Supplementary Material for A Memory-Efficient Federated Kernel Support Vector Machine for Edge Devices

Xiaochen Zhou and Xudong Wang, *Fellow, IEEE*

I. Proof of Theorem 1

Let $P_{m,l}(\mathbf{w}_m[l])$ denote the objective function of problem Ω_l, i.e.,

$$P_{m,l}(\mathbf{w}_m[l]) = C\sum_{i=1}^{N_m}\max\{0, 1-B_i-y_i\mathbf{w}_m[l]^\top\mathbf{a}(\mathbf{x}_i)[l]\} + \frac{1}{2}||\mathbf{w}_m[l]||^2$$

In the q-th iteration of block boosting, block $\mathbf{w}_m[l]$ of the local parameter vector $\mathbf{w}_m = [\mathbf{w}_m[1]^\top, ..., \mathbf{w}_m[L]^\top]^\top$ is optimized by first solving the dual problem of problem Ω_l:

$$\max_{\boldsymbol{\alpha}_{m,l}}\left\{-\frac{1}{2}||\mathbf{A}_m[l]\boldsymbol{\alpha}_{m,l}||^2 + \sum_{i}^{N_m}(1-B_i(q))\alpha_{m,l,i}\right\}$$
$$\text{s.t.}\quad 0\le\alpha_{m,l,i}\le C, i=1,2,...,N_m.$$

As the optimal solution $\boldsymbol{\alpha}^*_{m,l}(q)$ to the dual problem is obtained, it can be transformed to the optimal solution $\mathbf{w}_m[l](q)$ to problem Ω_l via $\mathbf{w}_m[l](q) = \mathbf{A}_m[l]\boldsymbol{\alpha}^*_{m,l}(q)$. Let $\mathbf{w}_m[l](q-1)$ denote the initial value of $\mathbf{w}_m[l]$ in the q-th iteration, then we have

$$P_{m,l}(\mathbf{w}_m[l](q-1)) \ge P_{m,l}(\mathbf{w}_m[l](q)). \tag{1}$$

Let

$$\begin{aligned}&\mathbf{w}_m(q-1)\\ =&[\mathbf{w}_m[1](q-1)^\top, ..., \mathbf{w}_m[l](q-1)^\top, ..., \mathbf{w}_m[L](q-1)^\top]^\top\\ &\mathbf{w}_m(q)\\ =&[\mathbf{w}_m[1](q-1)^\top, ..., \mathbf{w}_m[l](q)^\top, ..., \mathbf{w}_m[L](q-1)^\top]^\top,\end{aligned}$$

then we have

$$\begin{aligned}&P_m(\mathbf{w}_m(q-1)) - P_m(\mathbf{w}_m(q))\\ =&P_{m,l}(\mathbf{w}_m[l](q-1)) - P_{m,l}(\mathbf{w}_m[l](q)).\end{aligned}$$

Based on equation (1), we have

$$P_m(\mathbf{w}_m(q-1)) \ge P_m(\mathbf{w}_m(q)). \tag{2}$$

Equation (2) indicates that the sequence of local training loss $(P_m(\mathbf{w}_m(0), ..., P_m(\mathbf{w}_m(q), ...)$ is non-increasing when the local parameter vector is optimized by block boosting. Since $P_m(\cdot)$ is a strongly convex function, if the training loss cannot be further reduced, then the optimal local parameter vector $\mathbf{w}^*_m$ is obtained.

The authors are with UM-SJTU Joint Institute, Shanghai Jiao Tong University. Corresponding author: Xudong Wang, Email: wxudong@ieee.org

II. Proof of Theorem 2

In the $(t+1)$-th iteration of Fed-KSVM, edge device m initially holds a local parameter vector $\mathbf{w}_m(t) = \mathbf{A}_m\boldsymbol{\alpha}_m(t)$, and it employs block boosting to obtain the optimal solution $\mathbf{w}^*_m(t+1)$ to

$$\min_{\mathbf{w}_m} P_m(\mathbf{w}_m; \bar{\mathbf{w}}_m(t+1)) := C\sum_{i\in\mathcal{I}_m}\max\{0, 1-y_i\hat{f}_m(\mathbf{x}_i)\} + \frac{1}{2}||\mathbf{w}_m||^2,$$

where $\hat{f}_m(\mathbf{x}_i) = \frac{1}{\sqrt{D}}(\bar{\mathbf{w}}_m(t+1)+\mathbf{w}_m)^\top\mathbf{a}(\mathbf{x}_i)$. Based on the duality, $\mathbf{w}^*_m(t+1)$ can be also expressed by

$$\mathbf{w}^*_m(t+1) = \mathbf{A}_m\boldsymbol{\alpha}^*_m(t+1),$$

where $\boldsymbol{\alpha}^*_m(t+1)$ is the optimal solution to

$$\max_{\boldsymbol{\alpha}_m} D_m(\boldsymbol{\alpha}_m; \bar{\mathbf{w}}_m(t+1)) := -\frac{1}{2}||\bar{\mathbf{w}}_m(t+1)+\mathbf{A}_m\boldsymbol{\alpha}_m||^2 + \sum_{i\in\mathcal{I}_m}\alpha_i + \frac{1}{2}||\bar{\mathbf{w}}_m(t+1)||^2$$
$$\text{s.t.}\quad 0\le\alpha_i\le C, i\in\mathcal{I}_m.$$

By applying Lemma 1 in [1] to the dual objective function $D_m(\boldsymbol{\alpha}_m; \bar{\mathbf{w}}_m(t+1))$, we have

$$\begin{aligned}&\mathbb{E}[D_m(\boldsymbol{\alpha}^*_m(t+1); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1))]\\ \ge&\frac{s_m}{N}\mathbb{E}[P_m(\mathbf{w}_m(t); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1))],\end{aligned} \tag{3}$$

where $s_m = \min_{\mathbf{w}}\frac{\sum_{i\in\mathcal{I}_m}|\frac{1}{\sqrt{D}}\mathbf{w}^\top a(\mathbf{x}_i)-y_i|}{\sum_{i\in\mathcal{I}_m}(\frac{1}{D}||a(\mathbf{x}_i)||^2+|\frac{1}{\sqrt{D}}\mathbf{w}^\top a(\mathbf{x}_i)-y_i|)}$.

Note that

$$
\begin{aligned}
&\sum_{m=1}^{M} P_m(\mathbf{w}_m(t); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1)) \\
=& C\sum_{i=1}^{N} \max\{0, 1 - y_i \hat{f}(\mathbf{x}_i)\} - \sum_{i}^{N} \alpha_i(t) \\
&+ \sum_{m=1}^{M} (\frac{1}{2}||\mathbf{w}_m(t)||^2 + \frac{1}{2}||\bar{\mathbf{w}}_m(t+1) + \mathbf{A}_m\boldsymbol{\alpha}_m(t)||^2 \\
&- \frac{1}{2}||\bar{\mathbf{w}}_m(t+1)||^2) \\
=& C\sum_{i=1}^{N} \max\{0, 1 - y_i \hat{f}(\mathbf{x}_i)\} - \sum_{i}^{N} \alpha_i(t) \\
&+ \sum_{m=1}^{M} (\frac{1}{2}||\mathbf{w}_m(t)||^2 + \frac{1}{2}||\mathbf{w}(t)||^2 - \frac{1}{2}||\bar{\mathbf{w}}_m(t+1)||^2) \\
=& C\sum_{i=1}^{N} \max\{0, 1 - y_i \hat{f}(\mathbf{x}_i)\} - \sum_{i}^{N} \alpha_i(t) + ||\mathbf{w}(t)||^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&P(\mathbf{w}(t)) - D(\boldsymbol{\alpha}(t)) \\
=& C\sum_{i=1}^{N} \max\{0, 1 - y_i \hat{f}(\mathbf{x}_i)\} + \frac{1}{2}||\mathbf{w}(t)||^2 \\
&- (-\frac{1}{2}||\mathbf{A}\boldsymbol{\alpha}(t)||^2 + \sum_{i}^{N} \alpha_i(t)) \\
=& C\sum_{i=1}^{N} \max\{0, 1 - y_i \hat{f}(\mathbf{x}_i)\} - \sum_{i}^{N} \alpha_i(t) + ||\mathbf{w}(t)||^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{m=1}^{M} P_m(\mathbf{w}_m(t); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1)) \\
=& P(\mathbf{w}(t)) - D(\boldsymbol{\alpha}(t)).
\end{aligned}
$$

By summing up equation 3 from 1 to M, we have

$$
\begin{aligned}
&\sum_{m=1}^{M} \mathbb{E}[D_m(\boldsymbol{\alpha}^*_m(t+1); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1))] \\
\geq& \sum_{m=1}^{M} \frac{s_m}{N} \mathbb{E}[P_m(\mathbf{w}_m(t+1); \bar{\mathbf{w}}_m(t)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1))] \\
\geq& \frac{s}{N} \mathbb{E}[P(\mathbf{w}(t)) - D(\boldsymbol{\alpha}(t))],
\end{aligned}
$$

where $s = \min_m s_m$.

By using Jensen inequality, we have

$$
\begin{aligned}
&\frac{1}{M}\mathbb{E}[D_m(\boldsymbol{\alpha}^*_m(t+1); \bar{\mathbf{w}}_m(t+1)) - D_m(\boldsymbol{\alpha}_m(t); \bar{\mathbf{w}}_m(t+1))] \\
=& \frac{1}{M}\mathbb{E}[D([\boldsymbol{\alpha}_1(t), ..., \boldsymbol{\alpha}^*_m(t+1), ..., \boldsymbol{\alpha}_M(t)])] \\
&- \frac{1}{M}\mathbb{E}[D([\boldsymbol{\alpha}_1(t), ..., \boldsymbol{\alpha}_m(t), ..., \boldsymbol{\alpha}_M(t)])] \\
=& \frac{1}{M}\mathbb{E}[D([\boldsymbol{\alpha}_1(t), ..., \boldsymbol{\alpha}_m(t) + \Delta\boldsymbol{\alpha}_m(t), ..., \boldsymbol{\alpha}_M(t)])] \\
&- \frac{1}{M}\mathbb{E}[D([\boldsymbol{\alpha}_1(t), ..., \boldsymbol{\alpha}_m(t), ..., \boldsymbol{\alpha}_M(t)])] \\
\leq& \mathbb{E}[D([\boldsymbol{\alpha}_1(t) + \frac{\Delta\boldsymbol{\alpha}_1(t)}{M}, ..., \boldsymbol{\alpha}_M(t) + \frac{\Delta\boldsymbol{\alpha}_M(t)}{M}])] \\
&- \mathbb{E}[D([\boldsymbol{\alpha}_1(t), ..., \boldsymbol{\alpha}_m(t), ..., \boldsymbol{\alpha}_M(t)])] \\
=& \mathbb{E}[D(\boldsymbol{\alpha}(t+1)) - D(\boldsymbol{\alpha}(t))] \\
\leq& \mathbb{E}[D(\boldsymbol{\alpha}^*) - D(\boldsymbol{\alpha}(t))].
\end{aligned}
$$

Hence we have

$$
\mathbb{E}[D(\boldsymbol{\alpha}^*) - D(\boldsymbol{\alpha}(t))] \geq \frac{s}{NM}\mathbb{E}[P(\mathbf{w}(t)) - D(\boldsymbol{\alpha}(t))].
$$

Since $D(\boldsymbol{\alpha}(t)) \leq D(\boldsymbol{\alpha}^*) = P(\mathbf{w}^*)$, then

$$
\mathbb{E}[D(\boldsymbol{\alpha}^*) - D(\boldsymbol{\alpha}(t))] \geq \frac{s}{NM}\mathbb{E}[P(\mathbf{w}(t)) - P(\mathbf{w}^*)]. \quad (4)
$$

Let $\xi = \min_i |\frac{1}{\sqrt{D}}\mathbf{w}^{*\top} a(\mathbf{x}_i) - y_i|$ for all i that satisfy $|\frac{1}{\sqrt{D}}\mathbf{w}^{*\top} a(\mathbf{x}_i) - y_i| > 0$. Then, according to Proposition 1 in [1], we have

$$
\begin{aligned}
&D(\eta\boldsymbol{\alpha}^* + (1-\eta)\boldsymbol{\alpha}) \\
\geq& \eta D(\boldsymbol{\alpha}^*) + (1-\eta)D(\boldsymbol{\alpha}) + \frac{\xi\eta(1-\eta)}{2N}||\boldsymbol{\alpha}^* - \boldsymbol{\alpha}||^2.
\end{aligned}
$$

Thus, the convergence rate of CoCoA (denoted as Θ) is $\Theta = 1 - \frac{\xi}{M(\xi + C\tilde{N})}$ according to [2], where $\tilde{N} = \max_m N_m$. Then we have

$$
\begin{aligned}
&\mathbb{E}[D(\boldsymbol{\alpha}^*) - D(\boldsymbol{\alpha}(t))] \\
\leq& \Theta^t(\mathbb{E}[D(\boldsymbol{\alpha}^*) - D(\boldsymbol{\alpha}(0))]) \\
=& \Theta^t\mathbb{E}[D(\boldsymbol{\alpha}^*)].
\end{aligned} \quad (5)
$$

By combining equation 4 and equation 5, we finally obtain

$$
\begin{aligned}
&\mathbb{E}[P(\mathbf{w}(t)) - P(\mathbf{w}^*)] \\
\leq& \frac{\Theta^t NM}{s}\mathbb{E}[D(\boldsymbol{\alpha}^*)] \\
=& \frac{\Theta^t NM}{s}\mathbb{E}[P(\mathbf{w}^*)].
\end{aligned}
$$

References

[1] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization." *Journal of Machine Learning Research (JMLR)*, vol. 14, no. 2, 2013.

[2] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan, "Communication-efficient distributed dual coordinate ascent," in *Advances in Neural Information Processing Systems (Neurips)*, 2014, pp. 3068–3076.