Supplementary Material for Memory and Communication Efficient Federated Kernel k-Means

Xiaochen Zhou and Xudong Wang, *Fellow, IEEE*

APPENDIX A
PROOF OF THEOREM 1

We first briefly demonstrate the principle of stochastic kernel PCA. It is a centralized learning algorithm that determines a low-rank estimate of the kernel matrix $\mathbf{K}$ by solving a composite optimization problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times N}} \frac{1}{2}||\mathbf{Z}-\mathbf{K}||_F^2+\lambda||\mathbf{Z}||_*.$$

The optimal solution to this problem is

$$\mathbf{Z}^*=\sum_{\lambda_i>\lambda}(\lambda_i-\lambda)\mathbf{u}_i\mathbf{u}_i^\top=\sum_{i=1}^{s}(\lambda_i-\lambda)\mathbf{u}_i\mathbf{u}_i^\top,$$

where λ_i and $\mathbf{u}_i$ are the i-th eigenvalue and eigenvector of $\mathbf{K}$, respectively. A stochastic optimization method is developed to solve this problem as follows. In the t-th iteration, an unbiased estimate $\boldsymbol{\xi}_t$ of $\mathbf{K}$ is constructed based on a random feature method. The updated solution $\mathbf{Z}_t$ is then computed based on the current solution $\mathbf{Z}_{t-1}$ and $\boldsymbol{\xi}_t$ via stochastic proximal gradient descent

$$\begin{aligned}\mathbf{Z}_t=\arg\min_{\mathbf{Z}\in\mathbb{R}^{N\times N}}&\frac{1}{2}||\mathbf{Z}-\mathbf{Z}_{t-1}||_F^2+\eta_t\langle\mathbf{Z}-\mathbf{Z}_{t-1},\mathbf{Z}_{t-1}-\boldsymbol{\xi}_t\rangle\\&+\eta_t\lambda||\mathbf{Z}||_*,\end{aligned}$$

where η_t is the learning rate in the t-th iteration. $\mathbf{Z}_t$ has a closed-form expression, i.e.,

$$\mathbf{Z}_t=\sum_{\lambda_{i,t}>\eta_t\lambda}(\lambda_{i,t}-\eta_t\lambda)\tilde{\mathbf{u}}_{i,t}\tilde{\mathbf{u}}_{i,t}^\top, \tag{1}$$

where $\lambda_{i,t}$ and $\tilde{\mathbf{u}}_{i,t}$ are the i-th eigenvalue and eigenvector of the matrix $(1-\eta_t)\mathbf{Z}_{t-1}+\eta_t\boldsymbol{\xi}_t$.

We then show that the update rule

$$\mathbf{B}_t=\left[\sqrt{\sigma_{1,t}^2-\eta_t\lambda}\mathbf{u}_{1,t},...,\sqrt{\sigma_{I,t}^2-\eta_t\lambda}\mathbf{u}_{I,t}\right], \tag{2}$$

is actually equivalent to that in Eq. (1). For the estimate $\boldsymbol{\xi}_t$, it can be decomposed as $\boldsymbol{\xi}_t=\frac{1}{D}\mathbf{A}_t\mathbf{A}_t^\top$ according to the random feature method. Next, it is proved that $\mathbf{Z}_t=\mathbf{B}_t\mathbf{B}_t^\top$ via mathematical induction. Initially, $\mathbf{Z}_0=\mathbf{0}$ and $\mathbf{B}_0=\mathbf{0}$. Assume that $\mathbf{Z}_{t-1}=\mathbf{B}_{t-1}\mathbf{B}_{t-1}^\top$. $(1-\eta_t)\mathbf{Z}_{t-1}+\eta_t\boldsymbol{\xi}_t$ can then be written as

$$\begin{aligned}(1-\eta_t)\mathbf{Z}_{t-1}+\eta_t\boldsymbol{\xi}_t&=(1-\eta_t)\mathbf{B}_{t-1}\mathbf{B}_{t-1}^\top+\frac{\eta_t}{D}\mathbf{A}_t\mathbf{A}_t^\top\\&=\mathbf{W}_t\mathbf{W}_t^\top.\end{aligned}$$

Hence, in Eq. (1), $\lambda_{i,t}$ and $\tilde{\mathbf{u}}_{i,t}$ are also the i-th eigenvalue and eigenvector of the matrix $\mathbf{W}_t\mathbf{W}_t^\top$. Moreover, $\lambda_{i,t}=\sigma_{i,t}^2$ and $\tilde{\mathbf{u}}_{i,t}=\mathbf{u}_{i,t}$ where $\sigma_{i,t}$ and $\mathbf{u}_{i,t}$ are the i-th singular value and singular vector of $\mathbf{W}_t$. respectively. According to Eq. (2), $\mathbf{Z}_t$ can be rewritten as $\mathbf{Z}_t=\mathbf{B}_t\mathbf{B}_t^\top$, which completes the mathematical induction.

Similar to $\mathbf{B}_{t-1}=[\mathbf{B}_{t-1}^\top[1],...,\mathbf{B}_{t-1}^\top[M]^\top]$, the updated estimate $\mathbf{B}_t$ can also be rewritten in the form of M submatrices, i.e., $\mathbf{B}_t=[\mathbf{B}_t^\top[1],...,\mathbf{B}_t^\top[M]^\top]$ where $\mathbf{B}_t[m]$ is the updated submatrix at user device m.

Since in stochastic kernel PCA $\mathbf{Z}_t$ converges to $\mathbf{Z}^*=\sum_{i=1}^{s}(\lambda_i-\lambda)\mathbf{u}_i\mathbf{u}_i^\top$, $\mathbf{B}_t$ converges to

$$\mathbf{B}^*=\left[\sqrt{\lambda_1-\lambda}\mathbf{u}_1,...,\sqrt{\lambda_s-\lambda}\mathbf{u}_s\right].$$

APPENDIX B
PROOF OF THEOREM 2

Before the proof, we first define $F(\mathbf{Z})=\frac{1}{2}\mathbb{E}[||\mathbf{Z}-\boldsymbol{\xi}||_F^2]$ and $f_t(\mathbf{Z})=\frac{1}{2}||\mathbf{Z}-\boldsymbol{\xi}_t||_F^2$. For a μ-strongly convex function $l(\mathbf{Z})$, if $l(\mathbf{Z}_1)\geq l(\mathbf{Z}_2)$, then

$$l(\mathbf{Z}_1)-l(\mathbf{Z}_2)\geq\frac{\mu}{2}||\mathbf{Z}_1-\mathbf{Z}_2||_F^2. \tag{3}$$

Let $\mathbf{B}_{t+1}\mathbf{B}_{t+1}^\top=\mathbf{Z}_{t+1}$ and $\hat{\mathbf{B}}_{t+1}\hat{\mathbf{B}}_{t+1}^\top=\mathbf{Z}_{t+1}^*$ where $\mathbf{Z}_{t+1}^*$ is the optimal solution to the optimization problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times N}}\frac{1}{2}||\mathbf{Z}-\mathbf{Z}_t||_F^2+\eta_t\langle\mathbf{Z}-\mathbf{Z}_t,\nabla f_t(\mathbf{Z}_t)\rangle+\eta_t\lambda||\mathbf{Z}||_*. \tag{4}$$

The following lemma is a key step in this proof.

Lemma 1. *Before FEA converges, the following inequality holds, i.e.,*

$$\begin{aligned}&\frac{1}{2}||\mathbf{Z}_{t+1}-\mathbf{Z}_t||_F^2+\eta_t\langle\mathbf{Z}_{t+1}-\mathbf{Z}_t,\nabla f_t(\mathbf{Z}_t)\rangle+\eta_t\lambda||\mathbf{Z}_{t+1}||_*\\\leq&\frac{1}{2}||\mathbf{Z}^*-\mathbf{Z}_t||_F^2+\eta_t\langle\mathbf{Z}^*-\mathbf{Z}_t,\nabla f_t(\mathbf{Z}_t)\rangle+\eta_t\lambda||\mathbf{Z}^*||_*,\end{aligned} \tag{5}$$

where $\mathbf{Z}^*=\mathbf{B}^*\mathbf{B}^{*\top}$ is the optimal solution to

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times N}}F(\mathbf{Z})+\lambda||\mathbf{Z}||_*.$$

Proof. The objective function in (4) can be rewritten as

$$\begin{aligned}&\frac{1}{2}||\mathbf{Z}-\mathbf{Z}_t||_F^2+\eta_t\langle\mathbf{Z}-\mathbf{Z}_t,\nabla f_t(\mathbf{Z}_t)\rangle+\eta_t\lambda||\mathbf{Z}||_*\\=&\frac{1}{2}||\mathbf{Z}-[(1-\eta_t)\mathbf{Z}_t+\eta_t\boldsymbol{\xi}_t]||_F^2+\eta_t\lambda||\mathbf{Z}||_*-\frac{\eta_t^2}{2}||\nabla f_t(\mathbf{Z}_t)||_F^2.\end{aligned}$$

Since $\frac{\eta_t^2}{2}||\nabla f_t(\mathbf{Z}_t)||_F^2$ is a constant, we can only consider

$$l(\mathbf{Z})=\frac{1}{2}||\mathbf{Z}-[(1-\eta_t)\mathbf{Z}_t+\eta_t\boldsymbol{\xi}_t]||_F^2+\eta\lambda||\mathbf{Z}||_*$$

in the following part of the proof.

Now we first assume that $l(\mathbf{Z}^*) \leq l(\mathbf{Z}_{t+1})$, then we have

$$l(\mathbf{Z}_{t+1}) - l(\mathbf{Z}^*_{t+1}) \geq l(\mathbf{Z}^*) - l(\mathbf{Z}^*_{t+1}) \geq \frac{\mu}{2}||\mathbf{Z}^*_{t+1} - \mathbf{Z}^*||_F^2. \tag{6}$$

Let $\mathbf{R}_t$ denote $(1-\eta_t)\mathbf{Z}_t + \eta_t\boldsymbol{\xi}_t$, then $l(\mathbf{Z}_{t+1}) - l(\mathbf{Z}^*_{t+1})$ can be expanded as

$$\begin{aligned}
& l(\mathbf{Z}_{t+1}) - l(\mathbf{Z}^*_{t+1}) \\
=& \frac{1}{2}(||\mathbf{Z}_{t+1} - \mathbf{R}_t||_F - ||\mathbf{Z}^*_{t+1} - \mathbf{R}_t||_F) \\
& (||\mathbf{Z}_{t+1} - \mathbf{R}_t||_F + ||\mathbf{Z}^*_{t+1} - \mathbf{R}_t||_F) \\
& + \eta_t\lambda(||\mathbf{Z}_{t+1}||_* - ||\mathbf{Z}^*_{t+1}||_*) \\
\leq& \frac{1}{2}||\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1}||_F(||\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1}||_F + 2||\mathbf{Z}^*_{t+1} - \mathbf{R}_t||_F) \\
& + \eta_t\lambda||\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1}||_*
\end{aligned} \tag{7}$$

It is well known that given a matrix $\mathbf{M}$ the following inequality holds for its nuclear norm and its Frobenius norm, i.e., $||\mathbf{M}||_*^2 \leq \mathrm{rank}(\mathbf{M})||\mathbf{M}||_F^2$. By this inequality, we have

$$||\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1}||_* \leq \sqrt{r}||\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1}||_F \leq \sqrt{r}N\epsilon, \tag{8}$$

where r is the rank of $(\mathbf{Z}_{t+1} - \mathbf{Z}^*_{t+1})$. Substitute (8) into (7), we have

$$l(\mathbf{Z}_{t+1}) - l(\mathbf{Z}^*_{t+1}) \leq \frac{1}{2}N^2\epsilon^2 + n\epsilon||\mathbf{Z}^*_{t+1} - \mathbf{R}_t||_F + \eta_t\lambda\sqrt{r}n\epsilon.$$

Since $||\mathbf{Z}^*_{t+1} - \mathbf{R}_t||_F$ is a constant, this upper bound of $l(\mathbf{Z}_{t+1}) - l(\mathbf{Z}^*_{t+1})$ can become arbitrarily small if ϵ is arbitrarily small. Hence, according to (6), $||\mathbf{Z}^*_{t+1} - \mathbf{Z}^*||_F^2$ can also be arbitrarily small. However, this contradicts that $||\mathbf{Z}^*_{t+1} - \mathbf{Z}^*||_F^2$ cannot become arbitrarily small before the convergence of FEA. Therefore, the assumption $l(\mathbf{Z}^*) \leq l(\mathbf{Z}_{t+1})$ does not hold. In other words, $l(\mathbf{Z}^*) \geq l(\mathbf{Z}_{t+1})$ is satisfied before the convergence of FEA. □

The rest part then follows the proof of Theorem 1 in [1]. Based on Lemma 1 and the property of strongly convex function in (3), we have

$$\begin{aligned}
& \frac{1}{2}||\mathbf{Z}_{t+1} - \mathbf{Z}_t||_F^2 + \eta_t\langle\mathbf{Z}_{t+1} - \mathbf{Z}_t, \nabla f_t(\mathbf{Z}_t)\rangle + \eta_t\lambda||\mathbf{Z}_{t+1}||_* \\
\leq& \frac{1}{2}||\mathbf{Z}^* - \mathbf{Z}_t||_F^2 + \eta_t\langle\mathbf{Z}^* - \mathbf{Z}_t, \nabla f_t(\mathbf{Z}_t)\rangle + \eta_t\lambda||\mathbf{Z}^*||_* \\
& - \frac{1}{2}||\mathbf{Z}^* - \mathbf{Z}_{t+1}||_F^2.
\end{aligned} \tag{9}$$

Similarly, according to (3) we have

$$\frac{1}{2}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \leq F(\mathbf{Z}_t) + \lambda||\mathbf{Z}_t||_* - F(\mathbf{Z}^*) - \lambda||\mathbf{Z}^*||_*.$$

Since $F(\mathbf{Z})$ is 1-strongly convex, then

$$\begin{aligned}
& \frac{1}{2}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \\
\leq& \langle\mathbf{Z}_t - \mathbf{Z}^*, \nabla F(\mathbf{Z}_t)\rangle - \frac{1}{2}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 + \lambda(||\mathbf{Z}_t||_* - ||\mathbf{Z}^*||_*) \\
=& \langle\mathbf{Z}_t - \mathbf{Z}^*, \nabla f_t(\mathbf{Z}_t)\rangle - \lambda||\mathbf{Z}^*||_* - \frac{1}{2\eta_t}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \\
& + \lambda||\mathbf{Z}_t||_* - \frac{1}{2}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 + \frac{1}{2\eta_t}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \\
& + \langle\nabla F(\mathbf{Z}_t) - \nabla f_t(\mathbf{Z}_t), \mathbf{Z}_t - \mathbf{Z}^*\rangle.
\end{aligned}$$

Based on (9), we eventually obtain that

$$\begin{aligned}
& \frac{1}{2}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \\
\leq& \langle\mathbf{Z}_t - \mathbf{Z}_{t+1}, \nabla f_t(\mathbf{Z}_t)\rangle - \lambda||\mathbf{Z}_{t+1}||_* - \frac{1}{2\eta_t}||\mathbf{Z}_{t+1} - \mathbf{Z}_t||_F^2 \\
& - \frac{1}{2\eta_t}||\mathbf{Z}^* - \mathbf{Z}_{t+1}||_F^2 + \lambda||\mathbf{Z}_t||_* + \frac{1-\eta_t}{2\eta_t}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 \\
& + \langle\nabla F(\mathbf{Z}_t) - \nabla f_t(\mathbf{Z}_t), \mathbf{Z}_t - \mathbf{Z}^*\rangle \\
\leq& \frac{\eta_t}{2}||\nabla f_t(\mathbf{Z}_t)||_F^2 - \frac{1}{2\eta_t}||\mathbf{Z}_{t+1} - \mathbf{Z}^*||_F^2 + \lambda(||\mathbf{Z}_t||_* - ||\mathbf{Z}_{t+1}||_*) \\
& + \frac{1-\eta_t}{2\eta_t}||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 + \langle\nabla F(\mathbf{Z}_t) - \nabla f_t(\mathbf{Z}_t), \mathbf{Z}_t - \mathbf{Z}^*\rangle
\end{aligned} \tag{10}$$

By substituting $\delta_t = \langle\boldsymbol{\xi}_t - \mathbf{K}, \mathbf{Z}_t - \mathbf{Z}^*\rangle$ and $C^2 = \max_{t\in[T]}||\mathbf{Z}_t - \boldsymbol{\xi}_t||_F^2$ into (10),

$$\begin{aligned}
||\mathbf{Z}_{t+1} - \mathbf{Z}^*||_F^2 \leq& \eta_t^2C^2 + 2\eta_t\delta_t + 2\lambda\eta_t(||\mathbf{Z}_t||_* - ||\mathbf{Z}_{t+1}||_*) \\
& + (1-2\eta_t)||\mathbf{Z}_t - \mathbf{Z}^*||_F^2.
\end{aligned} \tag{11}$$

The inequality in (11) is the same as the result of Lemma 1 in [1]. Thus, the following lemmas [1] from [1] can be directly utilized to derive a probability bound for $||\mathbf{Z}_{t+1} - \mathbf{Z}^*||_F^2$.

Lemma 2 (Lemma 2 in [1]). *Define $\gamma = \max_{t\in[T]}||\mathbf{Z}_t||_*$. By setting $\eta_t = \frac{2}{t}$, an upper bound of $||\mathbf{Z}_{t+1} - \mathbf{Z}^*||_F^2$ can be written as*

$$\begin{aligned}
& ||\mathbf{Z}_{T+1} - \mathbf{Z}^*||_F^2 \\
\leq& \frac{2}{T(T-1)}\left[2\sum_{t=2}^{T}(t-1)\delta_t - \sum_{t=2}^{T}(t-1)||\mathbf{Z}_t - \mathbf{Z}^*||_F^2\right] \\
& + \frac{4(C^2 + \lambda\gamma)}{T}.
\end{aligned}$$

The upper bound of $\sum_{t=2}^{T}(t-1)\delta_t$ in Lemma 2 is then provided in Lemma 3.

Lemma 3 (Lemma 3 in [1]). *Assume $||\boldsymbol{\xi}_t - \mathbf{K}||_F \leq G$, and $||\mathbf{Z}_t - \mathbf{Z}^*||_F \leq H$, $\forall t > 2$. With a probability at least $1-\delta$, $\sum_{t=2}^{T}(t-1)\delta_t$ is upper bounded by*

$$\begin{aligned}
\sum_{t=2}^{T}(t-1)\delta_t \leq& \frac{1}{2}\sum_{t=2}^{T}(t-1)||\mathbf{Z}_t - \mathbf{Z}^*||_F^2 + 2G^2\tau(T-1) \\
& + \frac{2}{3}GH(T-1)\tau + GH(T-1),
\end{aligned}$$

where $\tau = \log\frac{\lceil 2\log_2 T\rceil}{\delta}$.

Based on Lemma 2 and Lemma 3, the following upper bound of $||\mathbf{Z}_{T+1} - \mathbf{Z}^*||_F^2$ holds with a probability at least $1-\delta$,

$$||\mathbf{Z}_{T+1} - \mathbf{Z}^*||_F^2 \leq \frac{4}{T}\left(C^2 + \lambda\gamma + 2G^2\tau + \frac{2}{3}GH\tau + GH\right). \tag{12}$$

In Lemma 4, the upper bounds for C, γ, G, and H are provided.

[1]These lemmas can be found in the supplementary material of [1] that can be downloaded from https://cs.nju.edu.cn/zlj/pdf/AAAI-2016-Zhang-S.pdf

Lemma 4 (Lemma 4 in [1])**.** *Assume $||\boldsymbol{\xi}||_F \le L$. By setting $\eta_t = \frac{2}{t}$, it can be obtained that*

$$C^2 \le 10L^2,\ \gamma \le 2L \max_{t\in[T]} \sqrt{r_t},\ G = 2L,\ and\ H = 3L,$$

where r_t is the rank of $\mathbf{Z}_t$.

By substituting the upper bounds in Lemma 4 into (12) and replacing $\mathbf{Z}_{T+1}$ and $\mathbf{Z}^*$ with $\mathbf{B}_{T+1}\mathbf{B}_{T+1}^\top$ and $\mathbf{B}^*\mathbf{B}^{*\top}$, respectively, we eventually obtain

$$\frac{||\mathbf{B}_{T+1}\mathbf{B}_{T+1}^\top - \mathbf{B}^*\mathbf{B}^{*\top}||_F^2}{N^2} \le \frac{8}{T}\left[\lambda \frac{L}{n^2}\max_{t\in[T]}\sqrt{r_t} + \frac{L^2}{n^2}\left(8 + 6\log\frac{\lceil 2\log_2 T\rceil}{\delta}\right)\right] = O(\frac{1}{T}).$$

APPENDIX C
PROOF OF THEOREM 3

In the t-th iteration of FEA, the following procedure is executed for Q_t iterations. The central server broadcasts a vector $\mathbf{c}_q$ to all M user devices. User device m computes a local vector $\mathbf{d}_m = \mathbf{W}_t[m]^\top \mathbf{W}_t[m]\mathbf{c}_q$ and then uploads $\mathbf{d}_m$ to the central server. Since $\mathbf{W}_t[m] = [\sqrt{\frac{\eta_t}{D}}\mathbf{A}_t[m], \sqrt{1-\eta_t}\mathbf{B}_{t-1}[m]] \in \mathbb{R}^{N_m\times(r_{t-1}+D)}$, $\mathbf{W}_t[m]^\top \mathbf{W}_t[m]$ is a matrix with dimensions of $(r_{t-1}+D)\times(r_{t-1}+D)$. Hence, the dimensions of both $\mathbf{c}_q$ and $\mathbf{d}_m$ also equal $(r_{t-1}+D)$. As a result, the communication cost equals $2Q_tM(r_{t-1}+D)$ in the t-th iteration of FEA, which shows that the communication cost is linear to $M(r_{t-1}+D)$.

For the centralized method, the user devices first uploads $\{\mathbf{W}_t[m], m \in \mathcal{M}\}$ to the central server, the central server then sends the updated submatrix $\mathbf{B}_t[m]$ to user device m for all $m \in \mathcal{M}$. Thus, its communication cost equals $N(r_{t-1}+r_t+D)$ in the t-th iteration of FEA. Thus, CELA reduces the communication cost of FEA with a rate

$$\eta_t = 1 - \frac{2Q_tM(r_{t-1}+D)}{N(r_{t-1}+r_t+D)} \ge 1 - \frac{2MQ_t}{N}.$$

Note that $Q_t \le r_{t-1}+D$ according to the analysis in Section IV-A. Eventually, we have

$$1 - \frac{2M(r_{t-1}+D)}{N} \le \eta_t.$$

APPENDIX D
PROOF OF THEOREM 4

Define $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, and $\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$. The low-rank approximation of $\mathbf{K}$ with rank s is denoted as $\mathbf{K}_s = \mathbf{U}\boldsymbol{\Lambda}_s\mathbf{U}^\top$, and $\mathbf{P}_s = \mathbf{U}\boldsymbol{\Lambda}_s^{\frac{1}{2}}$ where the diagonal of $\boldsymbol{\Lambda}_s$ contains the s largest eigenvalues of $\mathbf{K}$ while its rest diagonal entries are all zero. The output of FEA at iteration t is an estimation of $\mathbf{K}_s$, denoted as $\widetilde{\mathbf{K}}_t$, and $\widetilde{\mathbf{K}}_t = \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top$.

The following two lemmas will be used in the proof of Theorem 3.

Lemma 5. *Given $\widetilde{\mathbf{K}}_t$, the following inequality holds with a probability at least $1-\delta$ for any rank k projection matrix $\boldsymbol{\Pi} \in \mathbb{R}^{N\times N}$,*

$$\mathrm{Tr}(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)) \le O(\sqrt{\frac{s}{t\delta}}N)$$

Proof. Since $\boldsymbol{\Pi}$ is a rank-k projection matrix, it is obvious that $\mathrm{Tr}(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)) \le ||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_*$ For a rank-s matrix $\mathbf{A}$, $||\mathbf{A}||_*^2 \le s||\mathbf{A}||_F^2$ holds for its Nuclear norm and its Frobenius norm Hence, $||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_* \le \sqrt{s}||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_F$. By Lemma 4, $\mathbf{Z}_t$ converges to $\mathbf{Z}^*$ at an $O(N^2/t\delta)$ rate. Note $\mathbf{Z}^*$ has the same eigenvectors as $\mathbf{K}_s$. Thus, $\widetilde{\mathbf{K}}_t$ constructed based on $\mathbf{Z}_t$ also converges to $\mathbf{K}_s$ at an $O(N^2/t\delta)$ rate with a probability at least $1-\delta$, i.e., $||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_F^2$ has an upper bound as

$$||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_F^2 \le O(N^2/t\delta).$$

Hence, the following inequality holds with a probability of at least $1-\delta$

$$\mathrm{Tr}(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)) \le \sqrt{s}||\mathbf{K}_s - \widetilde{\mathbf{K}}_t||_F \le O(\sqrt{\frac{s}{t\delta}}N).$$

□

Lemma 6. *Fix an error parameter $\varepsilon \in (0,1)$. For any rank k projection matrix $\boldsymbol{\Pi} \in \mathbb{R}^{N\times N}$,*

$$\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \widetilde{\mathbf{K}}_t)\boldsymbol{\Pi}\right) \le (\varepsilon + \frac{k}{s})||\mathbf{P} - \boldsymbol{\Pi}\mathbf{P}||_F^2.$$

Proof. $\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \mathbf{K}_s)\boldsymbol{\Pi}\right)$ can be expanded as

$$\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \mathbf{K}_s)\boldsymbol{\Pi}\right) = \mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{P}\mathbf{P}^\top - \mathbf{P}_s\mathbf{P}_s^\top)\boldsymbol{\Pi}\right).$$

Note that $\mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\top - \mathbf{P}_s\mathbf{P}_s^\top\right) = \sum_{i=s+1}^N \sigma_i^2(\mathbf{P})$, where $\sigma_i(\mathbf{P})$ is the i-th singular value of $\mathbf{P}$. After multiplied with a rank-k projection matrix $\boldsymbol{\Pi}$, the maximal value of $\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \mathbf{K}_s)\right)$ is achieved when the largest k singular values in $\{\sigma_i(\mathbf{P}), i = s+1, ..., N\}$ are kept. Hence, we have

$$\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \mathbf{K}_s)\boldsymbol{\Pi}\right) \le \sum_{i=s+1}^{s+k} \sigma_i^2(\mathbf{P}). \tag{13}$$

$\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \widetilde{\mathbf{K}}_t)\boldsymbol{\Pi}\right)$ is then expanded as

$$\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \widetilde{\mathbf{K}}_t)\boldsymbol{\Pi}\right)\\
=&\mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K} - \mathbf{K}_s)\boldsymbol{\Pi}\right) + \mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)\boldsymbol{\Pi}\right)\\
\le& \sum_{i=s+1}^{s+k} \sigma_i^2(\mathbf{P}) + \mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)\boldsymbol{\Pi}\right)\\
=& \sum_{i=s+1}^{s+k} \sigma_i^2(\mathbf{P}) + \mathrm{Tr}\left(\boldsymbol{\Pi}(\mathbf{K}_s - \widetilde{\mathbf{K}}_t)\right)\\
\le& \sum_{i=s+1}^{s+k} \sigma_i^2(\mathbf{P}) + O(\sqrt{\frac{s}{t\delta}}N)\\
\le& \frac{k}{s}\sum_{i=k+1}^{s+k} \sigma_i^2(\mathbf{P}) + O(\sqrt{\frac{s}{t\delta}}N),
\end{aligned}$$

where the first inequality comes from Eq. (13), the second inequality comes from Lemma 5, and $k \le s$.

For $\sum_{i=k+1}^{s+k} \sigma_i^2(\mathbf{P})$, we have

$$\sum_{i=k+1}^{s+k} \sigma_i^2(\mathbf{P}) \le \sum_{i=k+1}^{N} \sigma_i^2(\mathbf{P}) = ||\mathbf{P} - \mathbf{P}_k||_F^2.$$

Since $||\mathbf{P}||_F^2 = \mathrm{Tr}K = O(N)$, we have $||\mathbf{P} - \mathbf{P}_k||_F^2 = O(N)$ Thus, $O(\sqrt{\frac{s}{t\delta}}N)$ can be rewritten as $\varepsilon||\mathbf{P} - \mathbf{P}_k||_F^2$ where ε is still at the order of $O(\sqrt{\frac{s}{t\delta}})$. As a result, we have

$$\mathrm{Tr}\left(\mathbf{\Pi}(\mathbf{K} - \widetilde{\mathbf{K}}_t)\mathbf{\Pi}\right) \leq (\varepsilon + \frac{k}{s})||\mathbf{P} - \mathbf{P}_k||_F^2.$$

Since $||\mathbf{P} - \mathbf{P}_k||_F^2$ is the minimal value of $||\mathbf{P} - \mathbf{\Pi}\mathbf{P}||_F^2$ for any $\mathbf{\Pi}$, we then have

$$\mathrm{Tr}\left(\mathbf{\Pi}(\mathbf{K} - \widetilde{\mathbf{K}}_t)\mathbf{\Pi}\right) \leq (\varepsilon + \frac{k}{s})||\mathbf{P} - \mathbf{\Pi}\mathbf{P}||_F^2.$$

□

After completing the proofs of Lemma 5 and Lemma 6, we then finish the rest proof of Theorem 3 as follows. It can be obtained that

$$\begin{aligned} &||(\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}||_F^2 - ||(\mathbf{I}_N - \mathbf{\Pi})\widetilde{\mathbf{P}}_t||_F^2 \\ =&\mathrm{Tr}((\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}\mathbf{P}^\top) - \mathrm{Tr}((\mathbf{I}_N - \mathbf{\Pi})\widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top) \\ =&\mathrm{Tr}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top) - \mathrm{Tr}(\mathbf{\Pi}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top)\mathbf{\Pi}). \end{aligned}$$

Let $\alpha = \mathrm{Tr}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top)$, and then the above equation can be rewritten as

$$||(\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}||_F^2 + \mathrm{Tr}(\mathbf{\Pi}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top)\mathbf{\Pi}) = \alpha + ||(\mathbf{I}_N - \mathbf{\Pi})\widetilde{\mathbf{P}}_t||_F^2.$$

After sufficient iterations, both α and $\mathrm{Tr}(\mathbf{\Pi}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top)\mathbf{\Pi})$ are non-negative with a high probability. Thus, by Lemma 6 it holds that

$$\begin{aligned} &||(\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}||_F^2 \\ \leq&\alpha + ||(\mathbf{I}_N - \mathbf{\Pi})\widetilde{\mathbf{P}}_t||_F^2 \\ =&||(\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}||_F^2 + \mathrm{Tr}(\mathbf{\Pi}(\mathbf{P}\mathbf{P}^\top - \widetilde{\mathbf{P}}_t\widetilde{\mathbf{P}}_t^\top)\mathbf{\Pi}) \\ \leq&(1 + \varepsilon + \frac{k}{s})||(\mathbf{I}_N - \mathbf{\Pi})\mathbf{P}||_F^2. \end{aligned} \tag{14}$$

Based on (14), Theorem 3 can be proved as follows. Let $\mathbf{\Pi} = \widetilde{\mathbf{Y}}_t\widetilde{\mathbf{L}}_t\widetilde{\mathbf{Y}}_t^\top$, where $\widetilde{\mathbf{Y}}_t$ is the indicator matrix obtained by applying a γ-approximate algorithm to $\widetilde{\mathbf{P}}_t$, then

$$\begin{aligned} ||(\mathbf{I}_N - \widetilde{\mathbf{Y}}_t\widetilde{\mathbf{L}}_t\widetilde{\mathbf{Y}}_t^\top)\mathbf{P}||_F^2 &\leq \alpha + ||(\mathbf{I}_N - \widetilde{\mathbf{Y}}_t\widetilde{\mathbf{L}}_t\widetilde{\mathbf{Y}}_t^\top)\widetilde{\mathbf{P}}_t||_F^2 \\ &\leq \alpha + \gamma||(\mathbf{I}_N - \widetilde{\mathbf{Y}}_t^*\widetilde{\mathbf{L}}_t^*\widetilde{\mathbf{Y}}_t^{*\top})\widetilde{\mathbf{P}}_t||_F^2, \end{aligned}$$

where $\widetilde{\mathbf{Y}}_t^*$ is the optimal indicator matrix for the linear k-means problem on $\widetilde{\mathbf{P}}_t$. Since $\gamma > 1$, it follows that

$$\begin{aligned} &\alpha + \gamma||(\mathbf{I}_N - \widetilde{\mathbf{Y}}_t^*\widetilde{\mathbf{L}}_t^*\widetilde{\mathbf{Y}}_t^{*\top})\widetilde{\mathbf{P}}_t||_F^2 \\ \leq&\alpha + \gamma||(\mathbf{I}_N - \mathbf{Y}^*\mathbf{L}^*\mathbf{Y}^{*\top})\widetilde{\mathbf{P}}_t||_F^2 \\ \leq&\gamma(1 + \varepsilon + \frac{k}{s})||(\mathbf{I}_N - \mathbf{Y}^*\mathbf{L}^*\mathbf{Y}^{*\top})\mathbf{P}||_F^2. \end{aligned}$$

Thus,

$$||(\mathbf{I}_N - \widetilde{\mathbf{Y}}_t\widetilde{\mathbf{L}}_t\widetilde{\mathbf{Y}}_t^\top)\mathbf{P}||_F^2 \leq \gamma(1+\varepsilon+\frac{k}{s})||(\mathbf{I}_N - \mathbf{Y}^*\mathbf{L}^*\mathbf{Y}^{*\top})\mathbf{P}||_F^2,$$

which is equivalent to $f_K(\widetilde{\mathbf{Y}}_t) \leq \gamma(1 + \varepsilon + \frac{k}{s}) \min_{\mathbf{Y}} f_K(\mathbf{Y})$.

APPENDIX E
DISCUSSION ON PRIVACY PRESERVATION

If the central server collects sufficient random feature vectors of a data samples, then it is possible for the central server to recover the data samples from these random features. The reason is as follows. A random feature vector of a data sample $\mathbf{x}_i$ has the form $\cos(\boldsymbol{\omega}^\top\mathbf{x}_i + b)$ where the $\boldsymbol{\omega}$ and b are known by the central server. Since the value of $\boldsymbol{\omega}^\top\mathbf{x}_i + b$ cannot be arbitrarily large, the central server can determine the value of $\boldsymbol{\omega}^\top\mathbf{x}_i + b$ for each random feature vector if sufficient such random features are collected. The data sample $\mathbf{x}_i$ can be recovered by solving a system of linear equations.

In FedKKM, the above recovering operation is infeasible, which is proved by the following lemma.

Lemma 7. *Based on the collected vectors $\{\mathbf{g}_q = \mathbf{W}_t^\top\mathbf{W}_t\mathbf{c}_q, q = 1, ..., Q\}$, the central server can at most recover the matrix $\mathbf{W}_t^\top\mathbf{W}_t$ via matrix operations. Moreover, recovering the matrix $\mathbf{A}_t$ from the matrix $\mathbf{W}_t^\top\mathbf{W}_t$ is an ill-posed problem with infinite solutions.*

Proof. In FedKKM, the central server collects the vectors $\{\mathbf{g}_q = \mathbf{W}_t^\top\mathbf{W}_t\mathbf{c}_q, q = 1, ..., Q\}$. If $\mathbf{W}_t^\top\mathbf{W}_t$ is a full-rank matrix, and Q equals the rank of $\mathbf{W}_t^\top\mathbf{W}_t$, the central server can compute $\mathbf{W}_t^\top\mathbf{W}_t$ based on the matrices $\mathbf{G} = [\mathbf{g}_1, ..., \mathbf{g}_Q]$ and $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_Q]$, i.e.,

$$\mathbf{W}_t^\top\mathbf{W}_t = \mathbf{G}\mathbf{C}^{-1},$$

where $\mathbf{C}^{-1}$ is the inverse matrix of $\mathbf{C}$.

Since

$$\mathbf{W}_t^\top\mathbf{W}_t = \begin{bmatrix} \frac{\eta_t}{D}\mathbf{A}_t^\top\mathbf{A}_t & \sqrt{\frac{\eta_t(1-\eta_t)}{D}}\mathbf{A}_t^\top\mathbf{B}_{t-1} \\ \sqrt{\frac{\eta_t(1-\eta_t)}{D}}\mathbf{B}_{t-1}^\top\mathbf{A}_t & (1-\eta_t)\mathbf{B}_t^\top\mathbf{B}_{t-1} \end{bmatrix},$$

$\mathbf{A}_t^\top\mathbf{A}_t$ can be recovered from $\mathbf{W}_t^\top\mathbf{W}_t$.

For a matrix $\mathbf{A}_t \in \mathbb{R}^{N\times D}$, a matrix $\mathbf{A}' \in \mathbb{R}^{N\times D}$ can be constructed via

$$\mathbf{A}' = \mathbf{U}_o\mathbf{A}_t,$$

where $\mathbf{U}_o \in \mathbb{R}^{N\times N}$ is an arbitrary orthogonal matrix with $\mathbf{U}_o^\top\mathbf{U}_o = \mathbf{I}_n$. By this construction, it can be derived that

$$\mathbf{A}'^\top\mathbf{A}' = \mathbf{A}_t^\top\mathbf{U}_o^\top\mathbf{U}_o\mathbf{A}_t = \mathbf{A}_t^\top\mathbf{A}_t.$$

Since there exist infinite matrices $\mathbf{U}_o$ satisfying $\mathbf{U}_o^\top\mathbf{U}_o = \mathbf{I}_n$, the problem $\mathbf{A}_t^\top\mathbf{A}_t = \mathbf{A}'^\top\mathbf{A}'$ has infinite solutions. Hence, recovering the random feature matrix $\mathbf{A}_t$ from $\mathbf{A}_t^\top\mathbf{A}_t$ is an ill-posed problem with infinite solutions. □

By Lemma 7, the central server cannot recover $\mathbf{A}_t$ from $\mathbf{A}_t^\top\mathbf{A}_t$. Without such random feature vectors, it is infeasible for the central server to recover users' data via matrix operations.

REFERENCES

[1] L. Zhang, T. Yang, J. Yi, R. Jin, and Z.-H. Zhou, "Stochastic optimization for kernel PCA," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 2316–2322.