Appendix

Roadmap: In Appendix A, we list several probability results. In Appendix B we prove our convergence result of FL-NTK. In Appendix C, we prove our generalization result of FL-NTK.

A. Probability Tools

Lemma A.1 (Bernstein inequality (Bernstein, 1924)). *Let $X_1, \dots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all $i \in [n]$. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^{n} X_i \leq t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^{n} \mathbb{E}[X_j^2] + Mt/3}\right).$$

Lemma A.2 (Anti-concentration inequality of Gaussian). *Let $X \sim N(0, \sigma^2)$, then for any $0 < t \leq \sigma$*

$$\Pr\left[|X| \leq t\right] \in \left(\frac{2t}{3\sigma}, \sqrt{\frac{2}{\pi}} \cdot \frac{t}{\sigma}\right).$$

Proof. For completeness, we provide a short proof. Since $X \sim N(0, \delta^2)$, the CDF of X^2 is $\Pr[X^2 \leq t^2] = \frac{\gamma(1/2, t^2/2\sigma^2)}{\Gamma(1/2)}$ where $\gamma(\cdot, \cdot)$ is the incomplete lower gamma function. This can be further simplified to $\Pr[X^2 \leq t^2] = \mathrm{erf}(\sqrt{t^2/2\sigma^2})$ where erf is the error function. For $z \leq 1$, we can sandwich the erf function by $2z/3 \leq \mathrm{erf}(z/\sqrt{2}) \leq \sqrt{2/\pi} z$, thus letting $z = t/\sigma$ complete the proof.

□

B. Convergence of Neural Networks in Federated Learning

This section is organized as follows:

- In Appendix B.1, we introduce some definitions.
- In Appendix B.2, we present the convergence result of FL-NTK.
- In Appendix B.3, we upper bound C_1, C_2, C_3, C_4 that appear in the proof.
- In Appendix B.4, we present the property at initialization of FL-NTK.
- In Appendix B.5, we show the properties of local steps.
- In Appendix B.6, we present several technical claims used in the proof.

B.1. Definitions

Definition B.1. *We let κ to denote the condition number of Gram matrix $H(0)$.*

Assumption B.2. *We assume $\|x_i\|_2 = 1$ and $\lambda = \lambda_{\min}(H(0)) \in (0, 1]$.*

B.2. Convergence Result

Theorem B.3. *Recall that $\lambda = \lambda_{\min}(H(0)) > 0$. Let $m = \Omega(\lambda^{-4} n^4 \log(n/\delta))$, we iid initialize $u_r(0) \sim \mathcal{N}(0, I)$, a_r sampled from $\{-1, +1\}$ uniformly at random for $r \in [m]$, and we set the step size $\eta_{\mathrm{local}} = O(\lambda/(\kappa K n^2))$ and $\eta_{\mathrm{global}} = O(1)$, then with probability at least $1 - \delta$ over the random initialization we have for $t = 0, 1, 2, \cdots$*

$$\|y(t) - y\|_2^2 \leq (1 - \frac{\eta_{\mathrm{global}} \eta_{\mathrm{local}} \lambda K}{2N})^t \cdot \|y(0) - y\|_2^2. \tag{6}$$

Notation	Dimension	Meaning
N	$\mathbb{N}$	#clients
c	$[N]$	its index
T	$\mathbb{N}$	#communication rounds
t	$[T]$	its index
K	$\mathbb{N}$	#local update steps
k	$[K]$	its index
$y(t)$	$\mathbb{R}^n$	aggregated server model after global round t
y_c	$\mathbb{R}^{\lvert S_c\rvert}$	ground truth of c-th client
$y_c^{(k)}(t)$	$\mathbb{R}^{\lvert S_c\rvert}$	c-th client's model in global round t and local step k
$y^{(k)}(t)$	$\mathbb{R}^n$	all client's model in global round t and local step k
$w_{k,c}(t)$	$\mathbb{R}^{d\times m}$	c-th client's model parameter in global round t and local step k
$u(t)$	$\mathbb{R}^{d\times m}$	aggregated server model parameter in global round t and local step k

Table 1: Summary of several notations

Proof. We prove by induction. The base case is $t = 0$ and it is trivially true. Assume for $\tau = 0, \cdots, t$ we have proved Eq. (6) to be true. We show Eq. (6) holds for $\tau = t + 1$.

Recall that the set $Q_i \subset [m]$ is defined as follow

$$Q_i := \left\{r \in [m] : \forall w \in \mathbb{R}^d \text{ s.t. } \|w - w_r(0)\|_2 \le R,\ \ \mathbf{1}_{w_r(0)^\top x_i \ge 0} = \mathbf{1}_{w^\top x_i \ge 0}\right\},$$

and $\overline{Q}_i$ denotes its complement.

Let $v_{1,i}, v_{2,i}$ be defined as follows

$$v_{1,i} = \frac{1}{\sqrt{m}} \sum_{r\in Q_i} a_r \bigg(\phi\big((u_r(t) + \eta_{\text{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i)\bigg),$$
$$v_{2,i} = \frac{1}{\sqrt{m}} \sum_{r\in \overline{Q}_i} a_r \bigg(\phi\big((u_r(t) + \eta_{\text{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i)\bigg).$$

Let $H(t,k,c)_{i,j}, H(t,k,c)^\perp_{i,j}$ be defined as follows

$$H(t,k,c)_{i,j} = \frac{1}{m}\sum_{r=1}^m x_i^\top x_j \mathbf{1}_{u_r^\top x_i \ge 0, w_{k,c,r}(t)^\top x_j \ge 0},$$
$$H(t,k,c)^\perp_{i,j} = \frac{1}{m}\sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \ge 0, w_{k,c,r}(t)^\top x_j \ge 0}.$$

Define $H(t)$ and $H(t)^\perp \in \mathbb{R}^{n\times n}$ as

$$H(t)_{i,j} = \frac{1}{m}\sum_{r=1}^m x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \ge 0, u_r(t)^\top x_j \ge 0},$$
$$H(t)^\perp_{i,j} = \frac{1}{m}\sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \ge 0, u_r(t)^\top x_j \ge 0}.$$

Let $y_c^{(k)}(t)_j$ ($j \in S_c$) be defined by

$$y_c^{(k)}(t)_j = f(w_{k,c}(t), x_j).$$

We can write $\Delta u_r(t)$ as follow

$$\Delta u_r(t) = \frac{a_r}{N} \sum_{c\in[N]} \sum_{k\in[K]} \frac{\eta_{\text{local}}}{\sqrt{m}} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j) x_j \mathbf{1}_{w_{k,c,r}(t)^\top x_j \geq 0}.$$

Thus we have

$$\begin{aligned} v_{1,i} &= \frac{\eta_{\text{global}}\eta_{\text{local}}}{mN} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j) x_i^\top x_j \sum_{r\in Q_i} \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \\ &= \frac{\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^\perp_{i,j}). \end{aligned}$$

We can therefore write $-2(y - y(t))^\top (y(t+1) - y(t))$ as follow

$$\begin{aligned} & -2(y - y(t))^\top (y(t+1) - y(t)) \\ = & -2(y - y(t))^\top (v_1 + v_2) \\ = & -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^\perp_{i,j}) \\ & -2\sum_{i\in[n]} (y_i - y_i(t)) v_{2,i}. \end{aligned}$$

Let

$$\begin{aligned} C_1 = & -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)_{i,j} \\ C_2 = & \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)^\perp_{i,j} \\ C_3 = & -2\sum_{i\in[n]} (y_i - y_i(t)) v_{2,i} \\ C_4 = & \|y(t+1) - y(t)\|_2^2. \end{aligned}$$

Then

$$\begin{aligned} & \|y - y(t+1)\|_2^2 \\ = & \|y - y(t)\|_2^2 - 2(y - y(t))^\top (y(t+1) - y(t)) + \|y(t+1) - y(t)\|_2^2 \\ = & \|y - y(t)\|_2^2 + C_1 + C_2 + C_3 + C_4. \end{aligned}$$

By Claim B.4, Claim B.5, Claim B.6 and Claim B.7 we have

$$\begin{aligned} \|y - y(t+1)\|_2^2 \leq & \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}(-K\lambda + 4nRK(1 + 2\eta_{\text{local}}Kn) + 2\eta_{\text{local}}\kappa\lambda K^2 n)\|y - y(y)\|_2^2 \\ & + \frac{16\eta_{\text{global}}\eta_{\text{local}}}{N} K(1 + 2\eta_{\text{local}}Kn)nR\|y - y(y)\|_2^2 \\ & + \frac{16\eta_{\text{global}}\eta_{\text{local}}}{N} K(1 + 2\eta_{\text{local}}Kn)nR\|y - y(y)\|_2^2 \\ & + \frac{4\eta_{\text{global}}^2\eta_{\text{local}}^2 n^2K^2(1 + 2\eta_{\text{local}}Kn)^2}{N^2}\|y - y(y)\|_2^2. \end{aligned}$$

By the choice of $\eta_{\text{local}} \leq \frac{\lambda}{1000\kappa n^2 K}$ and $\eta_{\text{local}}\eta_{\text{global}} \leq \frac{\lambda}{1000\kappa n^2 K}$ and $R \leq \lambda/(1000n)$ we come to

$$\|y - y(t+1)\|_2^2 \leq \|y - y(t)\|_2^2$$

$$
\begin{aligned}
&- \frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y-y(t)\|_2^2 \qquad (7)\\
&+ 40\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y-y(t)\|_2^2\\
&+ 40\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y-y(t)\|_2^2\\
&+ \frac{\eta_{\text{local}}^2\eta_{\text{global}}^2n^2K^2}{N^2}\|y-y(t)\|_2^2\\
\le &\|y-y(t)\|_2^2\\
&- (1-1/10)\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y-y(t)\|_2^2\\
&+ 80\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y-y(t)\|_2^2\\
\le &\|y-y(t)\|_2^2 - \frac{1}{2}\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y-y(t)\|_2^2 \qquad (8)
\end{aligned}
$$

where the second step follows from $\eta_{\text{local}} \le \frac{\lambda}{1000\kappa n^2 K}$, the third step follows from $R \le \lambda/(1000n)$. □

B.3. Bounding C_1, C_2, C_3, C_4

Claim B.4. *We have with probability at least* $1 - n^2 \cdot \exp(-mR/10)$ *over random initialization*

$$
C_1 \le \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\|y-y(t)\|_2^2(-K\lambda + 4nRK(1+2\eta_{\text{local}}Kn) + 2\eta_{\text{local}}\kappa\lambda K^2 n).
$$

Proof. We first calculate

$$
\begin{aligned}
&\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)H(t,k,c)_{i,j}\\
=&\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(0)_{i,j})\\
&+\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j(t) - y_c^{(k)}(t)_j)H(0)_{i,j}\\
&+K\sum_{i\in[n]}\sum_{j\in[n]}(y_i - y_i(t))(y_j - y_j(t))H(0)_{i,j}.
\end{aligned}
$$

From Lemma B.12 and Lemma B.9 we have $\|u_r(t) - u(0)\|_2 \le R$ and $\|w_{k,c,r}(t) - u(0)\|_2 \le R$. Let $H(t,k)$ be defined by

$$
H(t,k)_{i,j} = H(t,k,c)_{i,j}
$$

for $j \in S_c$. Then from Lemma B.11 we obtain

$$
\|H(t,k) - H(0)\|_F \le 2nR
$$

with probability at least $1 - n^2 \cdot \exp(-mR/10)$ over random initialization.

Therefore from direct calculations we have

$$
\begin{aligned}
&\left|\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(0)_{i,j})\right|\\
=&\sum_{k\in[K]}(y-y(t))^\top(H(t,k) - H(0))(y - y^{(k)}(t))\\
\le&\sum_{k\in[K]}\|y-y(t)\|_2\|y-y^{(k)}(t)\|_2\|H(t,k) - H(0)\|_F
\end{aligned}
$$

$$\leq 4nRK(1+2\eta_{\text{local}}Kn)\|y-y(t)\|_2^2.$$

where the last step comes from Eq (15).

By Lemma B.10 we have

$$\left|\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i-y_i(t))(y_j(t)-y_c^{(k)}(t)_j)H(0)_{i,j}\right| \leq \sum_{k\in[K]}\|y-y(t)\|_2\|H(0)\|\|y(t)-y^{(k)}(t)\|_2$$
$$\leq 2\eta_{\text{local}}\kappa\lambda K^2 n\|y-y(t)\|_2^2.$$

Finally we have

$$K\sum_{i\in[n]}\sum_{j\in[n]}(y_i-y_i(t))(y_j-y_j(t))H(0)_{i,j}\geq K\lambda\|y-y(t)\|_2^2.$$

Combining the above we conclude the proof with

$$\begin{aligned}
&C_1\\
=&-\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i-y_i(t))(y_j-y_c^{(k)}(t)_j)H(t,k,c)_{i,j}\\
\leq&-\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}(-4nRK(1+2\eta_{\text{local}}K^2n)\|y-y(t)\|_2^2+K\lambda\|y-y(t)\|_2^2-2\eta_{\text{local}}\kappa\lambda K^2n\|y-y(t)\|_2^2)\\
\leq&\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\|y-y(t)\|_2^2(-K\lambda+4nRK(1+2\eta_{\text{local}}K^2n)+2\eta_{\text{local}}\kappa\lambda K^2n).
\end{aligned}$$

□

Claim B.5. *The following holds with probability at least* $1-n\exp(-mR)$ *over random initialization*

$$C_2\leq\frac{16\eta_{\text{global}}\eta_{\text{local}}}{N}K(1+2\eta_{\text{local}}nK)nR\|y-y(t)\|_2^2.$$

Proof. We define matrix $H(t,k)^\perp\in\mathbb{R}^{n\times n}$ such that $H(t,k)^\perp_{i,j}=H(t,k,c)^\perp_{i,j}, j\in S_c$. Notice that

$$\begin{aligned}
C_2&=\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i-y_i(t))(y_j-y_c^{(k)}(t)_j)H(t,k,c)^\perp_{i,j}\\
&=\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{k\in[K]}(y-y(t))^\top H(t,k)^\perp(y-y^{(k)}(t))\\
&\leq\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{k\in[K]}\|y-y(t)\|_2\|H(t,k)^\perp\|_F\|y-y^{(k)}(t)\|_2\\
&\leq\frac{4\eta_{\text{global}}\eta_{\text{local}}}{N}K(1+2\eta_{\text{local}}nK)\|y-y(t)\|_2^2\|H(t,k)^\perp\|_F
\end{aligned}$$

where the last step comes from Eq (15).

It thus suffices to upper bound $\|H(t,k)^\perp\|_F$.

For each $i\in[n]$, we define ζ_i as follows

$$\zeta_i=\sum_{r=1}^m\mathbf{1}_{r\in\overline{Q}_i}.$$

It then follows from direct calculations that

$$\|H(t,k)^\perp\|_F^2=\sum_{i=1}^n\sum_{j=1}^n(H(t,k)^\perp_{i,j})^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \sum_{c\in[N]} \sum_{j\in S_c} \Big(\frac{1}{m} \sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \Big)^2 \\
&= \sum_{i=1}^{n} \sum_{c\in[N]} \sum_{j\in S_c} \Big(\frac{1}{m} \sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \cdot \mathbf{1}_{r\in\overline{Q}_i} \Big)^2 \\
&= \sum_{i=1}^{n} \sum_{c\in[N]} \sum_{j\in S_c} (\frac{x_i^\top x_j}{m})^2 \Big(\sum_{r=1}^{m} \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \cdot \mathbf{1}_{r\in\overline{Q}_i} \Big)^2 \\
&\leq \frac{1}{m^2} \sum_{i=1}^{n} \sum_{c\in[N]} \sum_{j\in S_c} \Big(\sum_{r=1}^{m} \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \cdot \mathbf{1}_{r\in\overline{Q}_i} \Big)^2 \\
&= \frac{n}{m^2} \sum_{i=1}^{n} \Big(\sum_{r=1}^{m} \mathbf{1}_{r\in\overline{Q}_i} \Big)^2 \\
&= \frac{n}{m^2} \sum_{i=1}^{n} \zeta_i^2.
\end{aligned}
$$

Fix $i \in [n]$. The plan is to use Bernstein inequality to upper bound ζ_i with high probability.

First by Eq. (10) we have $\mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}] \leq R$. We also have

$$
\begin{aligned}
\mathbb{E}\Big[(\mathbf{1}_{r\in\overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}])^2\Big] &= \mathbb{E}[\mathbf{1}^2_{r\in\overline{Q}_i}] - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}]^2 \\
&\leq \mathbb{E}[\mathbf{1}^2_{r\in\overline{Q}_i}] \\
&\leq R.
\end{aligned}
$$

Finally we have $|\mathbf{1}_{r\in\overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}]| \leq 1$.

Notice that $\{\mathbf{1}_{r\in\overline{Q}_i}\}_{r=1}^m$ are mutually independent, since $\mathbf{1}_{r\in\overline{Q}_i}$ only depends on $w_r(0)$. Hence from Bernstein inequality (Lemma A.1) we have for all $t > 0$,

$$
\Pr\left[\zeta_i > m \cdot R + t\right] \leq \exp\left(-\frac{t^2/2}{m \cdot R + t/3}\right).
$$

By setting $t = 3mR$, we have

$$
\Pr\left[\zeta_i > 4mR\right] \leq \exp(-mR). \tag{9}
$$

Hence by union bound, with probability at least $1 - n\exp(-mR)$,

$$
\|H(t,k)^\perp\|_F^2 \leq \frac{n}{m^2} \cdot n \cdot (4mR)^2 = 16n^2R^2.
$$

Putting all together we have

$$
\|H(t,k)^\perp\|_F \leq 4nR
$$

with probability at least $1 - n\exp(-mR)$ over random initialization. □

Claim B.6. *With probability at least* $1 - n\exp(-mR)$ *over random initialization the following holds*

$$
C_3 \leq \frac{16\eta_{\text{global}}\eta_{\text{local}}K}{N}(1 + 2\eta_{\text{local}}nK)nR\|y - y(t)\|_2^2
$$

Proof. We can upper bound $\|v_2\|_2$ in the following sense

$$
\|v_2\|_2^2 \leq \sum_{i=1}^{n} \left(\frac{\eta_{\text{global}}}{\sqrt{m}} \sum_{r\in\overline{Q}_i} \left|\Delta u_r(t)^\top x_i\right| \right)^2
$$

$$= \frac{\eta_{\text{global}}^2}{m} \sum_{i=1}^{n} \Big(\sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i} |\Delta u_r(t)^\top x_i| \Big)^2$$
$$\leq \frac{\eta_{\text{global}}^2 \eta_{\text{local}}^2}{m} \cdot \Big(\frac{2K(1+2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2 \Big)^2 \cdot \sum_{i=1}^{n} \left(\sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i} \right)^2$$

where the last step comes from Lemma B.10.

It is previously shown that $\sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i} \leq 4mR$ holds with probability at least $1 - n\exp(-mR)$ over random initialization, thus with probability at least $1 - n\exp(-mR)$ over random initialization

$$\|v_2\|_2^2 \leq \frac{\eta_{\text{global}}^2 \eta_{\text{local}}^2}{m} \cdot \frac{4K^2(1+2\eta_{\text{local}} nK)^2 n}{N^2 m} \|y - y(t)\|_2^2 \cdot n(4mR)^2$$
$$\leq \Big(\frac{8\eta_{\text{global}} \eta_{\text{local}} K}{N} (1 + 2\eta_{\text{local}} nK) nR \|y - y(t)\| \Big)^2.$$

Using Cauchy-Schwarz inequality, we complete the proof with

$$C_3 = -2 \sum_{i \in [n]} (y_i - y_i(t)) v_{2,i}$$
$$\leq 2\|y - y(t)\|_2 \cdot \|v_2\|_2$$
$$\leq \frac{16\eta_{\text{global}} \eta_{\text{local}} K}{N} (1 + 2\eta_{\text{local}} nK) nR \|y - y(t)\|_2^2.$$

□

Claim B.7. *We have*

$$C_4 \leq \frac{4\eta_{\text{local}}^2 \eta_{\text{global}}^2 n^2 K^2 (1 + 2\eta_{\text{local}} nK)^2}{N^2} \|y - y(t)\|_2^2.$$

Proof. Recall that $y(t+1) - y(t) = v_1 + v_2$, we have

$$\|y(t+1) - y(t)\|_2^2 \leq \sum_{i=1}^{n} \left(\frac{\eta_{\text{global}}}{\sqrt{m}} \sum_{r=1}^{m} |\Delta u_r(t)^\top x_i| \right)^2$$
$$= \frac{\eta_{\text{global}}^2}{m} \sum_{i=1}^{n} \Big(\sum_{r=1}^{m} |\Delta u_r(t)^\top x_i| \Big)^2$$
$$\leq \frac{\eta_{\text{global}}^2 \eta_{\text{local}}^2}{m} \cdot \Big(\frac{2K(1+2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2 \Big)^2 \cdot nm^2$$
$$\leq \frac{4\eta_{\text{local}}^2 \eta_{\text{global}}^2 n^2 K^2 (1 + 2\eta_{\text{local}} nK)^2}{N^2} \|y - y(t)\|_2^2$$

where the penultimate step comes from Lemma B.10. □

B.4. Random Initialization

Lemma B.8. *Let events E_1, E_2, E_3 be defined as follows*

$$E_1 = \Big\{ \phi(w_r(0)^\top x_i) \leq \sqrt{2\log(6mn/\delta)}, \forall r \in [m], \forall i \in [n] \Big\}$$
$$E_2 = \Big\{ \Big| \sum_{r=1}^{m} \frac{1}{\sqrt{m}} a_r \phi(w_r(0)^\top x_i) \mathbf{1}_{w_r(0)^\top x_i \leq \sqrt{2\log(6mn/\delta)}} \Big| \leq \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta), \forall i \in [n] \Big\}$$
$$E_3 = \Big\{ \sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i} \leq 4mR, \forall i \in [n] \Big\}.$$

Then $E_1 \cap E_2 \cap E_3$ is true with probability at least $1-\delta$ over the random initialization. Furthermore given $E_1 \cap E_2 \cap E_3$ the following holds

$$\|y - y(0)\|_2^2 = O(n \log(m/\delta) \log^2(n/\delta)).$$

Proof. First we bound $\Pr[\neg E_3]$. For each $i \in [n]$, we define ζ_i as follows

$$\zeta_i = \sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i}.$$

We use w as shorthand for $w(0)$. Define the event

$$A_{i,r} = \left\{ \exists u : \|u - w_r\|_2 \leq R, \mathbf{1}_{x_i^\top w_r \geq 0} \neq \mathbf{1}_{x_i^\top u \geq 0} \right\}.$$

Note this event happens if and only if $|w_r^\top x_i| < R$. Recall that $w_r \sim \mathcal{N}(0, I)$. By anti-concentration inequality of Gaussian (Lemma A.2), we have

$$\Pr[A_{i,r}] = \Pr_{z \sim \mathcal{N}(0,1)}[|z| < R] \leq \frac{2R}{\sqrt{2\pi}}. \tag{10}$$

It thus follows from Eq. (10) that $\mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}] \leq R$. We also have

$$\begin{aligned} \mathbb{E}\left[(\mathbf{1}_{r \in \overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}])^2 \right] &= \mathbb{E}[\mathbf{1}^2_{r \in \overline{Q}_i}] - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}]^2 \\ &\leq \mathbb{E}[\mathbf{1}^2_{r \in \overline{Q}_i}] \\ &\leq R. \end{aligned}$$

Therefore $|\mathbf{1}_{r \in \overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}]| \leq 1$.

Notice that $\{\mathbf{1}_{r \in \overline{Q}_i}\}_{r=1}^m$ are mutually independent, since $\mathbf{1}_{r \in \overline{Q}_i}$ only depends on w_r. Hence from Bernstein inequality (Lemma A.1) we have for all $t > 0$,

$$\Pr[\zeta_i > m \cdot R + t] \leq \exp\left(-\frac{t^2/2}{m \cdot R + t/3} \right).$$

By setting $t = 3mR$, we have

$$\Pr[\zeta_i > 4mR] \leq \exp(-mR). \tag{11}$$

Taking union bound and note the choice of R and m we have

$$\Pr[\neg E_3] \leq n \exp(-mR) \leq \delta/3.$$

Next we bound $\Pr[\neg E_1]$. Fix $r \in [m]$ and $i \in [n]$. Since $w_r \sim \mathcal{N}(0, I)$ and $\|x_i\|_2 = 1$, $w_r^\top x_i$ follows distribution $\mathcal{N}(0, 1)$. From concentration of Gaussian distribution, we have

$$\Pr_{w_r}[w_r^\top x_i \geq \sqrt{2 \log(6mn/\delta)}] \leq \frac{\delta}{6mn}.$$

Let E_1 be the event that for all $r \in [m]$ and $i \in [n]$ we have $\phi(w_r^\top x_i) \leq \sqrt{2 \log(6mn/\delta)}$. Then by union bound, $\Pr[\neg E_1] \leq \frac{\delta}{3}$,

Finally we bound $\Pr[\neg E_2]$. Fix $i \in [n]$. For every $r \in [m]$, we define random variable $z_{i,r}$ as

$$z_{i,r} := \frac{1}{\sqrt{m}} \cdot a_r \cdot \phi(w_r^\top x_i) \cdot \mathbf{1}_{w_r^\top x_i \leq \sqrt{2 \log(6mn/\delta)}}.$$

Then $z_{i,r}$ only depends on $a_r \in \{-1,1\}$ and $w_r \sim \mathcal{N}(0,I)$. Notice that $\mathbb{E}_{a_r,w_r}[z_{i,r}] = 0$, and $|z_{i,r}| \leq \sqrt{2\log(6mn/\delta)}$. Moreover,

$$\begin{aligned}
\mathop{\mathbb{E}}_{a_r,w_r}[z_{i,r}^2] &= \mathop{\mathbb{E}}_{a_r,w_r}\left[\frac{1}{m}a_r^2\phi^2(w_r^\top x_i)\mathbf{1}^2_{w_r^\top x_i \leq \sqrt{2\log(6mn/\delta)}}\right] \\
&= \frac{1}{m}\mathop{\mathbb{E}}_{a_r}[a_r^2]\cdot \mathop{\mathbb{E}}_{w_r}\left[\phi^2(w_r^\top x_i)\mathbf{1}^2_{w_r^\top x_i \leq \sqrt{2\log(6mn/\delta)}}\right] \\
&\leq \frac{1}{m}\cdot 1 \cdot \mathop{\mathbb{E}}_{w_r}[(w_r^\top x_i)^2] \\
&= \frac{1}{m},
\end{aligned}$$

where the second step uses independence between a_r and w_r, the third step uses $a_r \in \{-1,1\}$ and $\phi(t) = \max\{t,0\}$, and the last step follows from $w_r^\top x_i \sim \mathcal{N}(0,1)$.

Now we are ready to apply Bernstein inequality (Lemma A.1) to get for all $t > 0$,

$$\Pr\left[\sum_{r=1}^m z_{i,r} > t\right] \leq \exp\left(-\frac{t^2/2}{m\cdot\frac{1}{m} + \sqrt{2\log(6mn/\delta)}\cdot t/3}\right).$$

Setting $t = \sqrt{2\log(6mn/\delta)}\cdot\log(8n/\delta)$, we have with probability at least $1-\frac{\delta}{8n}$,

$$\sum_{r=1}^m z_{i,r} \leq \sqrt{2\log(6mn/\delta)}\cdot\log(8n/\delta).$$

Notice that we can also apply Bernstein inequality (Lemma A.1) on $-z_{i,r}$ to get

$$\Pr\left[\sum_{r=1}^m z_{i,r} < -t\right] \leq \exp\left(-\frac{t^2/2}{m\cdot\frac{1}{m} + \sqrt{2\log(6mn/\delta)}\cdot t/3}\right).$$

Let E_2 be the event that for all $i \in [n]$,

$$\left|\sum_{r=1}^m z_{i,r}\right| \leq \sqrt{2\log(2mn/\delta)}\cdot\log(8n/\delta).$$

By applying union bound on all $i \in [n]$, we have $\Pr[\neg E_2] \leq \delta/3$.

By union bound, $E_1 \cap E_2 \cap E_3$ will happen with probability at least $1-\delta$.

If both E_1 and E_2 happen, we have

$$\begin{aligned}
\|y - u(0)\|_2^2 &= \sum_{i=1}^n (y_i - f(W(0),a,x_i))^2 \\
&= \sum_{i=1}^n \Big(y_i - \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\phi(w_r^\top x_i)\Big)^2 \\
&= \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n \frac{y_i}{\sqrt{m}}\sum_{r=1}^m a_r\phi(w_r^\top x_i) + \sum_{i=1}^n \frac{1}{m}\Big(\sum_{r=1}^m a_r\phi(w_r^\top x_i)\Big)^2 \\
&= \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n y_i\sum_{r=1}^m z_{i,r} + \sum_{i=1}^n \Big(\sum_{r=1}^m z_{i,r}\Big)^2 \\
&\leq \sum_{i=1}^n y_i^2 + 2\sum_{i=1}^n |y_i|\sqrt{2\log(2mn/\delta)}\cdot\log(4n/\delta) + \sum_{i=1}^n \Big(\sqrt{2\log(2mn/\delta)}\cdot\log(4n/\delta)\Big)^2 \\
&= O(n\log(m/\delta)\log^2(n/\delta)),
\end{aligned}$$

where the first step uses E_1, the second step uses E_2, and the last step follows from $|y_i| = O(1), \forall i \in [n]$.

□

B.5. Local Steps

The following theorem is standard in neural tangent kernel theory (see e.g. (Song & Yang, 2019)).

Lemma B.9. *With probability at least $1-\delta$ over the random initialization, the following holds for all $k \in [K]$ and $c \in [N]$ and $r \in [m]$ in step t*

$$\|y_c^{(k)}(t) - y_c\|_2^2 \leq (1 - \eta_{\text{local}}\lambda/2)^k \cdot \|y_c^{(0)}(t) - y_c\|_2^2, \tag{12}$$

$$\|w_{k,c,r}(t+1) - w_{0,c,r}(t)\|_2 \leq \frac{4\sqrt{n}\|y_c^{(0)}(t) - y_c\|_2}{\sqrt{m}\lambda}, \tag{13}$$

$$\|y_c^{(k+1)}(t) - y_c^{(k)}(t)\|_2^2 \leq \eta_{\text{local}}^2 n^2 \cdot \|y_c^{(k)}(t) - y_c\|_2^2. \tag{14}$$

We then prove a Lemma that controls the updates in local steps.

Lemma B.10. *Given Eq (14) for all $k \in [K], c \in [N]$ in step t the following holds for all $k \in [K], c \in [N]$*

$$\|y_c(t) - y_c^{(k)}(t)\|_2 \leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|,$$
$$\|\Delta u_r(t)\|_2 \leq \frac{2\eta_{\text{local}} K(1 + 2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2.$$

Proof. For the first inequality, from Eq (14) we have

$$\begin{aligned}
\|y_c - y_c^{(k)}(t)\|_2 &\leq \|y_c^{(k)}(t) - y_c^{(k-1)}(t)\|_2 + \|y_c^{(k-1)}(t) - y_c\|_2 \\
&\leq (\eta_{\text{local}} n + 1)\|y_c - y_c^{(k-1)}(t)\|_2 \\
&\leq (\eta_{\text{local}} n + 1)^k \|y_c - y_c(t)\|_2.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|y_c(t) - y_c^{(k)}(t)\|_2 &\leq \sum_{i=1}^{k} \|y_c^{(j)}(t) - y_c^{(j-1)}(t)\|_2 \\
&\leq \sum_{i=1}^{k} \eta_{\text{local}} n \|y_c - y_c^{(j-1)}(t)\|_2 \\
&\leq \sum_{i=1}^{k} \eta_{\text{local}} n (\eta_{\text{local}} n + 1)^{j-1} \|y_c - y_c(t)\|_2 \\
&\leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|_2
\end{aligned}$$

where the last step comes from the choice of η_{local}.

For the second inequality, notice that

$$\begin{aligned}
\|\Delta u_r(t)\|_2 &= \eta_{\text{local}} \Big\| \frac{a_r}{N} \sum_{c\in[N]} \sum_{k\in[K]} \frac{1}{\sqrt{m}} \sum_{j\in S_c} (y_j - y^{(k)}(t)_j) x_j \mathbf{1}_{w_{r,k,c}(t)^\top x_j \geq 0} \Big\|_2 \\
&\leq \frac{\eta_{\text{local}}}{N\sqrt{m}} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} |y_j - y^{(k)}(t)_j| \\
&\leq \frac{\eta_{\text{local}}\sqrt{n}}{N\sqrt{m}} \sum_{k\in[K]} \|y - y^{(k)}(t)\|_2
\end{aligned}$$

where the second step comes form triangle inequality and $\|x_i\|_2 = 1$ and the last step comes from Cauchy-Schwartz inequality. From the $\|y_c(t) - y_c^{(k)}(t)\|_2 \leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|_2$ we have

$$\|y - y^{(k)}(t)\|_2^2 = \sum_{c\in[N]} \|y_c - y_c^{(k)}(t)\|_2^2 = \sum_{c\in[N]} 2(\|y_c - y_c(t)\|_2^2 + \|y_c(t) - y_c^{(k)}(t)\|_2^2)$$

$$\leq \sum_{c\in[N]} 2(\|y_c - y_c(t)\|_2^2 + (2\eta_{\text{local}} nK)^2 \|y_c(t) - y_c\|_2^2)$$
$$\leq 2(1 + 2\eta_{\text{local}} nK)^2 \|y - y(t)\|_2^2. \tag{15}$$

It thus follows that

$$\|\Delta u_r(t)\|_2 \leq \frac{\eta_{\text{local}}\sqrt{n}}{N\sqrt{m}} \sum_{k\in[K]} \|y - y^{(k)}(t)\|_2$$
$$\leq \frac{2\eta_{\text{local}} K(1 + 2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2.$$

□

B.6. Technical Lemma

Lemma B.11. *For any set of weight vectors $\widetilde{w}_1, \cdots, \widetilde{w}_m \in \mathbb{R}^d$ and $\widehat{w}_1, \cdots, \widehat{w}_m \in \mathbb{R}^d$ define $H(\widetilde{w}, \widehat{w}) \in \mathbb{R}^{n\times n}$ as*

$$H(\widetilde{w}, \widehat{w})_{i,j} = \frac{1}{m} x_i^\top x_j \sum_{r=1}^m \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widehat{w}_r^\top x_j \geq 0}.$$

Let $R \in (0, 1)$ and $w_1, \cdots, w_m$ be iid generated from $\mathcal{N}(0, I)$. Then we have with probability at least $1 - n^2 \cdot \exp(-mR/10)$ the following holds

$$\|H(w, w) - H(\widetilde{w}, \widehat{w})\|_F < 2nR$$

for any $\widetilde{w}_1, \cdots, \widetilde{w}_m \in \mathbb{R}^d$ and $\widehat{w}_1, \cdots, \widehat{w}_m \in \mathbb{R}^d$ such that $\|\widehat{w}_r - w_r\|_2 \leq R$ and $\|\widetilde{w}_r - w_r\|_2 \leq R$ for any $r \in [m]$.

Proof. For each $r \in [m]$ and $i, j \in [n]$, we define

$$s_{r,i,j} := \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widehat{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0}.$$

The random variable we consider can be rewritten as follows

$$\sum_{i=1}^n \sum_{j=1}^n |H(\widetilde{w}, \widehat{w})_{i,j} - H(w, w)_{i,j}|^2$$
$$\leq \frac{1}{m^2} \sum_{i=1}^n \sum_{j=1}^n \left(\sum_{r=1}^m \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widehat{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right)^2$$
$$= \frac{1}{m^2} \sum_{i=1}^n \sum_{j=1}^n \Big(\sum_{r=1}^m s_{r,i,j} \Big)^2.$$

It thus suffices to bound $\frac{1}{m^2}(\sum_{r=1}^m s_{r,i,j})^2$.

Fix i, j and we simplify $s_{r,i,j}$ to s_r. Then $\{s_r\}_{r=1}^m$ are mutually independent random variables.

We define the event

$$A_{i,r} = \left\{ \exists u : \|u - w_r\|_2 \leq R, \mathbf{1}_{x_i^\top w_r \geq 0} \neq \mathbf{1}_{x_i^\top u \geq 0} \right\}.$$

If $\neg A_{i,r}$ and $\neg A_{j,r}$ happen, then

$$\left| \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widehat{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right| = 0.$$

If $A_{i,r}$ or $A_{j,r}$ happen, then

$$\left| \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widehat{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right| \leq 1.$$

So we have

$$\begin{aligned}\mathop{\mathbb{E}}_{w_r}[s_r] \leq \mathop{\mathbb{E}}_{w_r}\left[\mathbf{1}_{A_{i,r}\vee A_{j,r}}\right] &\leq \Pr[A_{i,r}] + \Pr[A_{j,r}] \\ &\leq \frac{4R}{\sqrt{2\pi}} \\ &\leq 2R,\end{aligned}$$

and

$$\begin{aligned}\mathop{\mathbb{E}}_{w_r}\left[\left(s_r - \mathop{\mathbb{E}}_{w_r}[s_r]\right)^2\right] &= \mathop{\mathbb{E}}_{w_r}[s_r^2] - \mathop{\mathbb{E}}_{w_r}[s_r]^2 \\ &\leq \mathop{\mathbb{E}}_{w_r}[s_r^2] \\ &\leq \mathop{\mathbb{E}}_{w_r}\left[\left(\mathbf{1}_{A_{i,r}\vee A_{j,r}}\right)^2\right] \\ &\leq \frac{4R}{\sqrt{2\pi}} \\ &\leq 2R.\end{aligned}$$

We also have $|s_r| \leq 1$. So we can apply Bernstein inequality (Lemma A.1) to get for all $t > 0$,

$$\begin{aligned}\Pr\left[\sum_{r=1}^m s_r \geq 2mR + mt\right] &\leq \Pr\left[\sum_{r=1}^m (s_r - \mathbb{E}[s_r]) \geq mt\right] \\ &\leq \exp\left(-\frac{m^2t^2/2}{2mR + mt/3}\right).\end{aligned}$$

Choosing $t = R$, we get

$$\begin{aligned}\Pr\left[\sum_{r=1}^m s_r \geq 3mR\right] &\leq \exp\left(-\frac{m^2R^2/2}{2mR + mR/3}\right) \\ &\leq \exp\left(-mR/10\right).\end{aligned}$$

It follows that

$$\Pr\left[\frac{1}{m}\sum_{r=1}^m s_r \geq 3R\right] \leq \exp(-mR/10).$$

Similarly

$$\Pr\left[\frac{1}{m}\sum_{r=1}^m s_r \leq -3R\right] \leq \exp(-mR/10).$$

Therefore we complete the proof. □

Lemma B.12. *If Eq. (6) holds for $i = 0, \cdots, k$, then we have for all $r \in [m]$*

$$\|u_r(t) - u_r(0)\|_2 \leq \frac{8\sqrt{n}\|y - y(0)\|_2}{\sqrt{m}\lambda} := D.$$

Proof. We have

$$\|u_r(t) - u_r(0)\|_2 \leq \eta_{\text{global}} \sum_{\tau=0}^t \|\Delta u_r(\tau)\|_2$$

$$\leq \eta_{\text{global}} \sum_{\tau=0}^{t} \frac{2\eta_{\text{local}} K(1+2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(\tau)\|_2$$
$$\leq \eta_{\text{global}} \frac{2\eta_{\text{local}} K(1+2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \sum_{\tau=0}^{t} (1 - \frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{2N})^{\tau} \|y - y(0)\|_2$$
$$\leq \frac{8\sqrt{n}\|y - y(0)\|_2}{\sqrt{m}\lambda}.$$

where the second step comes from Lemma B.10 and the last step comes from the choice of η_{local}. □

C. Generalization

In this section, we generalize our initialization scheme to each $w_r(0) \sim \mathcal{N}(0, \sigma^2 I)$. Notice that this just introduces an extra σ^{-2} term to every occurrence of m. In addition, we use $U(t) = [u_1(t), \cdots, u_m(t)]^\top \in \mathbb{R}^{d\times m}$ to denote parameters in a matrix form. For convenience, we first list several definitions and results which will be used in the proof our generalization theorem. Our setting mainly follows (Arora et al., 2019a). This section is organized as follows:

- In Appendix C.1, we introduce several definitions.
- In Appendix C.2, we list some tools from previous work.
- In Appendix C.3, we upper bound the movement of weights which corresponds to the complexity of our model.
- In Appendix C.4, we present some technical claims used in the proof.
- In Appendix C.5, we show the generalization result of FL-NTK.

C.1. Definitions

Definition C.1 (Non-degenerate Data Distribution, Definition 5.1 in (Arora et al., 2019a))**.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$ is* (λ, δ, n)-non-degenerate*, if with probability at least $1-\delta$, for n iid samples $\{(x_i, y_i)\}_{i=1}^n$ chosen from $\mathcal{D}$, $\lambda_{\min}(H^\infty) \geq \lambda > 0$.*

Definition C.2 (Loss Functions)**.** *Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the loss function. For function $f : \mathbb{R}^d \to \mathbb{R}$, for distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$, the* population loss *is defined as*

$$L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(f(x), y)].$$

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be n samples. The empirical loss over S is defined as

$$L_S(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

Definition C.3 (Rademacher Complexity)**.** *Let $\mathcal{F}$ be a class of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$. Given n samples $S = \{x_1, \cdots, x_n\}$ where $x_i \in \mathbb{R}^d$ for $i \in [n]$, the* empirical Rademacher complexity *of $\mathcal{F}$ is defined as*

$$\mathcal{R}_S(\mathcal{F}) := \frac{1}{n} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i) \right].$$

where $\epsilon \in \mathbb{R}^d$ and each entry of ϵ are drawn from independently uniform at random from $\{\pm 1\}$.

C.2. Tools from Previous Work

Theorem C.4 (Theorem B.1 in (Arora et al., 2019a))**.** *Suppose the loss function $\ell(\cdot, \cdot)$ is bounded in $[0, c]$ for some $c > 0$ and is ρ-Lipschitz in its first argument. Then with probability at least $1-\delta$ over samples S of size n,*

$$\sup_{f\in\mathcal{F}} \{L_{\mathcal{D}}(f) - L_S(f)\} \leq 2\rho\mathcal{R}_S(\mathcal{F}) + 3c\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Lemma C.5 (Lemma 5.4 in (Arora et al., 2019a)). *Given $R > 0$, with probability at least $1-\delta$ over the random initialization on $U(0) \in \mathbb{R}^{m\times d}$ and $a \in \mathbb{R}^m$, for all $B > 0$, the function class*

$$\mathcal{F}_{R,B}^{U(0),a} = \{f(U,\cdot,a) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\phi(u_r^\top x) : \|u_r - u_r(0)\|_2 \le R, \forall r \in [m]; \|U - U(0)\|_F \le B\}$$

has bounded empirical Rademacher complexity

$$\mathcal{R}_S(\mathcal{F}_{R,B}^{U(0),a}) \le \frac{B}{\sqrt{2n}}\left(1 + \left(\frac{2\log(2/\delta)}{m}\right)^{1/4}\right) + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{2\log(2/\delta)}.$$

Lemma C.6 (Lemma C.3 in (Arora et al., 2019a)). *With probability at least $1-\delta$ we have*

$$\|H(0) - H^\infty\|_F \le O(n\sqrt{\log(n/\delta)/\sqrt{m}}).$$

C.3. Complexity Bound

To simplify the proof in the following sections, we define $\rho := \eta_{\text{local}}\eta_{\text{global}}K/N$.

Now we prove a key technical lemma which will be used to prove the main result.

Lemma C.7. *Let $\lambda = \lambda_{\min}(H^\infty) > 0$. Fix $\sigma > 0$, let $m = \Omega(\lambda^{-4}\sigma^{-2}n^4\log(n/\delta))$, we iid initialize $w_r \sim \mathcal{N}(0,\sigma^2 I)$, a_r sampled from $\{-1,+1\}$ uniformly at random for $r \in [m]$ and set $\eta_{\text{local}} = O(\frac{\lambda}{n^2K\kappa}), \eta_{\text{global}} = O(1)$. For weights $w_1,\cdots,w_m \in \mathbb{R}^d$, let $\text{vec}(W) = [w_1^\top\ w_2^\top\ \cdots w_m^\top]^\top \in \mathbb{R}^{md}$ be the concatenation of $w_1,\cdots,w_m$. Then with probability at least $1-6\delta$ over the random initialization, we have for all $t \ge 0$,*

- $\|u_r(t) - u_r(0)\|_2 \le \frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$,
- $\|U(t) - U(0)\|_F \le (y^\top(H^\infty)^{-1}y)^{1/2} + O\left((\frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\sigma^{1/2}m^{1/4}})\cdot\text{poly}(\log(m/\delta))\right)$.

Proof. Similarly to Appendix B, $\|u_r(t) - u_r(0)\|_2 \le \frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$ and $\|w_{k,c,r}(t) - u(0)\|_2 \le \frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$. For integer $k \ge 0$, define $J(k,t) \in \mathbb{R}^{md\times n}$ as the matrix

$$J(k,t) = \frac{1}{\sqrt{m}}\begin{pmatrix} a_1x_1\mathbf{1}_{w_{k,c_1,1}(t)^\top x_1\ge 0} & \cdots & a_1x_n\mathbf{1}_{w_{k,c_n,1}(t)^\top x_n\ge 0} \\ \vdots & \ddots & \vdots \\ a_mx_1\mathbf{1}_{w_{k,c_1,m}(t)^\top x_1\ge 0} & \cdots & a_mx_n\mathbf{1}_{w_{k,c_n,m}(t)^\top x_n\ge 0}\end{pmatrix}$$

where $c_i \in [N]$ denotes the unique client such that $i \in c_i$. We claim that

$$\|J(k,t) - J(0,0)\|_F \le O\Big(n\cdot\big(\delta + \frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}}\big)\Big)^{1/2}.$$

In fact, we can calculate $\|J(k,t) - J(0,0)\|_F^2$ in the following

$$\begin{aligned}\|J(k,t) - J(0,0)\|_F^2 &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\left(\|x_i\|_2\cdot a_i(\mathbf{1}_{w_{k,c,r}(t)^\top x_i\ge 0} - \mathbf{1}_{u_r(0)^\top x_i\ge 0})\right)^2 \\ &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\left(\mathbf{1}_{w_{k,c,r}(t)^\top x_i\ge 0} - \mathbf{1}_{u_r(0)^\top x_i\ge 0}\right)^2 \\ &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\mathbf{1}_{\mathbf{1}_{w_{k,c,r}(t)^\top x_i\ge 0}\ne\mathbf{1}_{u_r(0)^\top x_i\ge 0}}.\end{aligned}$$

Fix $c \in [N], i \in S_c$ and for $r \in [m]$ define t_r as follows

$$t_r = \mathbf{1}_{\mathbf{1}_{w_{k,c,r}(t)^\top x_i \geq 0} \neq \mathbf{1}_{u_r(0)^\top x_i \geq 0}}.$$

Consider the event

$$A_{i,r} = \{\exists w : \ \|u_r(0) - w\|_2 \leq R, \mathbf{1}_{w^\top x_i \geq 0} \neq \mathbf{1}_{u_r(0)^\top x_i \geq 0}\}$$

where $R = \frac{Cn\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\lambda\sqrt{m}}$ for sufficiently small constant $C > 0$. If $t_r = 1$ then either $A_{i,r}$ happens or $\|w_{k,c,r}(t) - u_r(0)\|_2 \leq R$, otherwise $t_r = 0$. Therefore

$$\mathbb{E}[t_r] \leq \Pr[A_{i,r}] + \Pr[\|u_r(0) - w_{k,c,r}(t)\|_2 < R] \leq R\sigma^{-1} + \delta.$$

And similarly $\mathbb{E}[(t_r - \mathbb{E}[t_r])^2] \leq \mathbb{E}[t_r^2] = R\sigma^{-1} + \delta$. Applying Bernstein inequality, we have for all $t > 0$,

$$\Pr\Big[\sum_{r=1}^m t_r \geq mR\sigma^{-1} + m\delta + mt\Big] \leq \exp\left(-\frac{m^2t^2}{2(mR\sigma^{-1} + m\delta + mt/3)}\right).$$

Choosing $t = R\sigma^{-1} + \delta$,

$$\Pr\Big[\sum_{r=1}^m t_r \geq 2m(R\sigma^{-1} + \delta)\Big] \leq \exp(-m(R\sigma^{-1} + \delta)/10).$$

By applying union bound over $i \in [n]$, we have with probability at least $1 - n\exp(-m(R\sigma^{-1} + \delta)/10)$, $\|J(k,t) - J(0,0)\|_F \leq 2n(R\sigma^{-1} + \delta)$. This is exactly what we need.

Notice that we can rewrite the update rule in federated learning as

$$\begin{aligned}\mathrm{vec}(U(t+1)) =& \ \mathrm{vec}(U(t)) - \frac{\eta_{\mathrm{global}}}{N}\sum_{k\in[K]}\eta_{\mathrm{local}}J(k,c)(y^{(k)}(t) - y) \\ =& \ \mathrm{vec}(U(t)) - \rho\frac{1}{K}\sum_{k\in[K]}J(k,c)(y^{(k)}(t) - y) \end{aligned} \tag{16}$$

where the last step follows from definition of $\rho = \eta_{\mathrm{global}}\eta_{\mathrm{local}}K/N$.

Recall from Appendix B that

$$\Delta u_r(t) = \frac{a_r}{N}\sum_{c\in[N]}\sum_{k\in[K]}\frac{\eta_{\mathrm{local}}}{\sqrt{m}}\sum_{j\in S_c}(y_j - y_c^{(k)}(t)_j)x_j\mathbf{1}_{w_{k,c,r}(t)^\top x_j \geq 0},$$

and

$$\begin{aligned}H(t,k,c)_{i,j} &= \frac{1}{m}\sum_{r=1}^m x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}, \\ H(t,k,c)^\perp_{i,j} &= \frac{1}{m}\sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}.\end{aligned}$$

We have

$$\begin{aligned}v_{1,i} =& \frac{1}{\sqrt{m}}\sum_{r\in Q_i} a_r\Big(\phi\big((u_r(t) + \eta_{\mathrm{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i)\Big) \\ =& -\frac{\eta_{\mathrm{local}}\eta_{\mathrm{global}}K}{N}\sum_{j\in S_c}(y(t)_j - y_j)H^\infty_{i,j} + (-\frac{\eta_{\mathrm{local}}\eta_{\mathrm{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y(t)_j)H^\infty_{i,j}\end{aligned}$$

$$+(-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j-y_j)(H(t,k,c)_{i,j}-H^{\infty}_{i,j})$$
$$+(-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j-y_j)H(t,k,c)^{\perp}_{i,j},$$
$$v_{2,i}=\frac{1}{\sqrt{m}}\sum_{r\in\overline{Q}_i}a_r\Big(\phi\big((u_r(t)+\eta_{\text{global}}\Delta u_r(t))^{\top}x_i\big)-\phi(u_r(t)^{\top}x_i)\Big).$$

Following the proof of Appendix B, let

$$\xi_i(t)=v_{2,i}(t)+(-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j-y(t)_j)H^{\infty}_{i,j}$$
$$+(-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j-y_j)(H(t,k,c)_{i,j}-H^{\infty}_{i,j})$$
$$+(-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j-y_j)H(t,k,c)^{\perp}_{i,j}.$$

Notice that

$$\sum_{i=1}^{n}|\overline{Q}_i|=\sum_{r=1}^{m}\sum_{i=1}^{n}\mathbf{1}_{r\in\overline{Q}_i}=\sum_{i=1}^{n}(\sum_{r=1}^{m}\mathbf{1}_{r\in\overline{Q}_i}).$$

Hence by Eq. (9), with probability at least $1-n\exp(-mR\sigma^{-1})$ we have

$$\sum_{i=1}^{n}|\overline{Q}_i|\le 4mnR\sigma^{-1}.$$

Similar to Appendix B, by the choice of $R=\frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$ and

$$\|y-y(0)\|_2=O\Big(\sqrt{n\log(m/\delta)\log^2(n/\delta)}\Big),$$

we can bound $\xi(t)=[\xi_1(t),\cdots,\xi_n]^{\top}\in\mathbb{R}^n$ as

$$\|\xi(t)\|_2\le O\Big(\frac{\eta_{\text{global}}\eta_{\text{local}}n^{5/2}K\kappa\sqrt{\log(m/\delta)\log^2(n/\delta)}}{N\sigma\lambda\sqrt{m}}\|y-y(t)\|_2\Big)$$
$$=O\Big(\frac{\rho n^{5/2}\kappa\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}}\|y-y(t)\|_2\Big)\tag{17}$$

where the last step follows from definition of $\rho=\eta_{\text{global}}\eta_{\text{local}}K/N$.

Notice that with probability at least $1-\delta$, for all $i\in[n]$,

$$|y_i(0)|\le\sigma\cdot\sqrt{2\log(2mn/\delta)}\cdot\log(4n/\delta),$$

which implies

$$\|y(0)\|_2^2\le n\sigma^2\cdot 2\log(2mn/\delta)\cdot\log^2(4n/\delta).\tag{18}$$

Therefore we can explicitly write the dynamics of the global model as

$$y(t)-y=(I-\frac{\eta_{\text{global}}\eta_{\text{local}}K}{N}H^{\infty})(y(t-1)-y)+\xi(t-1)$$

$$
\begin{aligned}
&= (I - \rho H^\infty)(y(t-1) - y) + \xi(t-1) \\
&= (I - \rho H^\infty)^t (y(0) - y) + \sum_{\tau=0}^{t-1} (I - \rho H^\infty)^\tau \xi(t-1-\tau) \\
&= -(I - \rho H^\infty)^t y + e(t).
\end{aligned}
$$

where the second step follows from definition of $\rho = \eta_{\text{global}} \eta_{\text{local}} K/N$, the third step comes from recursively applying the former step.y.

By Eq (17) and Eq (18) we have

$$
\begin{aligned}
e(t) &= (I - \rho H^\infty)^t y(0) + \sum_{\tau=0}^{t-1} (I - \rho H^\infty)^\tau \xi(t-1-\tau) \\
&= O\bigg((1-\rho)^t \cdot \sqrt{n\sigma^2} \cdot \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta) + t(1-\rho)^t \cdot \frac{\rho n^3 \log(m/\delta) \log^2(n/\delta)}{\lambda \sigma \sqrt{m}} \bigg)
\end{aligned}
$$

where we used $\|y(t) - y\|_2^2 \leq (1 - \frac{\eta_{\text{global}} \eta_{\text{local}} \lambda K}{2N})^t \cdot \|y(0) - y\|_2^2$ from Theorem B.3.

By Eq (16),

$$
\begin{aligned}
\operatorname{vec}(U(T)) - \operatorname{vec}(U(0)) &= \sum_{t=0}^{T-1} (\operatorname{vec}(U(t+1)) - \operatorname{vec}(U(t))) \\
&= \sum_{t=0}^{T-1} \left(-\rho \cdot \frac{1}{K} \sum_{k \in [K]} J(k,t)(y^{(k)}(t) - y) \right) \\
&= \sum_{t=0}^{T-1} \rho \cdot J(0,0)(I - \rho H^\infty)^t y \\
&\quad + \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k \in [K]} (J(k,t) - J(0,0))(I - \rho H^\infty)^t y \\
&\quad - \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k \in [K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)) \\
&= B_1 + B_2 + B_3
\end{aligned}
$$

where

$$
\begin{aligned}
B_1 &:= +\rho \cdot \sum_{t=0}^{T-1} J(0,0)(I - \rho H^\infty)^t y, \\
B_2 &:= +\rho \cdot \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k \in [K]} (J(k,t) - J(0,0))(I - \rho H^\infty)^t y, \\
B_3 &:= -\rho \cdot \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k \in [K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)).
\end{aligned}
$$

We bound these terms separately.

Putting Claim C.8, C.9 and C.10 together we have

$$
\begin{aligned}
&\|U(T) - U(0)\|_F \\
&= \|\operatorname{vec}(U(T)) - \operatorname{vec}(U(0))\|_2 \\
&= B_1 + B_2 + B_3
\end{aligned}
$$

$$\leq (y^\top (H^\infty)^{-1} y)^{1/2} + O\left((\frac{n^2\sqrt{\log(n/\delta)}}{\lambda^2\sqrt{m}})^{1/2} + (\frac{n^{3/2}}{m^{1/4}\sigma^{1/2}\lambda^{3/2}} + \frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3\sigma\sqrt{m}}) \cdot \mathrm{poly}(\log(m/\delta)) \right)$$

$$\leq (y^\top (H^\infty)^{-1} y)^{1/2} + O\left(\frac{n\sigma}{\lambda} \cdot \mathrm{poly}(\log(m/\delta)) + \frac{n^{7/2}}{\sigma^{1/2} m^{1/4}} \cdot \mathrm{poly}(\log(m/\delta)) \right)$$

which completes the proof of Lemma C.7. □

C.4. Technical Claims

Claim C.8 (Bounding B_1)**.** *With probability at least $1-\delta$ over the random initialization, we have*

$$\|B_1\|_2^2 \leq y^\top D^\top H^\infty D y + O(\frac{n^2\sqrt{\log(n/\delta)}}{\lambda^2\sqrt{m}})$$

where $D = \sum_{t=0}^{T-1} \rho(I - \rho H^\infty)^t \in \mathbb{R}^{n\times n}$.

Proof. Recall $D = \sum_{t=0}^{T-1} \rho(I - \rho H^\infty)^t \in \mathbb{R}^{n\times n}$, then we have

$$\begin{aligned}
\|B_1\|_2^2 &= \Big\| \sum_{t=0}^{T-1} \rho \cdot J(0,0) \cdot (I - \rho \cdot H^\infty)^t y \Big\|_2^2 \\
&= y^\top D^\top J(0,0)^\top J(0,0) D y \\
&= y^\top D^\top H^\infty D y + y^\top D^\top (H(0) - H^\infty) D y \\
&\leq y^\top D^\top H^\infty D y + \|H(0) - H^\infty\|_F \cdot \|D\|_2^2 \|y\|^2 \\
&\leq y^\top D^\top H^\infty D y + O(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}) \left(\sum_{t=0}^{T-1} \rho(1-\rho\lambda)^t \right)^2 n \\
&\leq y^\top D^\top H^\infty D y + O(\frac{n^2\sqrt{\log(n/\delta)}}{\lambda^2\sqrt{m}})
\end{aligned}$$

where the penultimate step comes from Lemma C.6 and $y_i = O(1)$. □

Claim C.9 (Bounding B_2)**.** *With probability at least $1-\delta$ over the random initialization, we have*

$$\|B_2\|_2 \leq \frac{n^{3/2}\,\mathrm{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{3/2}}.$$

Proof. For B_2, we have

$$\begin{aligned}
\|B_2\|_2 &= \Big\| \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} (J(k,t) - J(0,0))(I - \rho H^\infty)^t y \Big\|_2 \\
&\leq \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} \|J(k,t) - J(0,0)\|_F \cdot \|I - \rho H^\infty\|_2^t \cdot \|y\|_2 \\
&\leq O\left(\frac{n\,\mathrm{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{1/2}} \cdot \rho \cdot \sum_{k=0}^{K-1} (1-\rho\lambda)^k \cdot \sqrt{n} \right) \\
&= O\left(\frac{n^{3/2}\,\mathrm{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{3/2}} \right).
\end{aligned}$$

where in the third step we use

$$\|J(k,t) - J(0,0)\|_F \leq O\Big(n \cdot \big(\delta + \frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}} \big) \Big)^{1/2}$$

and without loss of generality, we can set δ sufficiently small. □

Claim C.10 (Bounding B_3)**.** *With probability at least $1-\delta$ over the random initialization, we have*

$$\|B_3\|_2 \leq (\frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3\sigma\sqrt{m}}) \cdot \mathrm{poly}(\log(m/\delta)).$$

Proof. Notice that for $k, t \geq 0$, $\|J(k,t)\|_F^2 \leq \frac{mn}{m} = n$. By Eq (17) and Eq (18) we have

$$\begin{aligned}
\|B_3\|_2 &= \left\| -\sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)) \right\|_2 \\
&\leq \rho \frac{1}{K} \cdot \sqrt{n} \cdot \sum_{t=0}^{T-1} O\bigg((1-\rho)^t \cdot \sqrt{n\sigma^2} \cdot \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta) \\
&\quad + t(1-\rho)^t \cdot \frac{\rho n^3 \log(m/\delta)\log^2(n/\delta)}{\lambda\sigma\sqrt{m}} \bigg) \\
&\leq (\frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3\sigma\sqrt{m}}) \cdot \mathrm{poly}(\log(m/\delta)),
\end{aligned}$$

here in the first step $-\sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} J(k,t)e(k)$ is the dominant term.

□

C.5. Main Results

Now we can present our main result in this section.

Theorem C.11. *Fix failure probability $\delta \in (0,1)$. Set $\sigma = O(\lambda\,\mathrm{poly}(\log n, \log(1/\delta))/n)$, $m = \Omega\left(\sigma^{-2}(n^{14}\,\mathrm{poly}(\log m, \log(1/\delta), \lambda^{-1}))\right)$, let the two layer neural network be initialized with w_r i.i.d sampled from $\mathcal{N}(0, \sigma^2 I)$ and a_r sampled from $\{-1,+1\}$ uniformly at random for $r \in [m]$. Suppose the training data $S = \{(x_i, y_i)\}_{i=1}^n$ are i.i.d samples from a $(\lambda, \delta/3, n)$-non-degenerate distribution $\mathcal{D}$. Let $\rho = \eta_{\mathrm{local}}\eta_{\mathrm{global}}K/N$ and train the two layer neural network $f(U(t), \cdot, a)$ by federated learning for*

$$T \geq \Omega\left(\rho^{-1}\lambda^{-1}\,\mathrm{poly}(\log(n/\delta))\right)$$

iterations. Consider loss function $\ell : \mathbb{R}\times\mathbb{R} \to [0,1]$ that is 1-Lipschitz in its first argument. Then with probability at least $1-\delta$ over the random initialization on $U(0) \in \mathbb{R}^{d\times m}$ and $a \in \mathbb{R}^m$ and the training samples,the population loss $L_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(U(T), x, a), y)]$ is upper bounded by

$$L_{\mathcal{D}}(f) \leq \sqrt{2y^\top (H^\infty)^{-1} y/n} + O(\sqrt{\log(n/(\lambda\delta))/(2n)}).$$

Proof. We will define a sequence of failing events and bound these failure probability individually, then we can apply the union bound to obtain the desired result.

Let E_1 be the event that $\lambda_{\min}(H^\infty) < \lambda$. Because $\mathcal{D}$ is $(\lambda, \delta/3, n)$-non-degenerate, $\Pr[E_1] \leq \epsilon/3$. In the remaining of the proof we assume E_1 does not happen.

Let E_2 be the event that $L_S(f(U(T), \cdot, a)) = \frac{1}{n}\sum_{i=1}^n \ell(f(U(T), x_i, a), y_i) > \frac{1}{\sqrt{n}}$. By Theorem B.3 with scaling δ properly, with probability $1-\delta/9$ we have $L_S(f(U(T), \cdot, a)) \leq \frac{1}{\sqrt{n}}$. So we have $\Pr[E_2] \leq \delta/9$.

Set $R, B > 0$ as

$$\begin{aligned}
R &= O(\frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\lambda\sqrt{m}}), \\
B &= (y^\top (H^\infty)^{-1} y)^{1/2} + O\left(\frac{n\sigma}{\lambda} \cdot \mathrm{poly}(\log(m/\delta)) + \frac{n^{7/2}}{\sigma^{1/2}m^{1/4}} \cdot \mathrm{poly}(\log(m/\delta))\right).
\end{aligned}$$

Notice that $\|y\|_2 = O(\sqrt{n})$ and $\|(H^\infty)^{-1}\|_2 = 1/\lambda$. By our setting of $\sigma = O(\frac{\lambda \operatorname{poly}(\log n, \log(1/\delta))}{n})$ and $m\sigma^2 \geq n^{14} > n^{12}$, $B = O(\sqrt{n/\lambda})$. Let E_3 be the event that there exists $r \in [m]$ so that $\|u_r - u_r(0)\|_2 > R$, or $\|U - U(0)\|_F > B$. By Lemma C.7, $\Pr[E_3] \leq \delta/9$.

For $i = 1, 2, \cdots$,, let $B_i = i$. Let E_4 be the event that there exists $i > 0$ so that

$$\mathcal{R}_S(\mathcal{F}_{R,B_i}^{U(0),a}) > \frac{B_i}{\sqrt{2n}}\left(1 + (\frac{2\log(18/\delta)}{m})^{1/4}\right) + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{2\log(18/\delta)}.$$

By Lemma C.5, $\Pr[E_4] \leq 1 - \delta/9$.

Assume neither of E_3, E_4 happens. Let i^* be the smallest integer so that $B_{i^*} = i^* \geq B$, then we have $B_{i^*} \leq B + 1$ and $i^* = O(\sqrt{n/\lambda})$. Since E_3 does not happen, we have $f(U(T), \cdot, a) \in \mathcal{F}_{R,B_{i^*}}^{U(0),a}$. Moreover,

$$\begin{aligned}
\mathcal{R}_S(\mathcal{F}_{R,B_{i^*}}^{U(0),a}) &\leq \frac{B+1}{\sqrt{2n}}\left(1 + (\frac{2\log(18/\delta)}{m})^{1/4}\right) + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{\log(18/\delta)} \\
&= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{1}{\sqrt{n}} + O(\frac{\sqrt{n}\sigma \cdot \operatorname{poly}(\log(m/\delta))}{\lambda}) \\
&+ \frac{n^3 \operatorname{poly}(\log m, \log(1/\delta), \lambda^{-1})}{m^{1/4}\sigma^{1/2}} + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{\log(18/\delta)} \\
&= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{1}{\sqrt{n}} + O(\frac{\sqrt{n}\sigma \cdot \operatorname{poly}(\log(m/\delta))}{\lambda}) + \frac{n^3 \operatorname{poly}(\log m, \log(1/\delta), \lambda^{-1})}{m^{1/4}\sigma^{1/2}} \\
&= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{2}{\sqrt{n}}
\end{aligned}$$

where the first step follows from E_4 does not happen and the choice of B, the second step follows from the choice of R, and the last step follows from the choice of m and σ.

Finally, let E_5 be the event so that there exists $i \in \{1, 2, \cdots, O(\sqrt{n/\lambda})\}$ so that

$$\sup_{f \in \mathcal{F}_{R,B_i}^{U(0),a}} \{L_{\mathcal{D}}(f) - L_S(f)\} > 2\mathcal{R}_S(\mathcal{F}_{R,B_i}^{U(0),a}) + \Omega\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right).$$

By Theorem C.4 and applying union bound on i, we have $\Pr[E_5] \leq \delta/3$.

In the case that all of the bad events E_1, E_2, E_3, E_4, E_5 do not happen,

$$\begin{aligned}
L_{\mathcal{D}}(f(U(T), \cdot, a)) &\leq L_S(f(U(T), \cdot, a)) + 2\mathcal{R}_S(\mathcal{F}_{R,B_{i^*}}^{U(0),a}) + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right) \\
&\leq \sqrt{\frac{2y^\top (H^\infty)^{-1} y}{n}} + \frac{5}{\sqrt{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right) \\
&= \sqrt{\frac{2y^\top (H^\infty)^{-1} y}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right).
\end{aligned}$$

which is exactly what we need. □