Supplemental File for "Energy or Accuracy? Near-Optimal User Selection and Aggregator Placement for Federated Learning in MEC"

Zichuan Xu, *Member, IEEE,* Dongrui Li, Weifa Liang, *Senior Member, IEEE,* Wenzheng Xu, *Member, IEEE,* Qiufen Xia, *Member, IEEE,* Pan Zhou, *Member, IEEE,* Omer F. Rana, *Senior Member, IEEE,* and Hao Li

Abstract—This is a supplemental file for the paper entitled "Energy or Accuracy? Near-Optimal User Selection and Aggregator Placement for Federated Learning in MEC", which consists of all the proof bodies of the theorems.

APPENDIX

Proof of Theorem 1

Proof: The probability of selecting a UE is jointly determined by the representative context selection and arm selection. These two steps of **Algorithm** `OL_MAB` are needed to find a bound on the regret $R(T)$. By Eqs. (18), and (20), the losses of UEs and experts depend on the loss of UEs.

Given the probability $p_{k,t}$ of selecting a UE in round t when all experts select their representative contexts, we show that

$$\mathbb{E}_{k\sim p_{k,t}}\tilde{l}_{k,t} = l_{K_t,t}, \tag{1}$$

where K_t is the index of the selected UE. The reason is that each UE only generates a loss when it is selected; otherwise, its loss is zero.

For each group g_w and its selected representative context $\Theta_{w,t}$, we use $\mathcal{UE}_{\Theta_{w,t}}$ to denote the UEs that are selected according to representative context $\Theta_{w,t}$. Notice that $\mathcal{UE}_{\Theta_{w,t}}$ is hard to determine because the UEs are selected according to the expected probabilities of all representative contexts. Instead, it serves as an 'auxiliary' set to derive the following upper bound,

$$\begin{aligned}\mathbb{E}_{\theta\sim p^w_{\theta,t}}\tilde{L}_{w,\theta,t} = L_{w,\theta,t} &= \sum\nolimits_{k\in\mathcal{UE}_{\Theta_{w,t}}}\mathbb{E}_{k\sim p^\theta_{k,t}}\tilde{l}_{k,t}\\ &\le \sum\nolimits_{k\in\mathcal{UE}_{\Theta_{w,t}}} l_{k,t},\end{aligned} \tag{2}$$

and

$$\begin{aligned}\mathbb{E}_{\theta\sim p^w_{\theta,t}}\tilde{L}^2_{w,\theta,t} &= \frac{L^2_{w,\Theta_{w,t},t}}{p^w_{\Theta_{m,t},t}} = \frac{(\sum_{k\in\mathcal{UE}_{\Theta_{w,t}}} l_{k,t})^2}{p^w_{\Theta_{m,t},t}}\\ &\le |\mathcal{UE}|^2/p^w_{\Theta_{m,t},t},\end{aligned} \tag{3}$$

since $l_{k,t}\le 1$ and $\mathcal{UE}_{\Theta_{w,t}}\subset\mathcal{UE}$.

Given the representative context of group g_w, we have

$$\mathbb{E}_{\Theta_{w,t}\sim p^w_{\Theta_{w,t},t}} 1/p^w_{\Theta_{w,t},t} = |g_w| \le |\mathcal{C}|. \tag{4}$$

The above inequalities mean that

$$\begin{aligned}&\sum\nolimits_{t=1}^T L_{w,\Theta_{m,t},t} - \sum\nolimits_{t=1}^T L_{w,\theta,t}\\ &= \sum_{t=1}^T \mathbb{E}_{\theta\sim p^w_{\theta,t}}\tilde{L}_{w,\theta,t} - \sum_{t=1}^T \sum_{K_t\in\mathcal{UE}_{\Theta_{w,t}}}\mathbb{E}_{K_t\sim p_{k,t}}\tilde{L}_{w,\theta,t}.\end{aligned} \tag{5}$$

We can rewrite the first term of Eq. (5) as

$$\begin{aligned}&\sum\nolimits_{t=1}^T \mathbb{E}_{\theta\sim p^w_{\theta,t}}\tilde{L}_{w,\theta,t}\\ &= (1/\eta_t)\ln\mathbb{E}_{\theta\sim p^w_{\theta,t}} exp(-\eta_t(\tilde{L}_{w,\theta,t} - \mathbb{E}_{\theta'\sim p^w_{\theta',t}}\tilde{L}_{w,\theta',t}))\\ &- (1/\eta_t)\ln\mathbb{E}_{\theta\sim p^w_{\theta,t}} exp(-\eta_t\tilde{L}_{w,\theta,t}).\end{aligned} \tag{6}$$

We now show the bound of Eq. (6) by

$$\begin{aligned}&\ln\mathbb{E}_{\theta\sim p^w_{\theta,t}} exp(-\eta_t(\tilde{L}_{w,\theta,t} - \mathbb{E}_{\theta'\sim p^w_{\theta',t}}\tilde{L}_{w,\theta',t}))\\ &= \ln\mathbb{E}_{\theta\sim p^w_{\theta,t}} exp(-\eta_t\tilde{L}_{w,\theta,t} + \eta_t\mathbb{E}_{\theta'\sim p^w_{\theta',t}}\tilde{L}_{w,\theta',t})\\ &\le \mathbb{E}_{\theta\sim p^w_{\theta,t}}(exp(-\eta_t\tilde{L}_{w,\theta,t}) - 1 + \eta_t\tilde{L}_{w,\theta,t})\\ &\qquad \text{due to } \ln x \le x-1.\\ &\le \mathbb{E}_{\theta\sim p^w_{\theta,t}}\eta_t^2\tilde{L}^2_{w,\theta,t}/2,\\ &\quad \text{due to } exp(-x) - 1 + x \le x^2/2 \text{ for all } 0\le x\le 1.\\ &\le (\eta_t^2|\mathcal{UE}|)/(2p^w_{\Theta_{m,t},t}), \quad \text{due to Eq. (3).}\end{aligned} \tag{7}$$

We proceed by showing the upper bound of Eq. (6). To this end, let $\tilde{L}^{cum}_{w,\theta,t}$ be the cumulative estimated loss of context θ in group g_w in round t. Clearly, $\tilde{L}^{cum}_{w,\theta,0} = 0$, $\Phi_0(\eta) = 0$, and $\Phi_t(\eta) = \frac{1}{\eta}\ln\frac{1}{|g_w|}\sum_{\theta=1}^{|g_w|} exp(\eta\tilde{L}^{cum}_{w,\theta,t})$. According to the definition of $p^w_{\theta,t}$ we have

$$\begin{aligned}&-1/\eta_t\ln\mathbb{E}_{\theta\sim p^w_{\theta,t}} exp(-\eta_t\tilde{L}_{w,\theta,t})\\ &= -1/\eta_t\ln\frac{\sum_\theta exp(\eta_t\tilde{L}^{cum}_{w,\theta,t})}{\sum_\theta exp(-\eta_t\tilde{L}^{cum}_{w,\theta,t-1})}\\ &= \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t).\end{aligned} \tag{8}$$

Summing up inequalities of (7) and (8), we have

$$\begin{aligned}(5) &\le \sum\nolimits_t^T \frac{\eta_t^2|\mathcal{UE}|}{2p^w_{\Theta_{m,t},t}} + \sum\nolimits_t^T(\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t))\\ &- \sum\nolimits_{t=1}^T\sum\nolimits_{K_t\in\mathcal{UE}_{\Theta_{w,t}}}\mathbb{E}_{K_t\sim p_{k,t}}\tilde{L}_{w,\theta,t}\end{aligned}$$

Thus

$$\mathbb{E}\Big(\sum_{t}^{T}\frac{\eta_t^2|\mathcal{UE}|}{2p^w_{\Theta_{m,t},t}}\Big)=\mathbb{E}\Big(\sum_{t}^{T}\mathbb{E}_{\Theta_{w,t}\sim p^w_{\theta,t}}\frac{\eta_t^2|\mathcal{UE}|}{2p^w_{\Theta_{m,t},t}}\Big)\le\frac{|\mathcal{C}|}{2}\sum_{t}^{T}\eta_t^2.$$

By applying the Abel transformation [1], we have

$$\sum_{t}^{T}(\Phi_{t-1}(\eta_t)-\Phi_t(\eta_t))=\sum_{t}^{T-1}(\Phi_t(\eta_{t+1})-\Phi_t(\eta_t))-\Phi_T(\eta_T),$$

considering that $\Phi_0(\eta_1)=0$.

By the definition of $\Phi_t(\eta)$, we have

$$\begin{aligned}&-\Phi_T(\eta_T)=(\ln|g_w|)/\eta_T-1/\eta_T\ln(\sum_{\theta'=1}^{|g_w|}exp(-\eta_T\tilde{L}^{cum}_{w,\theta',t}))\\&\le(\ln|g_w|)/\eta_T-1/\eta_T\ln exp(-\eta_T\tilde{L}^{cum}_{w,\theta,t})\\&=(\ln|g_w|)/\eta_T+\sum\nolimits_{t=1}^{T}\tilde{L}_{w,\theta,t}\end{aligned}\tag{9}$$

We thus have

$$\begin{aligned}&\mathbb{E}(\sum\nolimits_{t=1}^{T}L_{w,\Theta_{m,t},t}-\sum\nolimits_{t=1}^{T}L_{w,\theta,t})\\&\le\frac{|\mathcal{C}|}{2}\sum\nolimits_{t}^{T}\eta_t^2+\frac{\ln|g_w|}{\eta_T}+\mathbb{E}(\sum\nolimits_{t}^{T-1}(\Phi_t(\eta_{t+1})-\Phi_t(\eta_t)))\end{aligned}\tag{10}$$

Given the definition of $\Phi_t(\eta_t)$, we can easily obtain that $\Phi_t(\eta_{t+1})-\Phi_t(\eta_t))\le 0$, by the fact that its first order derivative is negative. Thus, for all groups the regret of the algorithm is

$$R(T)\le|\mathcal{C}|/2T\eta^2+\ln|\mathcal{C}|/\eta,\tag{11}$$

When $\eta_1=\eta_2=\cdots=\eta_T=\eta$.

We then analyze the time complexity of algorithm `OL_MAB`. Algorithm `OL_MAB` basically has three stages: (1) context grouping, (2) UE selection, and (3) Loss and probability updating. In context grouping, we have Θ contexts that should be divided into groups, which takes $O(\Theta)$ time. Within each context group, the selection of a representative context also takes $O(\Theta)$ time. In UE selection, each UE is selected according to the probability of the representative context. Since there are $|\mathcal{UE}|$ UEs, a randomly play of all UEs takes $O(|\mathcal{UE}|)$ time. Similarly, assuming that the loss and probability updating happen sequentially, it takes $O(|\mathcal{UE}|)$ time. In total, the total time complexity of `OL_MAB` is $O(T(\Theta+|\mathcal{UE}|))$. □

Proof of Theorem 2

Proof: We first show the solution feasibility of the algorithm, where a solution is feasible as long as the local model of each UE is aggregated and the resource capacity of each location Loc_q is not violated. It is not difficult to verify that FL request r_m can be admitted given the limited resource capacities on base stations and cloudlets, as the available resources of each location are split into multiple resource splits and each resource split has the minimum computing or bandwidth resource for each data unit. Each aggregator assigns the minimum amount of computing and bandwidth resources to aggregate trained models. Also, in the flow f each UE lies in a single path from itself to the virtual sink of G'. Its model thus is sent to the master aggregator for aggregation.

We then analyze the approximation ratio of **Algorithm** `Placement`. The basic idea is to treat each location as a number of resource splits and each resource split is used to process a single trained model w_m from a UE. However, assigning a lower amount of resource to each resource split may slow down the process speed of each aggregator, thereby increasing the energy consumption due to longer processing times. In the worse case, the energy consumption due to the aggregation of each trained model can be pushed up by a factor of ϑ, where $\vartheta=\max\{\frac{\max_q\kappa_q}{\min_q\kappa_q},\frac{\max_q\kappa_q}{\min_q\omega_q}\}$. Let OPT be the optimal energy consumption of the energy minimization problem for FL, and OPT' the optimal solution to the unsplittable minimum-cost multicommodity flow problem in auxiliary graph G'. We have $OPT'\le\vartheta OPT$, since the energy consumption due to aggregations is upper bounded by a factor of at most ϑ.

Using $e(f)$ to denote the energy consumption of solution obtained by **Algorithm** `Placement`, we have the energy consumption $e(f)=OPT'$ while the number of aggregated local model is $(0.075-\epsilon)|UE|$, if the approximation algorithm in [2] is used to find flow f in G'. Therefore, the approximation ratio of `Placement` is $e(T_t)/OPT\le\vartheta$.

We finally show the time complexity of algorithm `Placement`. The most time consuming parts of algorithm `Placement` consist of (1) constructing the auxiliary graph and finding a unsplittable minimum-cost multicommodity flow in the constructed auxiliary graph, and (2) finding assignment for each UE. For (1), it can be seen that the construction of auxiliary graph has four layers with one node in **layer 1**, $O(|\mathcal{BS}\cup\mathcal{CL}|)$ nodes in **layer 2**, $O(|\mathcal{BS}|)$ nodes in **layer 3**, and $O(|\mathcal{UE}|)$ nodes in **layer 4**. As such, the construction of the auxiliary graph takes $O((2\cdot|\mathcal{BS}|+|\mathcal{CL}|)^2)$ time. According the algorithm in [2], it takes $O(|\mathcal{UE}|\cdot(|\mathcal{BS}|+|\mathcal{CL}|)^2)$ time to find a feasible unsplittable minimum-cost multicommodity flow in the constructed auxiliary graph. For (2), finding the assignment for all UEs based on the found flow takes $O(|UE|)$ time. Thus, the time complexity of `Placement` is $O(|\mathcal{UE}|\cdot(|\mathcal{BS}|+|\mathcal{CL}|)^2)$. □

Proof of Theorem 3

Proof: We first show that the context grouping and loss range expansion method meets the Lipschitz condition by

$$|L_{w,\theta,t}-L_{w,\theta',t}|\le\chi|UB_w-LB_w+2\zeta|,\tag{12}$$

where χ is a given constant with $0<\chi\le 1$. The reason is that each context group has contexts whose losses are in the range of $[LB_w,UB_w]$. If its expert chooses to expand its loss range by ζ, there may or may not be contexts in the expanded range. Therefore, we have $0<\chi\le 1$.

Considering that there are R requests and the loss range of each expert is adjusted after the admission of each request, the loss of each expert may be pushed up by at most $\chi\zeta$. Following the result in theorem 1, we can get the regret

$$R(T)\le|\mathcal{C}|/2\sum\nolimits_{t}^{T}\eta^2+(\ln|\mathcal{C}|)/\eta+|R|\sigma\chi\zeta,$$

if $\eta_1=\eta_2=\cdots=\eta_T=\eta$.

We now analyze the time complexity of algorithm `OL_MULTI`. It must be mentioned that the major difference of `OL_MAB` and `OL_MULTI` is the latter allows each expert to expand its range, and considers multiple FL requests. Therefore, for each FL request, algorithm `OL_MULTI` has an additional time of $O(\Theta)$ for experts to make decisions. As such, the time complexity of `OL_MULTI` is $O(|R| \cdot T(2\Theta + |\mathcal{UE}|))$. $\square$

REFERENCES

[1] Abel transform. *https://en.wikipedia.org/wiki/Abel_transform*.
[2] S. G. Kolliopoulos and C. Stein. Approximation algorithms for single-source unsplittable flow. *SIAM Journal of Computing*, Vol. 31, No. 3, pp.919–946, SIAM, 2001.