1 Convergence Analysis

1.1 Basic Setup

Theoretical guarantees for the convergence of optimization towards the satisfaction of correctness properties are desired but very difficult to derive. In this section, we take a first step by proving the convergence theorem for a basic but non-trivial case over the interval domain.

Consider a 2-layer neural network N with input dimension d, output dimension 1, and m neurons in the hidden layer. The weight matrix of the hidden layer is W, where $\boldsymbol{w_r}$ is the weight vector connecting hidden neuron r and the input layer. The weight vector of the output layer is denoted by $\boldsymbol{a}$. The ReLU function is used as the activation function, defined as

$$\sigma(x) = \max(x, 0) = x\mathbb{1}_{x \geq 0},$$

where $\mathbb{1}_A$ is the event indicator function

$$\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

For each input variable $\boldsymbol{x}$, we assume $x_j, j \in [d]$ are bounded by intervals $[xl_j, xu_j]$. For a safety property ϕ whose input predicate $\phi_{in} = [xl^{safe}, xu^{safe}]$, $xl_j \leq xl_j^{safe} < xu_j^{safe} \leq xu_j, j \in [d]$ must hold. We normalize each interval by

$$xl_j = \frac{1}{\sqrt{d}} \frac{xl_j^{safe} - (xu_j + xl_j)/2}{(xu_j - xl_j)/2}$$

$$xu_j = \frac{1}{\sqrt{d}} \frac{xu_j^{safe} - (xu_j + xl_j)/2}{(xu_j - xl_j)/2}$$

for $j \in [d]$.

The inputs of N are vectors $\boldsymbol{x_i}^{[\boldsymbol{w_r}]}, i \in [n], r \in [m]$, where $\boldsymbol{x_i}$ is composed by elements from $\boldsymbol{xl_i}$ and $\boldsymbol{xu_i}$. The selection of elements depends on the signs of $\boldsymbol{w_r}$ and the desired safety property. For example, to find the lower bounds of $\boldsymbol{wx}$, we define for $j \in [d]$,

$$x_j^{[\boldsymbol{w}]} = \begin{cases} xl_j & \text{if } w_j \geq 0 \\ xu_j & \text{if } w_j < 0. \end{cases}$$

From the normalization of $\boldsymbol{xl}$ and $\boldsymbol{xu}$, we have $\left\| \boldsymbol{x}^{[w]} \right\|_2 \leq 1$.

Given the above definition, the output of N can be written as

$$f(W, \boldsymbol{x}, \boldsymbol{a}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\boldsymbol{w_r}^\intercal \boldsymbol{x}^{[\boldsymbol{w_r}]}).$$

We consider the general 2-norm loss function

$$L(W) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f(W, \boldsymbol{x_i}, \boldsymbol{a}))^2$$

where y_i is the label of x_i. A fixed step gradient descent algorithm is used to optimize W, which is

$$W(k+1) = W(k) - \eta \frac{\partial L(W(k))}{\partial W(k)}$$

where η is a preset step size and $k \in \mathbb{N}$ is the number of iterations. We use $u(k)$ to represent the output prediction at iteration k, i.e.,

$$u(k) = f(W(k), \boldsymbol{x}, \boldsymbol{a}).$$

Assumption 1. *We assume the network N satisfies the following properties:*

1. $\boldsymbol{w_r}(0) \sim \mathcal{N}(0, I)$, $r \in [m]$
2. $a_r = 1$, $r \in [m]$
3. $\left\| \boldsymbol{x_i}^{[\boldsymbol{w}]} \right\|_2 \leq 1$, $i \in [n]$

where $\mathcal{N}(0, I)$ represents normal distribution.

Remark 1. *Assumption 1 (2) is for clarity of the proof. We can adapt our result to $a_r \sim unif[-1, 1], r \in [m]$, as used in [?]. Assumption 1 (3) follows the normalization of $\boldsymbol{xl}$ and $\boldsymbol{xu}$.*

Definition 1.1. (Definition 1.1 in [?]) We define the following functions. Given $\boldsymbol{x}_{\boldsymbol{i}}^{[w]} \in \mathbb{R}^d$ and $\boldsymbol{w} \in \mathbb{R}^d$, we define the continuous Gram matrix $H^{cts} \in \mathbb{R}^{n\times n}$ as

$$H_{i,j}^{cts} = \mathbb{E}_{w\sim N(0,I)}\left[\boldsymbol{x}_{\boldsymbol{i}}^{[w]\top}\boldsymbol{x}_{\boldsymbol{j}}^{[w]}\mathbb{1}_{\boldsymbol{w}^\top \boldsymbol{x}_i^{[w]}\geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j^{[w]}\geq 0}\right], i,j \in [n]$$

and the discrete Gram matrix $H^{dis} \in \mathbb{R}^{n\times n}$ as

$$H_{i,j}^{dis} = \frac{1}{m}\sum_{r=1}^{m}\left[\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]\top}\boldsymbol{x}_{\boldsymbol{j}}^{[w_r]}\mathbb{1}_{\boldsymbol{w}_r^\top \boldsymbol{x}_i^{[w_r]}\geq 0,\, \boldsymbol{w}_r^\top \boldsymbol{x}_j^{[w_r]}\geq 0}\right], i,j \in [n].$$

By the definition of $\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]}$, we can write it as a function of $\boldsymbol{w_r}$

$$\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]} = \mathbb{D}_{\boldsymbol{w_r}\geq 0}\boldsymbol{xl_i} + \mathbb{D}_{\boldsymbol{w_r}<0}\boldsymbol{xu_i}$$

where $\mathbb{D}_{\boldsymbol{w_r}\geq 0} \in \mathbb{R}^{d\times d}$ is a diagonal matrix composed by $\mathbb{1}_{w_{r,j}\geq 0}, j \in [d]$. Similarly, we have

$$\begin{aligned}\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]\top}\boldsymbol{x}_{\boldsymbol{j}}^{w_r} =&(\mathbb{D}_{\boldsymbol{w_r}\geq 0}\ \boldsymbol{xl_i})^\top\ \mathbb{D}_{\boldsymbol{w_r}\geq 0}\ \boldsymbol{xl_j}\\ &+ (\mathbb{D}_{\boldsymbol{w_r}\geq 0}\ \boldsymbol{xl_i})^\top\ \mathbb{D}_{\boldsymbol{w_r}<0}\ \boldsymbol{xu_j}\\ &+ (\mathbb{D}_{\boldsymbol{w_r}<0}\ \boldsymbol{xu_i})^\top\ \mathbb{D}_{\boldsymbol{w_r}\geq 0}\ \boldsymbol{xl_j}\\ &+ (\mathbb{D}_{\boldsymbol{w_r}<0}\ \boldsymbol{xu_i})^\top\ \mathbb{D}_{\boldsymbol{w_r}<0}\ \boldsymbol{xu_j}\\ =&\boldsymbol{xl_i}^\top\ \mathbb{D}_{\boldsymbol{w_r}\geq 0}\ \boldsymbol{xl_j} + \boldsymbol{xu_i}^\top\ \mathbb{D}_{\boldsymbol{w_r}<0}\ \boldsymbol{xu_j},\end{aligned} \tag{1}$$

so H^{cts} and H^{dis} are functions of variable W.

Assumption 2. *(Assumption 1.2 in [?]). We make the following data-dependent assumption:*

$$\text{let } \lambda = \lambda_{\min}(H^{cts}), \text{ and } \lambda \in (0,1].$$

Our main result is

Theorem 1.2. *(Theorem 4.6 in [?]) Suppose network N satisfies Assumption 1 and 2. Let $m = \Omega(\lambda^{-4}n^4\log(\frac{n}{\delta}))$, and step size $\eta = O(\frac{\lambda}{n^2})$, then with probability at least $1-\delta$, we have*

$$\left\|u(k)-y\right\|_2^2 \leq (1-\frac{\eta\lambda}{2})^k\left\|u(0)-y\right\|_2^2.$$

In other words, under the over-parameterization of m, the optimization algorithm has linear convergence rate.

1.2 Detailed Proofs

The proof of Theorem 1.2 can be done by mathematical induction as used in [?]. Most proofs in our paper are similar to those in [?], except when $\boldsymbol{x}^{\boldsymbol{w}}$ is involved. Conclusions of depending theorems, lemmas, and claims are the same as [?] with our setup. But some proof details need to be modified. Here we list all necessary modifications in the proofs.

Lemma 1.3 (Lemma 4.1 in [?])**.** *We define $H^{cts}, H^{dis} \in \mathbb{R}^{n\times n}$ as follows:*

$$H_{i,j}^{cts} = \mathbb{E}_{w\sim N(0,I)}\left[\boldsymbol{x}_{\boldsymbol{i}}^{[w]\top}\boldsymbol{x}_{\boldsymbol{j}}^{[w]}\mathbb{1}_{\boldsymbol{w}^\top \boldsymbol{x}_i^{[w]}\geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j^{[w]}\geq 0}\right], i,j \in [n]$$

$$H_{i,j}^{dis} = \frac{1}{m}\sum_{r=1}^{m}\left[\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]\top}\boldsymbol{x}_{\boldsymbol{j}}^{[w_r]}\mathbb{1}_{\boldsymbol{w}_r^\top \boldsymbol{x}_i^{[w_r]}\geq 0,\, \boldsymbol{w}_r^\top \boldsymbol{x}_j^{[w_r]}\geq 0}\right], i,j \in [n].$$

Let $\lambda = \lambda_{\min}(H^{cts})$. If $m = \Omega(\lambda^{-2}n^2\log(n/\delta))$, we have

$$\left\|H^{dis}-H^{cts}\right\|_F \leq \frac{\lambda}{4}, \text{ and } \lambda_{\min}(H^{dis}) \geq \frac{3}{4}\lambda.$$

Proof. The key step in this proof is to use Hoeffding inequality to bound distance between H^{dis} and H^{cts}. To use Hoeffding inequality, we need to show that $E[H_{ij}^{dis}] = H_{ij}^{cts}$.

Let

$$z_r = \frac{1}{m}\boldsymbol{x}_{\boldsymbol{i}}^{[w_r]\top}\boldsymbol{x}_{\boldsymbol{j}}^{w_r}\mathbb{1}_{\boldsymbol{w}_r^\top \boldsymbol{x}_i^{[w_r]}\geq 0,\, \boldsymbol{w}_r^\top \boldsymbol{x}_j^{[w_r]}\geq 0}, r \in [m].$$

From Eq. (1),

$$z_r = \frac{1}{m}(\boldsymbol{xl}_i^\top \mathbb{D}_{\boldsymbol{w}_r \geq 0}\, \boldsymbol{xl}_j + \boldsymbol{xu}_i^\top \mathbb{D}_{\boldsymbol{w}_r < 0}\, \boldsymbol{xu}_j)\mathbb{1}_{\boldsymbol{w}_r^\top \boldsymbol{x}_i^{[\boldsymbol{w}_r]} \geq 0,\, \boldsymbol{w}_r^\top \boldsymbol{x}_j^{[\boldsymbol{w}_r]} \geq 0}, r \in [m].$$

Therefore, z_r is a random variable of w_r.

$$\begin{aligned}
E[H_{ij}^{dis}] &= \sum_{r=1}^{m} E_{\boldsymbol{w}_r \sim \mathcal{N}(0,I)} z_r \\
&= E_{\boldsymbol{w} \sim \mathbb{N}(0,I)} \frac{1}{m}(\boldsymbol{xl}_i^\top \mathbb{D}_{\boldsymbol{w} \geq 0}\, \boldsymbol{xl}_j + \boldsymbol{xu}_i^\top \mathbb{D}_{\boldsymbol{w} < 0}\, \boldsymbol{xu}_j)\mathbb{1}_{\boldsymbol{w}^\top \boldsymbol{x}_i^{[\boldsymbol{w}]} \geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j^{[\boldsymbol{w}]} \geq 0} \\
&= H^{cts}
\end{aligned}$$

The second equation holds because $\boldsymbol{w}_r, r \in [m]$ follows the same distribution, therefore the Expectations of z_r are the same. The subsequent the proof is the same starting from this step. □

. . .

1.3 Extension to DeepPoly

$$\begin{aligned}
f(W, x, a) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \Phi(\overline{g_i}, \underline{g_i}) \\
\underline{g_i} &= W^+ L + W^- U = w_r(\mathbb{1}_{w_r > 0} L + \mathbb{1}_{w_r < 0} U) \\
\overline{g_i} &= W^+ U + W^- L = w_r(\mathbb{1}_{w_r < 0} L + \mathbb{1}_{w_r > 0} U) \\
\underline{\Phi(\overline{g_i}, \underline{g_i})} &= \frac{\mathbb{1}_{w_r < 0} L + \mathbb{1}_{w_r > 0} U}{(\mathbb{1}_{w_r < 0} L + \mathbb{1}_{w_r > 0} U) - (\mathbb{1}_{w_r > 0} L + \mathbb{1}_{w_r < 0} U)} w_r^\top (\mathbb{1}_{w_r > 0} L + \mathbb{1}_{w_r < 0} U)
\end{aligned}$$

Denote $\mathbb{1}_{w_r < 0} = \alpha$, then $\mathbb{1}_{w_r > 0} = 1 - \alpha$.

$$\underline{\Phi(\overline{g_i}, \underline{g_i})} = w_r^\top \frac{\alpha L + (1-\alpha) U}{\left[\alpha L + (1-\alpha) U\right] - \left[(1-\alpha) L + \alpha U\right]}$$

Since $\alpha^2 - \alpha = 0$,

$$\begin{aligned}
\underline{\Phi(\overline{g_i}, \underline{g_i})} &= w_r^\top \frac{LU}{(L-U)(2\alpha - 1)} \\
&= w_r^\top \frac{LU}{(L-U)(2\mathbb{1}_{w_r < 0} - 1)}. \\
\frac{\partial f}{\partial w_r} &= \frac{1}{\sqrt{m}} a_r \frac{LU}{(L-U)(2\mathbb{1}_{w_r < 0} - 1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
H_{i,j}^{dis} &= \frac{1}{m} \sum_{r=1}^{m} \left[\frac{L_i U_i}{(L_i - U_i)(2\mathbb{1}_{w_r < 0} - 1)} \frac{L_j U_j}{(L_j - U_j)(2\mathbb{1}_{w_r < 0} - 1)} \right] \\
&= \frac{1}{m} \sum_{r=1}^{m} \left[\frac{L_i U_i L_j U_j}{(L_i - U_i)(L_j - U_j)} \right]
\end{aligned}$$

since $(2\mathbb{1}_{w_r < 0} - 1)^2 = 1$.