# MPLS: Stacking Diverse Layers into One Model for Decentralized Federated Learning

## A Convergence Proof of MPLS

The convergence proof of MPLS is summarized as follows,

**Assumption 1** *We first make some widely used assumptions [1, 2, 4] as follows:*

1) *Lipschitzian gradient.* The loss function $f_i$ is with $\mathcal{L}$-Lipschitzian gradients, *i.e.*,

$$\|\nabla f_i(w_1) - \nabla f_i(w_2)\| \leq \mathcal{L}\|w_1 - w_2\|, \forall w_1, w_2, i \tag{1}$$

where $\|\cdot\|$ is the vector $L_2$ normal.

2) *Connected topology.* The network topology $G$ is a connected topology.

3) *Unbiased estimation.* For worker $i$, we have the expectation for the stochastic gradient of $F$ on each data sample is equal to $f_i(w)$, *i.e.*,

$$\mathbb{E}_{\xi \in D_i} \nabla F(w; \xi) = \nabla f_i(w) \tag{2}$$

For all workers, we further assume that,

$$\mathbb{E}_{i \in V} \mathbb{E}_{\xi \in D_i} \nabla F(w; \xi) = \nabla f(w) \tag{3}$$

4) *Bounded variance.* Assume the variance of stochastic gradients is bounded, *i.e.*,

$$\mathbb{E}_{\xi \in D_i} \|\nabla F(w; \xi) - \nabla f_i(w)\|^2 \leq \sigma^2, \forall i, w \tag{4}$$

$$\mathbb{E}_{i \in V} \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \varsigma^2, \forall w \tag{5}$$

We also define $\epsilon^2$ as the upper bound of the variance between the local model and global model [3], *i.e.*,

$$\mathbb{E}_{i \in V} \|w^t - w_i^t\|^2 \leq \epsilon^2, \forall t, i. \tag{6}$$

We prove the convergence of MPLS in three steps. Firstly, we derive the convergence bound of the global model after an arbitrary worker $i$ has performed local training, *i.e.*, $\mathbb{E}[f(\overline{w}^t) - f(w^*)]$, where $\overline{w}^t$ is the global model after one worker has performed local training in epoch $t$. Secondly, we study the divergence between two loss values of the global models before and after model aggregation, *i.e.*, $\mathbb{E}[f(w^{t+1}) - f(\overline{w}^t)]$. Using the above two bounds, we finally obtain the average gradient after $T$ training epochs, *i.e.*, $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(w^t)\|^2$.

*Proof.* To express the relationship between the combined model and local model, we adopt an upper bound $\alpha^2$ of $\|\widetilde{w}_i^t - \overline{w}_i^t\|^2$, *i.e.*, $\|\widetilde{w}_i^t - \overline{w}_i^t\|^2 \leq \alpha^2, \forall t, i$. We also define $\|\nabla f(w^t)\|^2 - \|\nabla f(\overline{w}^t)\|^2 \leq \beta^2, \forall t$. The proof is as follows,

**Step 1:** Inspired by previous work [2], we can obtain the bound $\mathbb{E}[f(\overline{w}^t) - f(w^*)]$ by replacing the parameter in [2] (*i.e.*, $\hat{M}_k$) with $\epsilon^2$,

$$\begin{aligned}&\mathbb{E}[f(\overline{w}^t) - f(w^*)]\\&\leq \mathbb{E}[f(w^t) - f(w^*)] - \frac{\eta M}{2N}\mathbb{E}\|\nabla f(w^t)\|^2 + \lambda\end{aligned} \tag{7}$$

where $\lambda = \frac{\eta\tau\epsilon^2 M\mathcal{L}^2}{N} + \frac{6\eta^2\tau^2\epsilon^2\mathcal{L}^3M^2}{N^2} + \frac{12M^3\mathcal{L}^4\tilde{T}^2\eta^3\tau^3\epsilon^2}{N^3} + \frac{\eta^2\mathcal{L}M\tau(\sigma^2+6\varsigma^2\tau M)}{2N^2} + \frac{\mathcal{L}^2\tilde{T}^2\eta^3\tau^2M^2(\sigma^2+6\varsigma^2\tau M)}{N^3}$, $M$ is the batch size and $\tilde{T}$ is the maximum staleness during training.

**Step 2:** Based on the model averaging in Eq. (**??**), we further study the divergence $\mathbb{E}[f(w^{t+1}) - f(\overline{w}^t)]$,

$$\begin{aligned}&\mathbb{E}[f(w^{t+1}) - f(\overline{w}^t)] = \mathbb{E}[f(\overline{w}^t - \frac{1}{2N}(\widetilde{w}_i^t - \overline{w}_i^t)) - f(\overline{w}^t)]\\&\leq -\frac{1}{2N}\mathbb{E} < \nabla f(\overline{w}^t), \widetilde{w}_i^t - \overline{w}_i^t > + \frac{\mathcal{L}}{8N^2}\|\widetilde{w}_i^t - \overline{w}_i^t\|^2\\&= -\frac{1}{4N}\{\|\nabla f(\overline{w}^t) + \widetilde{w}_i^t - \overline{w}_i^t\|^2 - \|\nabla f(\overline{w}^t)\|^2\\&\quad - \|\widetilde{w}_i^t - \overline{w}_i^t\|^2\} + \frac{\mathcal{L}}{8N^2}\|\widetilde{w}_i^t - \overline{w}_i^t\|^2\\&\leq \frac{1}{4N}\mathbb{E}\|\nabla f(w^t)\|^2 + \frac{\mathcal{L}+2N}{8N^2}\alpha^2 + \frac{\beta^2}{4N}\end{aligned} \tag{8}$$

**Step 3:** By adding Eq. (7) and Eq. (8), we obtain the convergence bound between two consecutive epochs,

$$\begin{aligned}&\mathbb{E}[f(w^{t+1}) - f(w^*)] = \mathbb{E}[f(w^{t+1}) - f(\overline{w}^t) + f(\overline{w}^t) - f(w^*)]\\&\leq \mathbb{E}[f(w^t) - f(w^*)] - \frac{2\eta M - 1}{4N}\mathbb{E}\|\nabla f(w^t)\|^2 + \lambda\\&\quad + \frac{\mathcal{L}+2N}{8N^2}\alpha^2 + \frac{\beta^2}{4N}\end{aligned} \tag{9}$$

Thus, by subtracting the first term of the RHS from both sides of Eq. (9), we have

$$\begin{aligned}\mathbb{E}[f(w^{t+1}) - f(w^t)] \leq &-\frac{2\eta M - 1}{4N}\mathbb{E}\|\nabla f(w^t)\|^2\\&+ \lambda + \frac{\mathcal{L}+2N}{8N^2}\alpha^2 + \frac{\beta^2}{4N}\end{aligned} \tag{10}$$

We further sum the results in Eq. (10) from $t = 0$ to $t = T - 1$, and obtain

$$\begin{aligned}&\sum_{t=0}^{T-1}\mathbb{E}[f(w^{t+1}) - f(w^t)] = \mathbb{E}[f(w^{t+1}) - f(w^0)]\\&= -\frac{2\eta M - 1}{4N}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 + T(\lambda + \frac{\mathcal{L}+2N}{8N^2}\alpha^2 + \frac{\beta^2}{4N})\end{aligned}$$

By rearranging the above result, we complete the proof as shown in Theorem 1.

**Theorem 1.** *While* $\frac{1}{2M} < \eta < \frac{1}{\mathcal{L}}$*, after* $T$ *training epochs, the average gradient* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2$ *is bounded by,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 \leq \frac{4(f(w^0)-f(w^*))}{T(2\eta M-1)/N} + \frac{4N}{2\eta M-1}(\lambda+\frac{\beta^2}{4N}+\frac{\mathcal{L}+2N}{8N^2}\alpha^2) \tag{11}$$

Given a small value of learning rate, *e.g.*, $0 < \eta < \frac{\sqrt{N}}{L\sqrt{T\tau}}$, the bound will eventually converge after a sufficiently large number of training epochs. Beside of these constants, we also observe that the convergence bound is related to two variables, *i.e.*, $\alpha^2$ and $\beta^2$, which is mainly determined by the combined model $\widetilde{w}_i^t$. Since layers that were not selected before will have a higher probability to be selected and aggregated in MPLS, the fairness of peer and layer selection as well as the generalization of the combined model can be enhanced and the values of two variables can be reduced, which helps to improve the convergence performance, compared with that of other asynchronous DML approaches (*e.g.*, NetMax).

## References


1. Lian, X., Zhang, C., Zhang, H., Hsieh, C.J., Zhang, W., Liu, J.: Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. arXiv preprint arXiv:1705.09056 (2017)
2. Lian, X., Zhang, W., Zhang, C., Liu, J.: Asynchronous decentralized parallel stochastic gradient descent. In: International Conference on Machine Learning. pp. 3043–3052. PMLR (2018)
3. Yu, H., Yang, S., Zhu, S.: Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5693–5700 (2019)
4. Zhou, P., Lin, Q., Loghin, D., Ooi, B.C., Wu, Y., Yu, H.: Communication-efficient decentralized machine learning over heterogeneous networks. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE). pp. 384–395. IEEE (2021)