Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints - Supplementary Material -

Felix Sattler, Klaus-Robert Müller*, *Member, IEEE*, and Wojciech Samek*, *Member, IEEE*

A1. Supplement

A. Proving the Separation Theorem

The Separation Theorem makes a statement about the cosine similarities between the gradients of the empirical risk functions $\nabla_\theta r_i(\theta^*)$ and $\nabla_\theta r_j(\theta^*)$, which are noisy approximations of the true risk gradients $\nabla_\theta R_{I(i)}(\theta^*)$, respective $\nabla_\theta R_{I(j)}(\theta^*)$. To simplify the notation let us first re-define

$$v_l = \nabla_\theta R_l(\theta^*), l = 1, .., k \tag{1}$$

and

$$X_i = \nabla_\theta r_i(\theta^*) - \nabla_\theta R_{I(i)}(\theta^*), i = 1, .., m \tag{2}$$

Figure A1 shows a possible configuration in $d = 2$ with $k = 3$ different data generating distributions and their corresponding gradients v_1, v_2 and v_3. The empirical risk gradients $X_i + v_{i(i)}$, $i = 1, .., m$ are shown as dashed lines. The maximum angles between gradients from the same data generating distribution are shown green, blue and purple in the plot. Among these, the green angle is the largest one $\sphericalangle_{intra}^{max}$. The plot also shows the optimal bi-partitioning into clusters 1 and 2 and the minimum angle between the gradient updates from any two clients in different clusters $\sphericalangle_{cross}^{min}$ is displayed in red. As long as

$$\sphericalangle_{intra}^{max} < \sphericalangle_{cross}^{min} \tag{3}$$

or equivalently

$$\alpha_{intra}^{min} = \cos(\sphericalangle_{intra}^{max}) > \cos(\sphericalangle_{cross}^{min}) = \alpha_{cross}^{max} \tag{4}$$

the clustering will always be correct.

The proof of the Theorem can be organized into three separate steps:

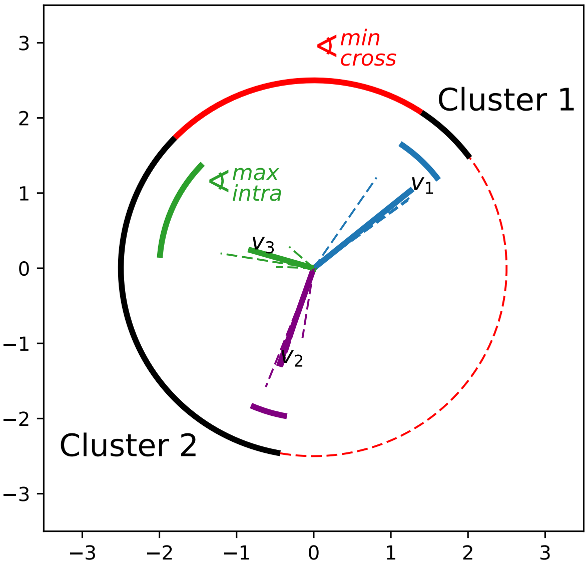

Fig. A1: Possible configuration in $d = 2$ with $k = 3$ different data generating distributions and their corresponding gradients v_1, v_2 and v_3. The empirical risk gradients $X_i + v_{i(i)}$, $i = 1, .., m$ are shown as dashed lines. The maximum angles between gradients from the same data generating distribution are shown green, blue and purple in the plot. Among these, the green angle is the largest one $\sphericalangle_{intra}^{max}$. The vectors are optimally bi-partitioned into clusters 1 and 2 and the minimum angle between the gradient updates from any two clients in different clusters $\sphericalangle_{cross}^{min}$ is displayed in red.

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center (01IS14013A) and the Berlin Center for Machine Learning (01IS18037I). Partial funding by DFG is acknowledged (EXC 2046/1, project-ID: 390685689). This work was also supported by the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451).

F. Sattler and W. Samek are with Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (e-mail: wojciech.samek@hhi.fraunhofer.de).

K.-R. Müller is with the Technische Universität Berlin, 10587 Berlin, Germany, with the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea (e-mail: klaus-robert.mueller@tu-berlin.de).

- In Lemma A1.1, we bound the cosine similarity between two noisy approximations of the same vector α_{intra}^{min} from below
- In Lemma A1.2, we bound the cosine similarity between two noisy approximations of two different vectors from above
- In Lemma A1.3, we show that every set of vectors that sums to zero can be separated into two groups such that the cosine similarity between any two vectors from separate groups can be bounded from above
- Lemma A1.2 and A1.3 together will allow us to bound the cross cluster similarity α_{cross}^{max} from above

Lemma A1.1. *Let $v, X, Y \in \mathbb{R}^d$ with $\|X\| < \|v\|$ and $\|Y\| < \|v\|$ then*

$$\alpha(v+X, v+Y) \geq -\frac{\|X\|\|Y\|}{\|v\|^2} + \sqrt{1-\frac{\|X\|^2}{\|v\|^2}}\sqrt{1-\frac{\|Y\|^2}{\|v\|^2}}. \tag{5}$$

Proof: We are interested in vectors X and Y which maximize the angle between $v+X$ and $v+Y$. Since

$$\alpha(v+X, v+Y) = \cos(\sphericalangle(v+X, v+Y)) \tag{6}$$

and cos is monotonically decreasing on $[0, \pi]$ such X and Y will minimize the cosine similarity α. As $\|X\| < \|v\|$ and $\|Y\| < \|v\|$ the angle will be maximized if and only if v, X and Y share a common 2-dimensional hyperplane, X is perpendicular to $v+X$ and Y is perpendicular to $v+Y$ and X and Y point into opposite directions (Figure A2). It then holds by the trigonometric property of the sine that

$$\sin(\sphericalangle(v, v+X)) = \frac{\|X\|}{\|v\|} \tag{7}$$

and

$$\sin(\sphericalangle(v, v+Y)) = \frac{\|Y\|}{\|v\|} \tag{8}$$

and hence

$$\begin{aligned}\cos(\sphericalangle(v+X, v+Y)) &= \cos(\sphericalangle(v+X) + \sphericalangle(v+Y)) &(9)\\ &\geq \cos(\sin^{-1}(\frac{\|X\|}{\|v\|}) + \sin^{-1}(\frac{\|Y\|}{\|v\|})). &(10)\end{aligned}$$

Since

$$\cos(\sin^{-1}(x) + \sin^{-1}(y)) = -xy + \sqrt{1-x^2}\sqrt{1-y^2} \tag{11}$$

the result follows after re-arranging terms.

■

Lemma A1.2. *Let $v, w, X, Y \in \mathbb{R}^d$ with $\|X\| < \|v\|$, $\|Y\| < \|w\|$ and define*

$$h(v, w, X, Y) := -\frac{\|X\|\|Y\|}{\|v\|^2} + \sqrt{1-\frac{\|X\|^2}{\|v\|^2}}\sqrt{1-\frac{\|Y\|^2}{\|v\|^2}} \tag{12}$$

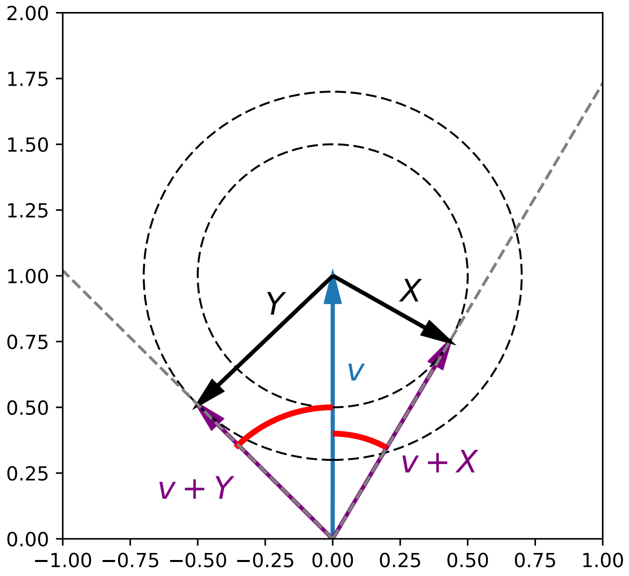

Fig. A2: We are interested in a configuration for which the angle between $v+X$ and $v+Y$ is maximized (red in the plot). As $\|X\| < \|v\|$ and $\|Y\| < \|v\|$ this is exactly the case if the line $\{\beta(v+X)|\beta \in \mathbb{R}\}$ is tangential to the circle with center v and radius $\|X\|$ and the line $\{\beta(v+Y)|\beta \in \mathbb{R}\}$ is tangential to the circle with center v and radius $\|Y\|$.

If

$$\frac{\langle v, w\rangle}{\|v\|\|w\|} \leq h(v, w, X, Y) \tag{13}$$

then it holds

$$\begin{aligned}\alpha(v+X, w+Y) \leq &\alpha(v, w)h(v, w, X, Y) &(14)\\ &+ \sqrt{1-\alpha(v, w)^2}\sqrt{1-h(v, w, X, Y)^2} &(15)\end{aligned}$$

Proof: Analogously to the argument in Figure A2, the angle between $v+X$ and $w+Y$ is minimized, when v, w, X and Y share a common 2-dimensional hyperplane, X is orthogonal to $v+X$, Y is orthogonal to $w+Y$, and X and Y point towards each other. The minimum possible angle is then given by

$$\sphericalangle(v+X,w+Y)=\sphericalangle(v,w)-\sphericalangle(v,v+X)-\sphericalangle(w,w+Y) \tag{16}$$

$$\geq \max(0, \tag{17}$$

$$\cos^{-1}(\frac{\langle v,w\rangle}{\|v\|\|w\|}) \tag{18}$$

$$-\sin^{-1}(\frac{\|X\|}{\|v\|})+ \tag{19}$$

$$-\sin^{-1}(\frac{\|Y\|}{\|v\|})) \tag{20}$$

which can be simplified to

$$\sphericalangle(v+X,w+Y)\geq \max(0,\cos^{-1}(\frac{\langle v,w\rangle}{\|v\|\|w\|}) \tag{21}$$

$$-\cos^{-1}(-\frac{\|X\|\|Y\|}{\|v\|^2}+\sqrt{1-\frac{\|X\|^2}{\|v\|^2}}\sqrt{1-\frac{\|Y\|^2}{\|v\|^2}})) \tag{22}$$

Under condition (13) then second term in the maximum is greater than zero and we get

$$\cos(\sphericalangle(v+X,v+Y)) \tag{23}$$

$$\leq \cos(\cos^{-1}(\frac{\langle v,w\rangle}{\|v\|\|w\|}) \tag{24}$$

$$-\cos^{-1}(-\frac{\|X\|\|Y\|}{\|v\|^2}+\sqrt{1-\frac{\|X\|^2}{\|v\|^2}}\sqrt{1-\frac{\|Y\|^2}{\|v\|^2}})) \tag{25}$$

$$\leq \cos(\cos^{-1}(\alpha(v,w))-\cos^{-1}(h(v,w,X,Y))) \tag{26}$$

Since

$$\cos(\cos^{-1}(x)-\cos^{-1}(y))=xy+\sqrt{1-x^2}\sqrt{1-y^2} \tag{27}$$

the result follows after re-arranging terms. ■

Lemma A1.3. *Let* $v_1,..,v_k\in\mathbb{R}^d$*,* $d\geq 2$*,* $\gamma_1,..,\gamma_k\in\mathbb{R}_{>0}$ *and*

$$\sum_{i=1}^{k}\gamma_i v_i=0\in\mathbb{R}^d \tag{28}$$

then there exists a bi-partitioning of the vectors $c_1\cup c_2=\{1,..,k\}$ *such that*

$$\max_{i\in c_1,j\in c_2}\alpha(v_i,v_j)\leq\cos(\frac{\pi}{k-1}) \tag{29}$$

Proof:

Lemma A1.3 can be equivalently stated as follows: Let $v_1,..,v_k\in\mathbb{R}^d$, $d\geq 2$, $\gamma_1,..,\gamma_k\in\mathbb{R}_{>0}$ and

$$\sum_{i=1}^{k}\gamma_i v_i=0\in\mathbb{R}^d \tag{30}$$

then there exists a bi-partitioning of the vectors $c_1\cup c_2=\{1,..,k\}$ such that

$$\min_{i\in c_1,j\in c_2}\sphericalangle(v_i,v_j)\geq\frac{\pi}{k-1} \tag{31}$$

As the angle between two vectors is invariant under multiplication with positive scalars $\gamma>0$ we can assume w.l.o.g that $\gamma_i=1$ $i=1,..,k$.

Let us first consider the case where $d=2$. Let $e_1\in\mathbb{R}^2$ be the first standard basis vector and assume w.l.o.g that the vectors $v_1,..,v_k$ are sorted w.r.t. their angular distance to e_1 (they are arranged circular as shows in Figure A3). As all vectors lie in the 2d plane, we know that the sum of the angles between all neighboring vectors has to be equal to 2π.

$$\sum_{i=1}^{k}\sphericalangle(v_i,v_{(i+1)\text{ mod }k})=2\pi \tag{32}$$

Now let

$$i_1^*=\arg\max_{i\in\{1,..,k\}}\sphericalangle(v_i,v_{(i+1)\text{ mod }k}) \tag{33}$$

and

$$i_2^*=\arg\max_{i\in\{1,..,k\}\backslash i_1^*}\sphericalangle(v_i,v_{(i+1)\text{ mod }k}) \tag{34}$$

be the indices of the largest and second largest angles between neighboring vectors and define the following clusters:

$$c_1=\{i\text{ mod }k|i_1^*<i\leq i_2^*+k[i_2^*<i_1^*]\} \tag{35}$$

$$c_2=\{i\text{ mod }k|i_2^*<i\leq i_1^*+k[i_2^*>i_1^*]\}\} \tag{36}$$

where $[x]=1$ if x is true and $[x]=0$ is x is false. Then by construction the second largest angle $\sphericalangle(v_{i_2^*},v_{(i_2^*+1)\text{ mod }k})$ minimizes the angle between any two vectors from the two different clusters c_1, c_2 (see Figure A3 for an illustration):

$$\min_{i\in c_1,j\in c_2}\sphericalangle(v_i,v_j)=\sphericalangle(v_{i_2^*},v_{(i_2^*+1)\text{ mod }k}) \tag{37}$$

Hence in $d=2$ we can always find a partitioning c_1,c_2 s.t. the minimum angle between any two vectors from different clusters is greater or equal to the 2nd largest angle between neighboring vectors. This means the worst case configuration of vectors is one where the 2nd largest angle between neighboring vectors is minimized. As the sum of all k angles between neighboring vectors is constant according to (32), this is exactly the case when the largest angle between neighboring vectors is maximized and all other $k-1$ angles are equal.

Assume now that the angle between two neighboring vectors is greater than π. That would mean that there exists a separating line l which passes through the origin and all vectors $v_1,..,v_k$ lie on one side of that line. This however is impossible since $\sum_{l=1}^{k}v_l=0$. This means that the largest angle between neighboring vectors can not be greater than π. Hence in the worst-case scenario

$$\sphericalangle(v_{i_2^*},v_{(i_2^*+1)\text{ mod }k})\geq\frac{2\pi-\pi}{k-1}=\frac{\pi}{k-1}. \tag{38}$$

This concludes the proof for $d=2$.

Now consider he case where $d > 2$. Let c_1, c_2 be a clustering which maximizes the minimum angular distance between any two clients from different clusters. Let

$$i^*, j^* = \arg \min_{i \in c_1, j \in c_2} \sphericalangle(v_i, v_j) \tag{39}$$

then v_{i^*} and v_{j^*} are the two vectors with minimal angular distance. Let $A = [v_{i^*}, v_{j^*}] \in \mathbb{R}^{d,2}$ and consider now the projection matrix

$$P = A(A^T A)^{-1} A^T \tag{40}$$

which projects all d-dimensional vectors onto the plane spanned by v_{i^*} and v_{j^*}. Then be linearity of the projection we have

$$0 = P0 = P(\sum_{i=1}^k v_i) = \sum_{i=1}^k P(v_i) \tag{41}$$

Hence the projected vectors also satisfy the condition of the Lemma. As

$$\sphericalangle(Pv_{i^*}, Pv_{j^*}) = \sphericalangle(v_{i^*}, v_{j^*}) \tag{42}$$

and

$$\sphericalangle(Pv_i, Pv_j) \geq \sphericalangle(v_i, v_j) \tag{43}$$

for all $i, j \notin \{i^*, j^*\}$ the clustering c_1, c_2 is still optimal after projecting and we have found a 2d configuration of vectors satisfying the assumptions of Lemma A1.3 with the same minimal cross-cluster angle. In other words, we have reduced the $d > 2$ case to the $d = 2$ case, for which we have already proven the result. This concludes the proof.

■

Theorem A1.4 (Separation Theorem)**.** *Let $D_1, .., D_m$ be the local training data of m different clients, each dataset sampled from one of k different data generating distributions $\varphi_1, .., \varphi_k$, such that $D_i \sim \varphi_{I(i)}(x, y)$. Let the empirical risk on every client approximate the true risk at every stationary solution of the Federated Learning objective θ^* s.t.*

$$\|\nabla R_{I(i)}(\theta^*)\| > \|\nabla R_{I(i)}(\theta^*) - \nabla r_i(\theta^*)\| \tag{44}$$

and define

$$\gamma_i := \frac{\|\nabla R_{I(i)}(\theta^*) - \nabla r_i(\theta^*)\|}{\|\nabla R_{I(i)}(\theta^*)\|} \in [0, 1) \tag{45}$$

Then there exists a bi-partitioning $c_1^* \cup c_2^* = \{1, .., m\}$ of the client population such that that the maximum similarity between the updates from any two clients from different clusters can be bounded from above according to

$$\alpha_{cross}^{max} := \min_{c_1 \cup c_2 = \{1,..,m\}} \max_{i \in c_1, j \in c_2} \alpha(\nabla r_i(\theta^*), \nabla r_j(\theta^*)) \tag{46}$$

$$= \max_{i \in c_1^*, j \in c_2^*} \alpha(\nabla r_i(\theta^*), \nabla r_j(\theta^*)) \tag{47}$$

$$\leq \begin{cases} \cos(\frac{\pi}{k-1}) H_{i,j} + \sin(\frac{\pi}{k-1}) \sqrt{1 - H_{i,j}^2} & \text{if } H \geq \cos(\frac{\pi}{k-1}) \\ 1 & \text{else} \end{cases} \tag{48}$$

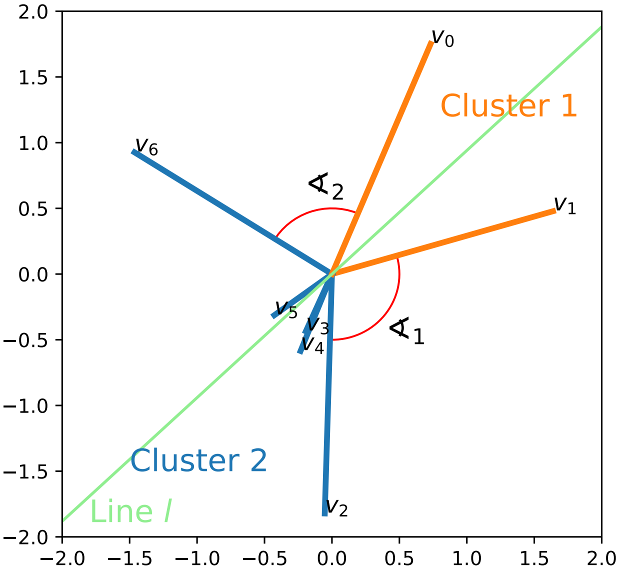

Fig. A3: Possible configuration in $d = 2$. The largest and 2nd largest angle between neighboring vectors (red) separate the two optimal clusters. The largest angle between neighboring vectors is never greater than π.

with

$$H_{i,j} = -\gamma_i \gamma_j + \sqrt{1 - \gamma_i^2} \sqrt{1 - \gamma_j^2} \in (-1, 1]. \tag{49}$$

At the same time the similarity between updates from clients which share the same data generating distribution can be bounded from below by

$$\alpha_{intra}^{min} := \min_{\substack{i,j \\ I(i)=I(j)}} \alpha(\nabla_\theta r_i(\theta^*), \nabla_\theta r_j(\theta^*)) \geq \min_{\substack{i,j \\ I(i)=I(j)}} H_{i,j}. \tag{50}$$

Proof: For the first result, we know that in every stationary solution of the Federated Learning objective θ^* it holds

$$\sum_{l=1}^k \gamma_i \nabla_\theta R_l(\theta^*) = 0 \tag{51}$$

and hence by Lemma A1.3 there exists a bi-partitioning $\hat{c}_1 \cup \hat{c}_2 = \{1, .., k\}$ such that

$$\max_{l \in \hat{c}_1, j \in \hat{c}_2} \alpha(\nabla_\theta R_l(\theta^*), \nabla_\theta R_j(\theta^*)) \leq \cos(\frac{\pi}{k-1}) \tag{52}$$

Let

$$c_1 = \{i | I(i) \in \hat{c}_1, i = 1, .., m\} \tag{53}$$

and

$$c_2 = \{i | I(i) \in \hat{c}_2, i = 1, .., m\} \tag{54}$$

and set for some $i \in c_1$ and $j \in c_2$:

$$v = \nabla_\theta R_{I(i)}(\theta^*) \tag{55}$$

$$X = \nabla_\theta r_i(\theta^*) - \nabla_\theta R_{I(i)}(\theta^*) \tag{56}$$

$$w = \nabla_\theta R_{I(j)}(\theta^*) \tag{57}$$

$$Y = \nabla_\theta r_j(\theta^*) - \nabla_\theta R_{I(j)}(\theta^*) \tag{58}$$

Then $\alpha(v, w) \leq \cos(\frac{\pi}{k-1})$ and the result follows directly from Lemma A1.2.

The second result follows directly from Lemma A1.1 by setting

$$v = \nabla_\theta R_{I(i)}(\theta^*) \tag{59}$$

$$X = \nabla_\theta r_i(\theta^*) - \nabla_\theta R_{I(i)}(\theta^*) \tag{60}$$

$$Y = \nabla_\theta r_j(\theta^*) - \nabla_\theta R_{I(i)}(\theta^*) \tag{61}$$

■