Enabling Communication-efficient and Robust Federated Learning over Packet Lossy Networks via Random Interleaved Vector Quantization

IEEE International Conference on Multimedia and Expo (ICME) 2025

APPENDIX

Proof of Theorem 1. Define $\boldsymbol{g} = (g_1, g_2, ..., g_n)$ as raw gradient, the received lossy gradient compressed by RIVQ as $\tilde{\boldsymbol{g}}$, erasure rate as r and the compression error as $\boldsymbol{e} = (e_1, e_2, ..., e_n)$, then the cosine distance between $\boldsymbol{g}$ and $\tilde{\boldsymbol{g}}$ satisifies:

$$\begin{aligned}
\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}, \boldsymbol{g}) &\overset{(a)}{=} 1 - \frac{\boldsymbol{g}\tilde{\boldsymbol{g}}}{\|\boldsymbol{g}\|\|\tilde{\boldsymbol{g}}\|} \\
&\overset{(b)}{=} 1 - \frac{\|\tilde{\boldsymbol{g}}\|^2}{\|\boldsymbol{g}\|\sqrt{\sum_{i\in S_e^c}(g_i^2 + e_i^2 + 2g_i e_i)}} \\
&\overset{(c)}{\leq} 1 - \frac{(1-r)\|\boldsymbol{g}\|^2}{\|\boldsymbol{g}\|\sqrt{\sum_{i\in S_e^c} 2(g_i^2 + e_i^2)}} \\
&\overset{(d)}{\leq} 1 - \frac{(1-r)\|\boldsymbol{g}\|^2}{\|\boldsymbol{g}\|\sqrt{2(1-r)\sum_{i=1}^n (g_i^2 + e_i^2)}} \\
&\overset{(e)}{\leq} 1 - \frac{\sqrt{(1-r)}\|\boldsymbol{g}\|^2}{\sqrt{2(\|\boldsymbol{e}\|^2 + \|\boldsymbol{g}\|^2)}}
\end{aligned} \tag{1}$$

where inequality (b) and (c) follow from the fact that $\langle \boldsymbol{g}, \tilde{\boldsymbol{g}} \rangle = \sum_{i\in S_e^c} g_i^2 = \|\tilde{\boldsymbol{g}}\|^2$, $\|\tilde{\boldsymbol{g}}\|^2 = \sum_{i\in S_e^c} g_i^2 \approx (1-r)\sum_{i=1}^n g_i^2$ and Young's inequality. Inequality (d) holds under $\sum_{i\in S_e^c} e_i^2 \approx (1-r)\sum_{i=1}^n e_i^2 = (1-r)\|\boldsymbol{e}\|^2$. S_e^c refers to the set of received gradient entries, i.e., complement of erasure entry set S_e. □

Proof of Theorem 2. Supposing global loss function $F(\boldsymbol{w}_t) = \frac{1}{K}\sum_{k=1}^K f_k(\boldsymbol{w}_t)$ satisifies L-smooth (Assumption 1), we have

$$\begin{aligned}
&F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}_t) \\
&\overset{(a)}{\leq} \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\
&\overset{(b)}{\leq} -\eta \langle \nabla F(\boldsymbol{w}_t), \tilde{\boldsymbol{g}}_t \rangle + \frac{L\eta^2}{2}\|\tilde{\boldsymbol{g}}_t\|^2 \\
&\overset{(c)}{\leq} -\eta \langle \nabla F(\boldsymbol{w}_t), (\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t) + \boldsymbol{g}_t \rangle + \frac{L\eta^2}{2}\|(\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t) + \boldsymbol{g}_t\|^2 \\
&\overset{(d)}{\leq} -\eta\|\nabla F(\boldsymbol{w}_t)\|^2 + \eta\|\nabla F(\boldsymbol{w}_t)\|\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\| \\
&\qquad + L\eta^2\|(\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t)\|^2 + L\eta^2\|\boldsymbol{g}_t\|^2 \\
&\overset{(e)}{\leq} -\eta\|\nabla F(\boldsymbol{w}_t)\|^2 + \eta\|\nabla F(\boldsymbol{w}_t)\|^2\sqrt{2\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}), \boldsymbol{g}} \\
&\qquad + 2L\eta^2\|\nabla F(\boldsymbol{w}_t)\|^2\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}, \boldsymbol{g}) + L\eta^2\|\nabla F(\boldsymbol{w}_t)\|^2 \\
&\overset{(f)}{\leq} \|\nabla F(\boldsymbol{w}_t)\|^2[\eta^2 L - \eta + \eta\sqrt{2\tilde{d}_{\cos}(\boldsymbol{g}, \tilde{\boldsymbol{g}})} + 2\eta^2 L\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}, \boldsymbol{g})]
\end{aligned} \tag{2}$$

where inequality (a) holds under Taylor's theorem [1] and L-smooth (Assumption 1); (b) commences from $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\tilde{\boldsymbol{g}}_t$ and $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t + \boldsymbol{e}_t$; (d) is derived by Cauchy-Schwarz inequality and triangle inequality; (e) originates from the fact that $\|\boldsymbol{g} - \tilde{\boldsymbol{g}}_t\| \approx \|\boldsymbol{g}_t\|\sqrt{2\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}_t, \boldsymbol{g}_t)}$.

Then, we compute the telescoping sums of inequality (2) over $t = 1, 2, ..., T$ and obtain:

$$\begin{aligned}
F(\boldsymbol{w}^*) - F(\boldsymbol{w}_1) \leq F(\boldsymbol{w}_{T+1}) - F(\boldsymbol{w}_1) \leq [\eta - \eta^2 L \\
- \eta\sqrt{2\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}, \boldsymbol{g})} - 2\eta^2 L\tilde{d}_{\cos}(\tilde{\boldsymbol{g}}, \boldsymbol{g})]\sum_{t=0}^{T-1}\|\nabla F(\boldsymbol{w}_t)\|^2
\end{aligned} \tag{3}$$

By substituting $\eta = \sqrt{\frac{K}{LT}}$ into inequality (3), then multiplying by $\frac{1}{T}$, taking expectations on its both sides and rewriting, we get desired results (omit some subscripts for brevity). □

Proof of the Effectiveness of Gradient Random Interleaving. Define the raw gradient as $\boldsymbol{g} = (g_1, g_2, ..., g_n)$, $\boldsymbol{g}$ is equally partitioned into M packets, i.e., $\boldsymbol{g} = (P_1, P_2, ..., P_M)$, P_i/g_i is described as ~~P_i~~ / ~~g_i~~ when packet i is erased.

(1). *Gradient Entry Erasure after Random Interleaving*

For g_i, whether it is erased or not depends on which packet is assigned to after interleaving. Designate the packet index of g_i after interleaving as J_i, where:

$$P(J_i = j) = \frac{1}{M}, \forall j \in 1, 2, ..., M$$

g_i is erased only when the packet $J_i = j$ is erased, hence:

$$P(\text{~~}g_i\text{~~}) = \sum_{i=i}^{M}\frac{1}{M}\cdot P(\text{~~}P_j\text{~~}) = r$$

(2). *Independence of Gradient Entry Erasure Events*

For two arbitrary gradient entries, random interleaving enables that whether or not they are assigned to the same packet is independent. Therefore, they are conditionally independent as well when it comes to packet erasure.

$$\begin{aligned}
P(g_i \in P_j | g_{i'} \in P_j) &= P(g_i \in P_j) \\
P(\text{~~}g_i\text{~~}|\text{~~}g_{i'}\text{~~}) &= P(\text{~~}g_i\text{~~}) \\
P(g_i|g_{i'}) &= P(g_i)
\end{aligned}$$

In summary, through eliminating the dependencies among gradient entires, random interleaving renders packet erasure equivalent to random sparsification successfully. □

Assumption 1 (L-Smooth [2])**.** *The loss function f in clients' local updating is Lipschitz-smooth with a constant L and uniformly bounded by $f(\boldsymbol{w}^*)$, that is, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

REFERENCES

[1] Gilbert Strang, *Calculus*, vol. 1, SIAM, 1991.

[2] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, "On the convergence of FedAvg on Non-IID data," in *ICLR*, 2020.