1 Exponential Convergence of FedAvg with MaSS for Linear Regression Problem with 0 Loss

The MaSS algorithm is a variant of the Nesterov acceleration, where in the update a non-negative multiple of the gradient is added to the Nesterov parameter update, to correct for the "over-descent" of the Nesterov update. In [Liu&Belkin], the authors show that for stochastic gradient methods in the interpolation setting, Nesterov update is not able to achieve accleration over vanilla SGD, whereas the MaSS algorithm is able to achieve acceleration both theoretically and empirically. We adapt the MaSS algorithm to our federated learning setting, and prove that for certain class of convex objectives with 0 global objective value, the FedAvg with MaSS update achieves exponential convergence and acceleration over SGD.

We first show that FedAvg with MaSS has exponential convergence for linear regression when the global objective has 0 as its global minimum. Since the global loss is given by

$$F(w) = \sum_{k=1}^{N} p_k F_k(w)$$
$$F_k(w) = \frac{1}{2n_k} \sum_{j=1}^{n_k} (w^T x_{k,j} - z_{k,j})^2$$

and there exists w^* such that $F(w^*) = 0$, in particular this implies that $F_k^* = F_k(w^*) = F^* = 0$ for all k.

1.1 Notation and Definitions

Following Liu&Belkin and Jain et al., we define some condition number related quantities. Let $H^k = \frac{1}{n_k} \sum_{j=1}^{n_k} x_{k,j} x_{k,j}$ be the Hessian matrix of F_k. Let L^k and μ^k be the largest and smallest non-zero eigenvalues of the Hessians H^k. For a mini-batch $\{\tilde{x}_j\}_{j=1}^m$ of m samples from device k, let $\tilde{H}_m^k = \frac{1}{m} \sum_{j=1}^m \tilde{x}_j \tilde{x}_j^T$ be the unbiased mini-batch estimate of H^k. Let L_1^k be the smallest positive number such that

$$\mathbb{E}\left[\|\tilde{x}\|^2 \tilde{x}\tilde{x}^T\right] \preceq L_1^k H^k$$

and define

$$L_m^k = L_1^k/m + (m-1)L/m$$

Define the m-stochastic condition number as $\kappa_m^k := L_m^k/\mu$. Define the statistical condition number $\tilde{\kappa}^k$ as the smallest positive real number such that

$$\mathbb{E}\left[\langle \tilde{x}(H^k)^{-1}, \tilde{x}\rangle \tilde{x}\tilde{x}^T\right] \preceq \tilde{\kappa}^k H^k$$

Note that $L_m^k \geq L$, and $\kappa_m^k \geq \kappa^k$, while $\tilde{\kappa}^k \leq \kappa^1$.

1.2 FedAvg with MaSS

The FedAvg algorithm with MaSS follows the updates

$$w_{t+1}^k = \begin{cases} u_t^k - \eta_1^k g_{t,k} & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^N p_k \left[u_t^k - \eta_1^k g_{t,k}\right] & \text{if } t+1 \in \mathcal{I}_E \end{cases}$$
$$u_{t+1}^k = w_{t+1}^k + \gamma^k (w_{t+1}^k - w_t^k) + \eta_2^k g_{t,k}$$

where we note that the natural parameter is w_t, while u_t is an auxiliary parameter, which we initialize to be $u_0^k \equiv u_0$, and

$$g_{t,k} := \nabla F_k(w_t^k, \xi_t^k)$$

is the stochastic gradient and

$$g_t = \sum_{k=1}^N p_k g_{t,k}$$

is the averaged stochastic gradient. When $\eta_2^k \equiv 0$, this reduces to the FedAvg algorithm with Nesterov updates.

We note that the update can equivalently be written as

$$v_{t+1}^k = (1-\alpha^k) v_t^k + \alpha^k w_t^k - \delta^k g_{t,k}$$
$$w_{t+1}^k = \begin{cases} u_t^k - \eta^k g_{t,k} & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^N p_k \left[u_t^k - \eta^k g_{t,k}\right] & \text{if } t+1 \in \mathcal{I}_E \end{cases}$$
$$u_{t+1}^k = \frac{\alpha^k}{1+\alpha^k} v_{t+1}^k + \frac{1}{1+\alpha^k} w_{t+1}^k$$

where there is a bijection between the parameters

$$\frac{1-\alpha^k}{1+\alpha^k} = \gamma^k$$
$$\eta^k = \eta_1^k$$
$$\frac{\eta^k - \alpha^k \delta^k}{1+\alpha^k} = \eta_2^k$$

and we further introduce an auxiliary parameter v_t^k, which is initialized at $v_0^k \equiv v_0$. We also note that when $\delta_t = \frac{\eta_t}{\alpha_t}$, the update reduces to the Nesterov accelerated SGD. This version of the FedAvg with MaSS algorithm is used for analyzing the exponential convergence.

As before, define the virtual sequences $\overline{w}_t = \sum_{k=1}^N p_k w_t^k$, $\overline{v}_t = \sum_{k=1}^N p_k v_t^k$, $\overline{u}_t = \sum_{k=1}^N p_k u_t^k$, and $\overline{g}_t = \sum_{k=1}^N p_k \mathbb{E} g_{t,k}$. We have $\mathbb{E} g_t = \overline{g}_t$ and $\overline{w}_{t+1} = \overline{u}_t - \eta_t g_t$, $\overline{v}_{t+1} = (1-\alpha^k)\overline{v}_t + \alpha^k \overline{w}_t - \delta^k g_t$, and $\overline{w}_{t+1} = \frac{\alpha^k}{1+\alpha^k}\overline{v}_{t+1} + \frac{1}{1+\alpha^k}\overline{y}_{t+1}$.

For the linear regression problem in the interpolation setting, we can write

$$\begin{aligned} F(w) &= \frac{1}{2}(w-w^*)^T H(w-w^*) \\ &= \frac{1}{2}\|w-w^*\|_H^2 \end{aligned}$$

and similarly $F^k(w) = \frac{1}{2}\|w-w^*\|_{H^k}^2$, so that

$$\begin{aligned} g_{t,k} &= \tilde{H}_t^k(w_t^k - w^*) \\ g_t &= \sum_{k=1}^N p_k \tilde{H}_t^k(w_t^k - w^*) \end{aligned}$$

1.3 Exponential Convergence when Global Loss is 0

We now present the exponential convergence result using FedAvg with MaSS updates. On each device, local data is stored and mini-batch gradient descent with batch size m^k is performed. We allow for varying batch sizes across devices.

Theorem 1. *(FedAvg with MaSS, Full Participation) Let the hyperparameters satisfy the requirements of [Liu&Belkin]. More precisely, let $\tilde{\kappa}_m := \tilde{\kappa}/m^k + (m^k-1)/m^k$, and let the hyper parameters satisfy*

$$\eta^k(m) = \frac{1}{L_m}, \alpha^k(m) = \frac{1}{\sqrt{\kappa_m^k \tilde{\kappa}_m^k}}, \delta^k(m) = \frac{\eta^k}{\alpha^k \tilde{\kappa}_m^k}$$

Let $\mathcal{I}_E = \ell E$ where $\ell \in \mathbb{N}$ be the set of communication rounds, then the updates

$$\begin{aligned} v_{t+1}^k &= (1-\alpha^k)v_t^k + \alpha^k w_t^k - \delta^k g_{t,k} \\ w_{t+1}^k &= \begin{cases} u_t^k - \eta^k g_{t,k} & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^N p_k \left[u_t^k - \eta^k g_{t,k}\right] & \text{if } t+1 \in \mathcal{I}_E \end{cases} \\ u_{t+1}^k &= \frac{\alpha^k}{1+\alpha^k} v_{t+1}^k + \frac{1}{1+\alpha^k} w_{t+1}^k \end{aligned}$$

where $g_{t,k}$ are the mini-batch stochastic gradients achieves the convergence

$$\|\overline{w}_t - w^*\|^2 \leq C \cdot (1 - 1/\sqrt{\kappa_m}$$

Proof. Note that at each communication round we update the w_{t+1}^k parameters to be the average across devices while fixing v_{t+1}^k. This automatically adjusts the u_{t+1}^k parameter at each device by the relation

$$u_{t+1}^k = \frac{\alpha^k}{1+\alpha^k} v_{t+1}^k + \frac{1}{1+\alpha^k} w_{t+1}^k$$

valid for all $t \geq 0$.

For now assume $\delta^k = \delta$ and $\alpha^k = \alpha$ is equal across devices. Need to relax later.

Theorems 2 and 3 of the Liu&Belkin paper gives the bound

$$\mathbb{E}\left[\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|u_E^k - \eta^k g_{E,k} - w^*\|^2\right] \leq (1-\alpha^k)^E(\|v_0^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_0^k - w^*\|^2)$$

for all k, where E is the first communication round. Note that $w_E^k = \overline{w}_E^k \neq u_E^k - \eta^k g_{t,k}$.

It follows from convexity that

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^N p_k\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_E - w^*\|^2\right] &\leq \sum_{k=1}^N p_k\mathbb{E}\left[\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|u_E^k - \eta^k g_{E,k} - w^*\|^2\right] \\ &\leq \sum_{k=1}^N p_k(1-\alpha^k)^E(\|v_0^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_0^k - w^*\|^2)\end{aligned}$$

Since $w_E^k = \overline{w}_E$ for all devices, applying the per-device result again starting at $t = E$ instead of $t = 0$, for each device we have the bound

$$\begin{aligned}\mathbb{E}\left[\|v_{2E}^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|u_{2E}^k - \eta^k g_{2E,k} - w^*\|^2\right] &\leq (1-\alpha^k)^E\mathbb{E}(\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_E^k - w^*\|^2) \\ &= (1-\alpha^k)^E\mathbb{E}(\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_E - w^*\|^2)\end{aligned}$$

Here we emphasize that w_E^k results from broadcasting and so is the same across all devices, while v_E^k remains distinct on each device (and is only auxiliary). Then by convexity and summing the above inequalities across devices we have

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^N p_k\|v_{2E}^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_{2E} - w^*\|^2\right] &\leq \sum_{k=1}^N p_k\mathbb{E}\left[\|v_{2E}^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|u_{2E}^k - \eta^k g_{2E,k} - w^*\|^2\right] \\ &\leq \sum_{k=1}^N p_k(1-\alpha^k)^E\mathbb{E}(\|v_E^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_E - w^*\|^2) \\ &\leq \sum_{k=1}^N p_k(1-\alpha^k)^{2E}(\|v_0^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_0^k - w^*\|^2)\end{aligned}$$

and by induction we can show that

$$\mathbb{E}\left[\sum_{k=1}^N p_k\|v_{\ell E}^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_{\ell E} - w^*\|^2\right] \leq \sum_{k=1}^N p_k(1-\alpha^k)^{\ell E}(\|v_0^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_0^k - w^*\|^2)$$

and more generally

$$\mathbb{E}\left[\sum_{k=1}^N p_k\|v_t^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|\overline{w}_t - w^*\|^2\right] \leq \sum_{k=1}^N p_k(1-\alpha^k)^t(\|v_0^k - w^*\|^2_{H_k^{-1}} + \frac{\delta^k}{\alpha^k}\|w_0^k - w^*\|^2)$$

In particular, this implies

$$\mathbb{E}\|\overline{w}_t - w^*\|^2 \leq C \cdot \sum_{k=1}^{N} p_k (1-\alpha^k)^t$$

which is exponential convergence. □