adding this bound of $|X|$ with (26), we get the bound for the $LKBM$ case in Theorem 4.

Bounding $|X|$ for RM: RM is very similar to FLoRA which is analyzed in Zhou et al. [2022]. The same analysis as in Theorem 4.5 of Zhou et al. [2022] applied to our case gives,

$$|X| \leq \beta_3 \sum_i \alpha_i \left[\nu(\mathcal{D}_i^a, \mathcal{D}^a) + \gamma \min_{k \in [K]} d_j(\boldsymbol{\lambda}, \boldsymbol{\lambda}_k) \right], \tag{32}$$

for algorithm $A^{(j)}$. Combining this with (26), we get the bound for the RM case in Theorem 4. □

Note that in Theorem 4, σ is taken to be the max of the upper bounds $\sigma(a)$ over different values of the fraction a used by FLASH, namely $a \in [a_0, \cdots, a_m]$. When the dataset $\mathcal{D}$ is sufficiently large, then for any $a \in [a_0, \cdots, a_m]$, the datasets $\mathcal{D}^a$ can be assumed to have a distribution that is similar (very close) to that of $\mathcal{D}$. In that case, $\mathcal{D}^a$ and $\underline{\mathcal{D}}^a$ in the bound can both be replaced by (closely approximated by) by $\mathcal{D}$. This implies $\nu(\mathcal{D}^a, \underline{\mathcal{D}}^a) \approx 0$. Therefore, for large datasets $\mathcal{D}$, the loss calculation error bound σ is well approximated as

$$\hat{\sigma} = \mu + \begin{cases} \beta_1 \sum_i \alpha_i \, \nu(\mathcal{D}_i, \mathcal{D}) & (LBM) \\ \beta_2 \max_{i,i'} \nu(\mathcal{D}_i, \mathcal{D}_{i'}) & (LKBM) \\ \beta_3 \sum_i \alpha_i \, \nu(\mathcal{D}_i, \mathcal{D}) + \gamma \bar{D} & (RM) \end{cases}$$

which is only in terms of the full training dataset $\mathcal{D}$.

References

Yi Zhou, Parikshit Ram, Theodoros Salonidis, Nathalie Baracaldo, Horst Samulowitz, and Heiko Ludwig. Single-shot hyper-parameter optimization for federated learning: A general algorithm and analysis. *arXiv preprint arXiv:2202.08338*, 2022. doi: 10.48550/ARXIV.2202.08338. URL `https://arxiv.org/abs/2202.08338`.