7. Details for the Toy Example

Suppose we train the network with n neurons for T time using gradient descent with random initialization, i.e., the network we obtain is $f_{\rho_T^n}$ using the terminology in Section 3.3. As shown by Song et al. (2018); Mei et al. (2019), $\mathcal{L}[f_{\rho_T^n}]$ is actually $\mathcal{O}(1/n+\epsilon)$ with high probability, where $\epsilon = \mathcal{L}[f_{\rho_T^\infty}]$ is the loss of the mean field limit network at training time T. Song et al. (2018) shows that $\lim_{T\to\infty}\mathcal{L}[f_{\rho_T^\infty}] = 0$ under some regularity conditions and this implies that if the training time T is sufficient, $\mathcal{L}[f_{\rho_T^\infty}]$ is generally a smaller term compared with the $\mathcal{O}(1/n)$ term.

To generate the synthesis data, we first generate a neural network $f_{\text{gen}}(\boldsymbol{x}) = \frac{1}{1000}\sum_{i=1}^N b_i \text{sigmoid}(\boldsymbol{a}_i^\top \boldsymbol{x})$, where $\boldsymbol{a}_i$ are i.i.d. sample from a 10 dimensional standard Gaussian distribution and b_i are i.i.d. sample from a uniform distribution Unif$(-5,5)$. The training data $\boldsymbol{x}$ is also generated from a 10 dimensional standard Gaussian distribution. We choose $f_{\text{gen}}(\boldsymbol{x}) = y$ as the label of data. Our training data consists of 100 data points. The network we use to fit the data is $f = \frac{1}{n}\sum_{i=1}^n b_i' \tanh(\boldsymbol{a}_i'^\top \boldsymbol{x})$. We use network with 1000 neurons for pruning and the pruned models will not be finetuned. All networks are trained for same and sufficiently large time to converge.

8. Finding Sub-Networks on CIFAR-10/100

In this subsection, we display the results of applying our proposed algorithm to various model structures on CIFAR-10 and CIFAR-100. On CIFAR-10 and CIFAR-100, we apply our algorithm to the networks already pruned by network slimming (Liu et al., 2017) provided by Liu et al. (2019b) and show that we can further compress models which have already pruned by the L_1 regularization. We apply our algorithm on the pretrained models, and finetune the model with the same experimental setting as ImageNet.

As demonstrated in Table 4, our proposed algorithm can further compress a model pruned by Liu et al. (2019b) without or only with little drop on accuracy. For example, on the pretrained VGG19 on CIFAR-10, Liu et al. (2017) can prune 30% channels and get $93.81\% \pm 0.14\%$ accuracy. Our algorithm can prune 44% channels of the original VGG19 and get $93.78\% \pm 0.16\%$ accuracy, which is almost the same as the strong baseline number reported by Liu et al. (2019b).

DataSet	Model	Method	Prune Ratio (%)	Accuracy (%)
CIFAR10	VGG19	Liu et al. (2017)	70	93.81 ± 0.14
		Ours	**56**	$\mathbf{93.78 \pm 0.16}$
	PreResNet-164	Liu et al. (2017)	60	94.90 ± 0.04
		Ours	**51**	$\mathbf{94.91 \pm 0.06}$
		Liu et al. (2017)	40	94.71 ± 0.21
		Ours	**33**	$\mathbf{94.68 \pm 0.17}$
CIFAR100	VGG19	Liu et al. (2017)	50	73.08 ± 0.22
		Ours	**44**	$\mathbf{73.05 \pm 0.19}$
	PreResNet-164	Liu et al. (2017)	60	76.68 ± 0.35
		Ours	**53**	$\mathbf{76.63 \pm 0.37}$
		Liu et al. (2017)	40	75.73 ± 0.29
		Ours	**37**	$\mathbf{75.74 \pm 0.32}$

Table 4. Accuracy on CIFAR100 and CIFAR10. "Prune ratio" stands for the total percentage of channels that are pruned in the whole network. We apply our algorithm on the models pruned by Liu et al. (2017) and find that our algorithm can further prune the models. The performance of Liu et al. (2017) is reported by Liu et al. (2019b). Our reported numbers are averaged by five runs.

9. Discussion on Assumption 2 and 5

Let $\phi_j(\boldsymbol{\theta}) = \sigma(\boldsymbol{x}^{(j)}, \boldsymbol{\theta})/\sqrt{m}$ and $\boldsymbol{\phi}(\boldsymbol{\theta}) = [\phi_1(\boldsymbol{\theta}), ..., \phi_m(\boldsymbol{\theta})]$ to be the vector of the outputs of the neuron $\sigma(\boldsymbol{x};\boldsymbol{\theta})$ scaled by $1/\sqrt{m}$, realized on a dataset $\mathcal{D}_m := \{\boldsymbol{x}^{(j)}\}_{j=1}^m$. We call $\boldsymbol{\phi}(\boldsymbol{\theta})$ the feature map of $\boldsymbol{\theta}$. Given a large network $f_{[N]}(x) = \sum_{i=1}^N \sigma(\boldsymbol{x};\boldsymbol{\theta}_i)/N$, define the marginal polytope of the feature map to be

$$\mathcal{M}_N := \text{conv}\left\{\boldsymbol{\phi}(\boldsymbol{\theta}_i) \mid i \in \{1,\dots,N\}\right\},$$

where conv denotes the convex hull. Then it is easy to see that Assumption 2 is equivalent to saying that $\mathbf{y} := [y^{(1)},\dots,y^{(m)}]/\sqrt{m}$ is in the interior of the marginal polytope $\mathcal{M}_N$, i.e., there exists $\gamma > 0$ such that $\mathcal{B}(\mathbf{y},\gamma) \subseteq \mathcal{M}_N$.

Here we denote by $\mathcal{B}(\boldsymbol{\mu}, r)$ the ball with radius r centered at $\boldsymbol{\mu}$. Similar to Assumption 2, Assumption 5 is equivalent to require that $\mathcal{B}(\mathbf{y}, \gamma^*) \subseteq \mathcal{M}$, where

$$\mathcal{M} := \operatorname{conv}\left\{\boldsymbol{\phi}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \operatorname{supp}(\rho_T^\infty)\right\}.$$

We may further relax the assumption to assuming $\mathbf{y}$ is in the relative interior (instead of interior) of $\mathcal{M}_N$ and $\mathcal{M}$. However, this requires some refined analysis and we leave this as future work.

It is worth mention that when $\mathcal{M}$ has dimension m and $f_{\rho_T^\infty}$ gives zero training loss, then assumption 5 holds. Similarly, if $\mathcal{M}_N$ has dimension m and $f_{\rho_T^N}$ gives zero training loss, then assumption 2 holds.

10. Pruning Randomly Weighted Networks

Our theoretical analysis is also applicable for pruning randomly weighted networks. Here we give the following corollary.

Corollary 4. *Under Assumption 1 and suppose that the weights $\{\boldsymbol{\theta}_i\}$ of the large neurons $f_{[N]}(\boldsymbol{x})$ are i.i.d. drawn from an absolutely continuous distribution ρ_0 with a bounded support in $\mathbb{R}^d$, without further gradient descent training. Suppose that Assumption 5 and 6 hold for ρ_0 (changing ρ_T^∞ to ρ_0). Let S_n^{Random} be the subset obtained by the proposed greedy forward selection (2) on such $f_{[N]}$ at the n-th step. For any $\delta > 0$ and $\gamma < \gamma^*/2$, when N is sufficiently large, with probability at least $1-\delta$, we have*

$$\mathcal{L}[f_{S_n^{Random}}] = \mathcal{O}\left(1/\left(\min\left(1, \gamma\right) n\right)^2\right).$$

This corollary is a special case of Theorem 3 when taking the training time to be zero ($T = 0$). And as the network is not trained, Assumption 4 are not needed for this corollary.

11. Forward Selection is Better Than Backward Elimination

A greedy backward elimination can be developed analogous to our greedy forward selection, in which we start with the full network and greedily eliminate neurons that gives the smallest increase in loss. Specifically, starting from $S_0^{\mathrm{B}} = [N]$, we sequentially delete neurons via

$$S_{n+1}^{\mathrm{B}} \leftarrow S_n^{\mathrm{B}} \setminus \{i_n\}^*, \quad \text{where } i_n^* = \operatorname*{arg\,min}_{i \in S_n^{\mathrm{B}}} \mathcal{L}[f_{S_n^{\mathrm{B}} \setminus \{i\}}], \tag{7}$$

where $\setminus$ denotes set minus. In this section, we demonstrate that the forward selection has significant advantages over backward elimination, from both theoretical and empirical perspectives.

Theoretical Comparison of Forward and Backward Methods Although greedy forward selection guarantees $\mathcal{O}(1/n)$ or $\mathcal{O}(1/n^2)$ error rate as we show in the paper, backward elimination does not enjoy similar theoretical guarantees. This is because the "effective search space" of backward elimination is more limited than that of forward selection, and gradually shrinkage over time. Specifically, at each iteration of backward elimination (7), the best neuron is chosen among S_n^{B}, which shrinks as more neurons are pruned. In contrast, the new neurons in greedy selection (2) are always selected from the full set $[N]$, which permits each neuron to be selected at every iteration, for multiple times. We now elaborate the theoretical advantages of forward selection vs. backward elimination from 1) the best achievable loss by both methods and 2) the decrease of loss across iterations.

• *On the lower bound.* In greedy forward selection, one neuron can be selected for multiple times at different iterations, while in backward elimination one neuron can only be deleted once. As a result, the best possible loss achievable by backward elimination is worse than that of greedy elimination. Specifically, because backward elimination yields a subnetwork in which each neuron appears at most once. We have an immediate lower bound of

$$\mathcal{L}[S_n^{\mathrm{B}}] \geq \mathcal{L}_N^{\mathrm{B}*}, \quad \forall n \in [N],$$

where

$$\mathcal{L}_N^{\mathrm{B}*} = \min_{\boldsymbol{\alpha}} \left\{ \mathcal{L}[f_{\boldsymbol{\alpha}}] : \ \alpha_i = \bar{\alpha}_i / \sum_{i=1}^N \bar{\alpha}_i, \quad \bar{\alpha}_i \in \{0, 1\} \right\}.$$

In comparison, for S_n^* from forward selection (2), we have from Theorem 1 that

$$\mathcal{L}[S_n^*] = \mathcal{O}(1/n) + \mathcal{L}_N^*,$$

where $\mathcal{L}_N^*$ equals (from Eq 3)

$$\mathcal{L}_N^* = \min_{\boldsymbol{\alpha}} \left\{ \mathcal{L}[f_{\boldsymbol{\alpha}}] : \ \ a_i \geq 0, \ \ \sum_{i=1}^N \alpha_i = 1 \right\}.$$

This yields a simple comparison result of

$$\mathcal{L}[S_n^{\mathrm{B}}] \geq \mathcal{L}[S_n^*] + (\mathcal{L}_N^{\mathrm{B}*} - \mathcal{L}_N^*) + \mathcal{O}(1/n).$$

Obviously, we have $\mathcal{L}_N^{\mathrm{B}*} \geq \mathcal{L}_N^*$ because $\mathcal{L}_N^*$ optimizes on a much larger set of $\boldsymbol{\alpha}$, indicating that backward elimination is inferior to forward selection. In fact, because $\mathcal{L}_N^{\mathrm{B}*}$ is most likely to be strictly larger than $\mathcal{L}_N^*$ in practice, we can conclude that $\mathcal{L}[S_n^{\mathrm{B}}] = \Omega(1) + \mathcal{L}_N^*$ where Ω is the Big Omega notation. This shows that it is impossible to prove bounds similar to $\mathcal{L}[S_n^*] = \mathcal{O}(1/n) + \mathcal{L}_N^*$ in Theorem 1 for backward elimination.

• *On the loss descend.* The key ingredient for proving the $\mathcal{O}(n^{-1})$ convergence of greedy forward selection is a recursive inequality that bounds $\mathcal{L}[f_{S_n}]$ at iteration n using $\mathcal{L}[f_{S_{n-1}}]$ from the previous iteration $n-1$. Specifically, we have

$$\mathcal{L}[f_{S_n}] \leq \mathcal{L}_N^* \ + \ \frac{\mathcal{L}_N^* - \mathcal{L}[f_{S_{n-1}}]}{n} \ + \ \frac{C}{n^2}, \tag{8}$$

where $C = \max_{\boldsymbol{u},\boldsymbol{v}} \left\{ \|\boldsymbol{u} - \boldsymbol{v}\|^2 : \boldsymbol{u}, \boldsymbol{v} \in \mathcal{M}_N \right\}$; see Appendix 12.1 for details. And inequality (8) directly implies that

$$\mathcal{L}[f_{S_n}] \leq \mathcal{L}_N^* \ + \ \frac{\mathcal{L}[f_{S_0}] - \mathcal{L}_N^*}{n}, \quad \forall n \in [N].$$

An importance reason for this inequality to hold is that the best neuron to add is selected from the whole set $[N]$ at each iteration. However, similar result does not hold for backward elimination, because the neuron to eliminate is selected from S_n^{B}, whose size shrinks when n grows. In fact, for backward elimination, we guarantee to find counter examples that violate a counterpart of (8), as shown in the following result, and thus fail to give the $\mathcal{O}(n^{-1})$ convergence rate.

Theorem 5. *For the S_n^B constructed by backward elimination in (7). There exists a full network $f_{[N]}(\boldsymbol{x}) = \sum_{i=1}^N \sigma(\boldsymbol{x};\ \boldsymbol{\theta}_i)/N$ and a dataset $\mathcal{D}_m = (\boldsymbol{x}^{(i)}, y^{(i)})_{i=1}^m$ that satisfies Assumption 1, 2, such that $\mathcal{L}_N^{B*} > 0$ and $\exists n \in [N]$*

$$\mathcal{L}[f_{S_{N-n}^B}] > \mathcal{L}_N^{B*} \ + \ \frac{\mathcal{L}[f_{S_N^B}] - \mathcal{L}_N^{B*}}{n},$$

In comparison, the S_n from greedy forward selection satisfies

$$\mathcal{L}[f_{S_n}] \leq \mathcal{L}_N^* \ + \ \frac{\mathcal{L}[f_{S_0}] - \mathcal{L}_N^*}{n}, \quad \forall n \in [N]. \tag{9}$$

In fact, on the same instance, we have $\mathcal{L}_N^* = 0$, and the faster rate $\mathcal{L}[f_{S_n}] \leq \mathcal{L}_N^* = \mathcal{O}(n^{-2})$ also holds for greedy forward selection.

Proof. Suppose the data set contains 2 data points and we represent the neurons as the feature map as in section 9. Suppose that $N = 43$, $\boldsymbol{\phi}(\boldsymbol{\theta}_1) = [0, 1.5]$, $\boldsymbol{\phi}(\boldsymbol{\theta}_2) = [0, 0]$, $\boldsymbol{\phi}(\boldsymbol{\theta}_3) = [-0.5, 1]$, $\boldsymbol{\phi}(\boldsymbol{\theta}_4) = [2, 1]$ and $\boldsymbol{\phi}(\boldsymbol{\theta}_i) = [(-1.001)^{i-3} + 2, 1]$, $i \in \{5, 6, ..., 43\}$ and the target $\mathbf{y} = [0, 1]$ (it is easy to construct the actual weights of neurons and data points such that the above feature maps hold). Deploying greedy backward elimination on this case gives that

$$\mathcal{L}[f_{S_{N-n}^{\mathrm{B}}}] > \frac{\mathcal{L}[f_{S_N^{\mathrm{B}}}] - \mathcal{L}_N^{\mathrm{B}*}}{n} + \mathcal{L}_N^{\mathrm{B}*},$$

for $n \in [38]$, where $\mathcal{L}_N^{\mathrm{B}*} = \min_{n \in [N]} \mathcal{L}_{N,n}^{\mathrm{B}*} > 0.03$. In comparison, for greedy forward selection, (9) holds from the proof of Theorem 1. In addition, on the same instance, we can verify that $\mathcal{L}_N^* = 0$, and the faster $\mathcal{O}(n^{-2})$ convergence rate also holds for greedy forward selection. In deed, the greedy forward selection is able to achieve 0 loss using two neurons (by selecting $\boldsymbol{\phi}(\boldsymbol{\theta}_3)$ for four times and $\boldsymbol{\phi}(\boldsymbol{\theta}_4)$ once). □

Model	Method	Top1 Acc	FLOPs
ResNet34	Backward	73.1	2.81G
	Forward	73.5	2.64G
	Backward	72.4	2.22G
	Forward	72.9	2.07G
MobileNetV2	Backward	71.4	257M
	Forward	71.9	258M
	Backward	70.8	215M
	Forward	71.2	201M

Table 5. Comparing greedy forward selection and backward elimination on Imagenet.

Empirical Comparison of Forward and Backward Methods We compare forward selection and backward elimination to prune Resnet34 and MobilenetV2 on Imagenet. As shown in Table 5, forward selection tends to achieve better top-1 accuracy in all the cases, which is consistent with the theoretical analysis above. The experimental settings of the greedy backward elimination is the same as that of the greedy forward selection.

12. Proofs

Our proofs use the definition of the convex hulls defined in Section 9 of Appendix.

12.1. Proof of Proposition 1

The proof of Proposition 1 follows the standard argument of proving the convergence rate of Frank-Wolfe algorithm with some additional arguments. Our algorithm is not a Frank-Wolfe algorithm, but as illustrated in the subsequent proof, we can essentially use the Frank-Wolfe updates to control the error of our algorithm.

Define $\ell(\boldsymbol{u}) = \|\boldsymbol{u} - \mathbf{y}\|^2$, then the subnetwork selection problem can be viewed as solving

$$\min_{\boldsymbol{u}\in\mathcal{M}_N} \ell(\boldsymbol{u}),$$

with $\mathcal{L}_N^* = \min_{\boldsymbol{u}\in\mathcal{M}_N} \ell(\boldsymbol{u})$. And our algorithm can be viewed as starting from $\boldsymbol{u}^0 = 0$ and iteratively updating $\boldsymbol{u}$ by

$$\boldsymbol{u}^k = (1-\xi_k)\boldsymbol{u}^{k-1} + \xi_k \boldsymbol{q}^k, \qquad \boldsymbol{q}^k = \operatorname*{arg\,min}_{\boldsymbol{q}\in \mathrm{Vert}(\mathcal{M}_N)} \ell\left((1-\xi_k)\boldsymbol{u}^{k-1} + \xi_k \boldsymbol{q}\right), \tag{10}$$

where $\mathrm{Vert}(\mathcal{M}_N) := \{\boldsymbol{\phi}(\boldsymbol{\theta}_1), ..., \boldsymbol{\phi}(\boldsymbol{\theta}_N)\}$ denotes the vertices of $\mathcal{M}_N$, and we shall take $\xi_k = 1/k$. We aim to prove that $\ell(\boldsymbol{u}^k) = O(1/k) + \mathcal{L}_N^*$. Our proof can be easily extended to general convex functions $\ell(\cdot)$ and different ξ_k schemes.

By the convexity and the quadratic form of $\ell(\cdot)$, for any $\boldsymbol{s}$, we have

$$\ell(\boldsymbol{s}) \geq \ell(\boldsymbol{u}^{k-1}) + \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s} - \boldsymbol{u}^{k-1}) \tag{11}$$

$$\ell(\boldsymbol{s}) \leq \ell(\boldsymbol{u}^{k-1}) + \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s} - \boldsymbol{u}^{k-1}) + \left\|\boldsymbol{s} - \boldsymbol{u}^{k-1}\right\|^2. \tag{12}$$

Minimizing $\boldsymbol{s}$ in $\mathcal{M}_N$ on both sides of (11), we have

$$\begin{aligned}\mathcal{L}_N^* = \min_{\boldsymbol{s}\in\mathcal{M}_N} \ell(\boldsymbol{s}) &\geq \min_{\boldsymbol{s}\in\mathcal{M}_N} \left\{\ell(\boldsymbol{u}^{k-1}) + \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s} - \boldsymbol{u}^{k-1})\right\} \\ &= \ell(\boldsymbol{u}^{k-1}) + \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s}^k - \boldsymbol{u}^{k-1}).\end{aligned} \tag{13}$$

Here we define

$$\begin{aligned}\boldsymbol{s}^k &= \operatorname*{arg\,min}_{\boldsymbol{s}\in\mathcal{M}_N} \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s} - \boldsymbol{u}^{k-1}) \\ &= \operatorname*{arg\,min}_{\boldsymbol{s}\in\mathrm{Vert}(\mathcal{M}_N)} \nabla\ell(\boldsymbol{u}^{k-1})^\top(\boldsymbol{s} - \boldsymbol{u}^{k-1}),\end{aligned} \tag{14}$$

where the second equation holds because we optimize a linear objective on a convex polytope $\mathcal{M}_N$ and hence the solution must be achieved on the vertices $\mathrm{Vert}(\mathcal{M}_N)$. Note that if we update $\boldsymbol{u}^k$ by $\boldsymbol{u}^k = (1-\xi_k)\boldsymbol{u}^{k-1} + \xi_k \boldsymbol{s}^k$, we would get the standard Frank-Wolfe (or conditional gradient) algorithm. The difference between our method and Frank-Wolfe is that we greedily minimize the loss $\ell(\boldsymbol{u}^k)$, while the Frank-Wolfe minimizes the linear approximation in (14).

Define $D_{\mathcal{M}_N} := \max_{\boldsymbol{u},\boldsymbol{v}}\{\|\boldsymbol{u}-\boldsymbol{v}\| : \ \boldsymbol{u},\boldsymbol{v} \in \mathcal{M}_N\}$ to be the diameter of $\mathcal{M}_N$. Following (17), we have

$$\begin{aligned}
\ell(\boldsymbol{u}^k) &= \min_{\boldsymbol{q}\in \mathrm{Vert}(\mathcal{M}_N)} \ell\left((1-\xi_k)\boldsymbol{u}^{k-1} + \xi_k \boldsymbol{q}\right) \\
&\le \ell\left((1-\xi_k)\boldsymbol{u}^{k-1} + \xi_k \boldsymbol{s}^k\right) \\
&\le \ell\left(\boldsymbol{u}^{k-1}\right) + \xi_k \nabla \ell(\boldsymbol{u}^{k-1})^\top \left(\boldsymbol{s}^k - \boldsymbol{u}^{k-1}\right) + C\xi_k^2 \qquad (15) \\
&\le (1-\xi_k)\ell\left(\boldsymbol{u}^{k-1}\right) + \xi_k \mathcal{L}_N^* + C\xi_k^2, \qquad (16)
\end{aligned}$$

where we define $C := D^2_{\mathcal{M}_N}$, (15) follows (12), and (16) follows (13). Rearranging this, we get

$$\ell(\boldsymbol{u}^k) - \mathcal{L}_N^* - C\xi_k \le (1-\xi_k)\left(\ell(\boldsymbol{u}^{k-1}) - \mathcal{L}_N^* - C\xi_k\right)$$

By iteratively applying the above inequality, we have

$$\ell(\boldsymbol{u}^k) - \mathcal{L}_N^* - C\xi_k \le \left(\prod_{i=1}^{k}(1-\xi_i)\right)\left(\ell(\boldsymbol{u}^0) - \mathcal{L}_N^* - C\xi_1\right).$$

Taking $\xi_k = 1/k$. We get

$$\ell(\boldsymbol{u}^k) - \mathcal{L}_N^* - \frac{C}{k} \le \frac{1}{k}\left(\ell(\boldsymbol{u}^0) - \mathcal{L}_N^* - C\right).$$

And thus

$$\ell(\boldsymbol{u}^k) \le \frac{1}{k}\left(\ell(\boldsymbol{u}^0) - \mathcal{L}_N^*\right) + \mathcal{L}_N^* = \mathcal{O}\left(\frac{1}{k}\right) + \mathcal{L}_N^*.$$

This completes the proof.

12.2. Proof of Theorem 2

The proof leverages the idea from the proof of Proposition 1 of Chen et al. (2012) for analyzing their *Herding* algorithm, but contains some extra nontrivial argument.

Following the proof of Proposition 1, our problem can be viewed as

$$\min_{\boldsymbol{u}\in\mathcal{M}_N}\left\{\ell(\boldsymbol{u}) := \|\boldsymbol{u}-\mathbf{y}\|^2\right\},$$

with $\mathcal{L}_N^* = \min_{\boldsymbol{u}\in\mathcal{M}_N}\ell(\boldsymbol{u})$, our greedy algorithm can be viewed as starting from $\boldsymbol{u}^0 = 0$ and iteratively updating $\boldsymbol{u}$ by

$$\boldsymbol{u}^k = \frac{k-1}{k}\boldsymbol{u}^{k-1} + \frac{1}{k}\boldsymbol{q}^k, \qquad \boldsymbol{q}^k = \operatorname*{arg\,min}_{\boldsymbol{q}\in\mathrm{Vert}(\mathcal{M}_N)}\left\|\frac{k-1}{k}\boldsymbol{u}^{k-1} + \frac{1}{k}\boldsymbol{q} - \mathbf{y}\right\|^2 \qquad (17)$$

where $\mathrm{Vert}(\mathcal{M}_N) := \{\boldsymbol{\phi}(\boldsymbol{\theta}_1), ..., \boldsymbol{\phi}(\boldsymbol{\theta}_N)\}$ denotes the vertices of $\mathcal{M}_N$. We aim to prove that $\ell(\boldsymbol{u}^k) = O(1/(k\max(1,\gamma))^2)$, under Assumption 2.

Define $\boldsymbol{w}^k = k(\mathbf{y}-\boldsymbol{u}^k)$, then $\ell(\boldsymbol{u}^k) = \left\|\boldsymbol{w}^k\right\|^2/k^2$. Therefore, it is sufficient to prove that $\left\|\boldsymbol{w}^k\right\| = \mathcal{O}(1/(\max(1,\gamma)))$.

Similar to the proof of Proposition 1, we define

$$\begin{aligned}\boldsymbol{s}^{k+1} &= \arg\min_{\boldsymbol{s}\in\mathcal{M}_N} \nabla\ell(\boldsymbol{u}^k)^\top(\boldsymbol{s}-\boldsymbol{u}^k)\\ &= \arg\min_{\boldsymbol{s}\in\mathcal{M}_N} \nabla\ell(\boldsymbol{u}^k)^\top\boldsymbol{s}\\ &= \arg\min_{\boldsymbol{s}\in\mathcal{M}_N} \langle \boldsymbol{w}^k,\ \boldsymbol{s}\rangle.\\ &= \arg\min_{\boldsymbol{s}\in\mathcal{M}_N} \langle \boldsymbol{w}^k,\ (\boldsymbol{s}-\mathbf{y})\rangle.\end{aligned}$$

Because $\mathcal{B}(\mathbf{y},\gamma)$ is included in $\mathcal{M}_N$ by Assumption 2, we have $\boldsymbol{s}' := \mathbf{y} - \gamma\boldsymbol{w}^k/\left\|\boldsymbol{w}^k\right\| \in \mathcal{M}_N$. Therefore

$$\langle \boldsymbol{w}^k,\ (\boldsymbol{s}^{k+1}-\mathbf{y})\rangle = \min_{\boldsymbol{s}\in\mathcal{M}_N}\langle \boldsymbol{w}^k,\ (\boldsymbol{s}-\mathbf{y})\rangle \le \langle \boldsymbol{w}^k,\ (\boldsymbol{s}'-\mathbf{y})\rangle = -\gamma\left\|\boldsymbol{w}^k\right\|.$$

Note that

$$\begin{aligned}\left\|\boldsymbol{w}^{k+1}\right\|^2 &= \min_{\boldsymbol{q}\in\mathrm{Vert}(\mathcal{M}_N)}\left\|k\boldsymbol{u}^k+\boldsymbol{q}-(k+1)\mathbf{y}\right\|^2\\ &= \min_{\boldsymbol{q}\in\mathrm{Vert}(\mathcal{M}_N)}\left\|\boldsymbol{w}^k+\boldsymbol{q}-\mathbf{y}\right\|^2\\ &\le \left\|\boldsymbol{w}^k+\boldsymbol{s}^{k+1}-\mathbf{y}\right\|^2\\ &= \left\|\boldsymbol{w}^k\right\|^2 + 2\langle \boldsymbol{w}^k,\ (\boldsymbol{s}^{k+1}-\mathbf{y})\rangle + \left\|\boldsymbol{s}^{k+1}-\mathbf{y}\right\|^2\\ &\le \left\|\boldsymbol{w}^k\right\|^2 - 2\gamma\left\|\boldsymbol{w}^k\right\| + D_{\mathcal{M}_N}^2,\end{aligned}$$

where $D_{\mathcal{M}_N}$ is the diameter of $\mathcal{M}_N$. Because $\boldsymbol{w}^0 = 0$, using Lemma 6, we have

$$\left\|\boldsymbol{w}^k\right\| \le \max(D_{\mathcal{M}_N},\ D_{\mathcal{M}_N}^2/2, D_{\mathcal{M}_N}^2/(2\gamma)) = \mathcal{O}\left(\frac{1}{\min(1,\gamma)}\right),\quad \forall k=1,2,\ldots,$$

This proves that $\ell(\boldsymbol{u}^k) = \frac{\left\|\boldsymbol{w}^k\right\|^2}{k^2} = \mathcal{O}\left(\frac{1}{k^2\min(1,\gamma)^2}\right)$.

Lemma 6. *Assume $\{z_k\}_{k\ge 0}$ is a sequence of numbers satisfying $z_0 = 0$ and*

$$|z_{k+1}|^2 \le |z_k|^2 - 2\gamma|z_k| + C,\quad \forall k = 0,1,2,\ldots$$

where C and γ are two positive numbers. Then we have $|z_k| \le \max(\sqrt{C},\ C/2, C/(2\gamma))$ for all $k = 0,1,2,\ldots$.

Proof. We prove $|z_k| \le \max(\sqrt{C},\ C/2, C/(2\gamma)) := u_*$ by induction on k. Because $z_0 = 0$, the result holds for $k = 0$. Assume $|z_k| \le u_*$, we want to prove that $|z_{k+1}| \le u_*$ also holds.

Define $f(z) = z^2 - 2\gamma z + C$. Note that the maximum of $f(z)$ on an interval is always achieved on the vertices, because $f(z)$ is convex.

Case 1: If $|z_k| \le C/(2\gamma)$, then we have

$$|z_{k+1}|^2 \le f(|z_k|) \le \max_z\left\{f(z):\ \ z\in[0,\ C/(2\gamma)]\right\} = \max\left\{f(0),\ f(C/(2\gamma))\right\} = \max\left\{C,\ C^2/(4\gamma^2)\right\} \le u_*^2.$$

Case 2: If $|z_k| \ge C/(2\gamma)$, then we have

$$|z_{k+1}|^2 \le |z_k|^2 - 2\gamma|z_k| + C \le |z_k|^2 \le u_*^2.$$

In both cases, we have $|z_{k+1}| \le u_*$. This completes the proof. □

12.3. Proof of Theorem 3

We first introduce the following Lemmas.

Lemma 7. *Under the Assumption 1, 3, 4, 5 and 6. For any $\delta > 0$, when N is sufficient large, with probability at least $1-\delta$,*

$$\mathcal{B}\left(\boldsymbol{y}, \frac{1}{2}\gamma^*\right) \subseteq \textit{conv}\left\{\boldsymbol{\phi}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \textit{supp}(\rho_T^N)\right\}.$$

Here ρ_T^N is the distribution of the weight of the large network with N neurons trained by gradient descent.

12.3.1. Proof of Theorem 3

The above lemmas directly imply Theorem 3.

12.3.2. Proof of Lemma 7

In this proof, we simplify the statement that 'for any $\delta > 0$, when N is sufficiently large, event E holds with probability at least $1-\delta$' by 'when N is sufficiently large, with high probability, event E holds'.

By the Assumption 5, there exists $\gamma^* > 0$ such that

$$\mathcal{B}(\mathbf{y}, \gamma^*) \subseteq \text{conv}\left\{\boldsymbol{\phi}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \text{supp}(\rho_T^\infty)\right\} = \mathcal{M}.$$

Given any $\boldsymbol{\theta} \in \text{supp}(\rho_T^\infty)$, define

$$\boldsymbol{\phi}^N(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}' \in \text{supp}\left(\rho_T^N\right)}{\arg\min} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}') - \boldsymbol{\phi}(\boldsymbol{\theta})\right\|$$

where $\boldsymbol{\phi}^N(\boldsymbol{\theta})$ is the best approximation of $\boldsymbol{\phi}(\boldsymbol{\theta})$ using the points $\boldsymbol{\phi}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i \in \text{supp}(\rho_T^N)$.

Using Lemma 11, by choosing $\epsilon = \gamma^*/6$, when N is sufficiently large, we have

$$\sup_{\boldsymbol{\theta} \in \text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}^N(\boldsymbol{\theta})\right\| \leq \gamma^*/6, \tag{18}$$

with high probability. (18) implies that $\mathcal{M}_N$ can approximate $\mathcal{M}$ for large N. Since $\mathcal{M}$ is assumed to contain the ball centered at $\mathbf{y}$ with radius γ^*, as $\mathcal{M}_N$ approximates $\mathcal{M}$, intuitively $\mathcal{M}_N$ would also contain the ball centered at $\mathbf{y}$ with a smaller radius. And below we give a rigorous proof for this intuition.

Step 1: $\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \gamma^*/6$**.** When N is sufficiently large, with high probability, we have

$$\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \sum_{i=1}^{M} q_i \left\|\boldsymbol{\phi}^N(\boldsymbol{\theta}_i^*) - \boldsymbol{\phi}(\boldsymbol{\theta}_i^*)\right\| \leq \gamma^*/6.$$

Step 2 $\mathcal{B}\left(\hat{\mathbf{y}}, \frac{5}{6}\gamma^*\right) \subseteq \mathcal{M}$ By step one, with high probability, $\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \gamma^*/4$, which implies that $\hat{\mathbf{y}} \in \mathcal{B}(\mathbf{y}, \gamma^*/4) \subseteq \mathcal{B}(\mathbf{y}, \gamma^*) \subseteq \mathcal{M}$. Also, for any $A \in \partial\mathcal{M}$ (here $\partial\mathcal{M}$ denotes the boundary of $\mathcal{M}$), we have

$$\|\hat{\mathbf{y}} - A\| \geq \|\mathbf{y} - A\| - \|\mathbf{y} - \hat{\mathbf{y}}\| \geq \gamma^* - \gamma^*/4.$$

This gives that $\mathcal{B}\left(\hat{\mathbf{y}}, \frac{5}{6}\gamma^*\right) \subseteq \mathcal{M}$.

Step 3 $\mathcal{B}\left(\hat{\mathbf{y}}, \frac{2}{3}\gamma^*\right) \subseteq \mathcal{M}_N$ Notice that $\hat{\mathbf{y}}$ is a point in $\mathbb{R}^m$ and suppose that A belongs to the boundary of $\mathcal{M}_N$ (denoted by $\partial\mathcal{M}_N$) such that

$$\|\hat{\mathbf{y}} - A\| = \min_{\tilde{A} \in \partial\mathcal{M}_N} \left\|\hat{\mathbf{y}} - \tilde{A}\right\|.$$

We prove by contradiction. Suppose that we have $\|\hat{\mathbf{y}} - A\| < \frac{2}{3}\gamma^*$.

Using support hyperplane theorem, there exists a hyperplane $P = \{\boldsymbol{u} : \langle \boldsymbol{u} - A, \boldsymbol{v}\rangle = 0\}$ for some nonempty vector $\boldsymbol{v}$, such that $A \in P$ and

$$\sup_{\boldsymbol{q} \in \mathcal{M}_N} \langle \boldsymbol{q}, \boldsymbol{v}\rangle \leq \langle A, \boldsymbol{v}\rangle.$$

We choose $A' \in P$ such that $A' - \hat{\mathbf{y}} \perp P$ (A and A' can be the same point). Notice that

$$\|\hat{\mathbf{y}} - A'\|^2 = \|\hat{\mathbf{y}} - A + A - A'\|^2 = \|\hat{\mathbf{y}} - A\|^2 + \|A - A'\|^2 + 2\langle \hat{\mathbf{y}} - A, A - A' \rangle .$$

Since $A' - \hat{\mathbf{y}} \perp P$ and $A, A' \in P$, we have $\langle \hat{\mathbf{y}} - A, A - A' \rangle = 0$ and thus $\|\hat{\mathbf{y}} - A'\| \leq \|\hat{\mathbf{y}} - A\| < \frac{2}{3}\gamma^*$. We have

$$A' \in \mathcal{B}\left(\hat{\mathbf{y}}, \|\hat{\mathbf{y}} - A\|\right) \subseteq \mathcal{B}\left(\hat{\mathbf{y}}, \frac{2}{3}\gamma^*\right) \subseteq \mathcal{B}\left(\hat{\mathbf{y}}, \frac{5}{6}\gamma^*\right) \subseteq \mathcal{M}.$$

Notice that as both $\hat{\mathbf{y}}, A' \in \mathcal{M}$ we choose $\lambda \geq 1$ such that $\hat{\mathbf{y}} + \lambda(A' - \hat{\mathbf{y}}) \in \partial\mathcal{M}$, where $\partial\mathcal{M}$ denotes the boundary of $\mathcal{M}$. Define $B = \hat{\mathbf{y}} + \lambda(A' - \hat{\mathbf{y}})$. As we have shown that $\mathcal{B}\left(\hat{\mathbf{y}}, \frac{5}{6}\gamma^*\right) \subseteq \mathcal{M}$, we have $\|\hat{\mathbf{y}} - B\| \geq \frac{5}{6}\gamma^*$. And thus

$$\begin{aligned}
\|B - A'\| &= \|B - \hat{\mathbf{y}}\| - \|\hat{\mathbf{y}} - A'\| \\
&> \frac{5}{6}\gamma^* - \frac{2}{3}\gamma^* \\
&> \frac{1}{6}\gamma^*.
\end{aligned}$$

Also notice that

$$\begin{aligned}
\langle B - A, \boldsymbol{v} \rangle &= \langle \hat{\mathbf{y}} + \lambda(A' - \hat{\mathbf{y}}) - A, \boldsymbol{v} \rangle \\
&= (1 - \lambda)\langle \hat{\mathbf{y}} - A, \boldsymbol{v} \rangle + \lambda \langle A' - A, \boldsymbol{v} \rangle \\
&= (1 - \lambda)\langle \hat{\mathbf{y}} - A, \boldsymbol{v} \rangle \\
&\geq 0.
\end{aligned}$$

This implies that B and $\mathcal{M}$ are on different side of P.

With high probability, we are able to find $D \in \{\boldsymbol{\phi}(\boldsymbol{\theta}); \boldsymbol{\theta} \in \text{supp}(\rho_T^N)\}$ such that

$$\|D - B\| \leq \frac{\gamma^*}{6}.$$

By the definition, $D \in \mathcal{M}_N$ and thus $\langle D - A, \boldsymbol{v} \rangle \leq 0$ as shown by the supporting hyperplane theorem. Also remind that $\langle B - A, \boldsymbol{v} \rangle \geq 0$. These allow us to choose $\lambda' \in [0, 1]$ such that

$$\langle \lambda' D + (1 - \lambda')B - A, \boldsymbol{v} \rangle = 0.$$

We define $E = \lambda' D + (1 - \lambda')B$ and thus $E \in P$. Notice that

$$\|B - E\| = \|B - \lambda' D - (1 - \lambda')B\| = \lambda' \|B - D\| \leq \|B - D\| \leq \frac{\gamma^*}{6}.$$

Also,

$$\|B - E\|^2 = \|B - A' + A' - E\|^2 = \|B - A'\|^2 + \|A' - E\|^2 + 2\langle B - A', A' - E \rangle .$$

As $B - A' \perp P$ and $A', E \in P$, we have $\langle B - A', A' - E \rangle = 0$, which implies that $\|B - E\| \geq \|B - A'\| > \frac{1}{6}\gamma^*$, which makes contradiction.

Step 4 $\mathcal{B}\left(\mathbf{y}, \frac{1}{2}\gamma^*\right) \subseteq \mathcal{M}_N$ As for sufficiently large N, we have $\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \frac{1}{6}\gamma^*$ and thus

$$\mathcal{B}\left(\mathbf{y}, \frac{1}{2}\gamma^*\right) \subseteq \mathcal{B}\left(\hat{\mathbf{y}}, \frac{2}{3}\gamma^*\right) \subseteq \mathcal{M}_N.$$

13. Technical Lemmas

Lemma 8. *Under assumption 1 and 3, for any N, at training time $T < \infty$, for any $\boldsymbol{\theta} \in$ supp(ρ_T^N) or $\boldsymbol{\theta} \in$ supp(ρ_T^∞), we have $\|\boldsymbol{\theta}\| \leq C$, $\|\boldsymbol{\phi}(\boldsymbol{\theta})\| \leq C$ and $\|\boldsymbol{\phi}(\boldsymbol{\theta})\|_{Lip} \leq C$ for some constant $C < \infty$.*

Lemma 9. *Suppose $\boldsymbol{\theta}_i \in \mathbb{R}^d$, $i = 1, ..., N$ are i.i.d. samples from some distribution ρ and $\Omega \subseteq \mathbb{R}^d$ is bounded. For any radius $r_B > 0$ and $\delta > 0$, define the following two sets*

$$A = \left\{ \boldsymbol{\theta}_B \in \Omega \,\middle|\, \mathbb{P}_{\boldsymbol{\theta} \sim \rho} \left(\boldsymbol{\theta} \in \mathcal{B}\left(\boldsymbol{\theta}_B, r_B \right) \right) > \frac{4}{N} \left((d+1) \log\left(2N\right) + \log\left(8/\delta\right) \right) \right\}$$
$$B = \left\{ \boldsymbol{\theta}_B \in \Omega \,\middle|\, \left\| \boldsymbol{\theta}_B - \boldsymbol{\theta}_B^N \right\| \leq r_B \right\},$$

where $\boldsymbol{\theta}_B^N = \underset{\boldsymbol{\theta}' \in \{\boldsymbol{\theta}_i\}_{i=1}^N}{\arg\min} \left\| \boldsymbol{\theta}_B - \boldsymbol{\theta}' \right\|$. With probability at least $1 - \delta$, $A \subseteq B$.

Lemma 10. *For any $\delta > 0$ and $\epsilon > 0$, when N is sufficiently large (N depends on δ), with probability at least $1 - \delta$, we have*

$$\sup_{\boldsymbol{\theta} \in supp(\rho_T^\infty)} \left\| \boldsymbol{\phi}(\boldsymbol{\theta}) - \bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) \right\| \leq \epsilon,$$

where $\bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) = \underset{\boldsymbol{\phi}(\bar{\boldsymbol{\theta}}') \in \{\boldsymbol{\phi}(\bar{\boldsymbol{\theta}}_i)\}_{i=1}^N}{\arg\min} \left\| \boldsymbol{\phi}(\bar{\boldsymbol{\theta}}') - \boldsymbol{\phi}(\boldsymbol{\theta}) \right\|$ and $\bar{\boldsymbol{\theta}}_i$ are i.i.d. samples from ρ_T^∞.

Lemma 11. *For any $\delta > 0$ and $\epsilon > 0$, when N is sufficiently large (N depends on δ), with probability at least $1 - \delta$, we have*

$$\sup_{\boldsymbol{\theta} \in supp(\rho_T^\infty)} \left\| \boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}^N(\boldsymbol{\theta}) \right\| \leq \epsilon,$$

where $\boldsymbol{\phi}^N(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}' \in supp\left(\rho_T^N\right)}{\arg\min} \left\| \boldsymbol{\phi}(\boldsymbol{\theta}') - \boldsymbol{\phi}(\boldsymbol{\theta}) \right\|$.

13.1. Proof of Lemma 8

We prove the case of training network with N neurons. Notice that

$$\begin{aligned}
\left\| \frac{\partial}{\partial t} \boldsymbol{\theta}(t) \right\| &= \left\| \boldsymbol{g}[\boldsymbol{\theta}(t), \rho_t^N] \right\| \\
&= \left\| \mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left(y - f_{\rho_t^N}(\boldsymbol{x}) \right) \nabla_{\boldsymbol{\theta}} \sigma(\boldsymbol{\theta}(t), \boldsymbol{x}) \right\| \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left(y - f_{\rho_t^N}(\boldsymbol{x}) \right)^2} \sqrt{\mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left\| \nabla_{\boldsymbol{\theta}} \sigma(\boldsymbol{\theta}(t), \boldsymbol{x}) \right\|^2} \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left(y - f_{\rho_0^N}(\boldsymbol{x}) \right)^2} \sqrt{\mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left\| \nabla_{\boldsymbol{\theta}} \sigma(\boldsymbol{\theta}(t), \boldsymbol{x}) \right\|^2}
\end{aligned}$$

Notice that by the assumption 1, we have $\sqrt{\mathbb{E}_{\boldsymbol{x}, y \sim \mathcal{D}} \left(y - f_{\rho_0^N}(\boldsymbol{x}) \right)^2} \leq C$. Remind that $\boldsymbol{\theta}(t) = [\boldsymbol{a}(t), b(t)]$, $\sigma(\boldsymbol{\theta}(t), \boldsymbol{x}) = b(t)\sigma_+(\boldsymbol{a}^\top(t)\boldsymbol{x})$. Thus we have

$$\left| \frac{\partial}{\partial t} b(t) \right| \leq C \left\| \sigma_+ \right\|_\infty.$$

And thus for any $i \in \{1, ..., N\}$, $\sup_{t \in [0,T]} \left\| b_i(t) \right\| \leq \int_0^T \left\| \frac{\partial}{\partial t} b_i(s) \right\| ds \leq TC$. Also

$$\begin{aligned}
\left\| \frac{\partial}{\partial t} \boldsymbol{a}(t) \right\| &\leq C |b(t)| \left\| \sigma_+' \right\|_\infty \sqrt{\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left\| \boldsymbol{x} \right\|^2} \\
&\leq TC.
\end{aligned}$$

By assumption 3, that $\left\| \boldsymbol{\theta}_0(t) \right\| \leq C$, we have

$$\sup_{t \in [0,T]} \left\| \boldsymbol{\theta}_i(t) \right\| \leq \int_0^T \left\| \frac{\partial}{\partial t} \boldsymbol{\theta}_i(s) \right\| ds \leq T^2 C.$$

Notice that this also holds to training the network with infinite number of neurons. Notice that $\|\boldsymbol{\phi}(\boldsymbol{\theta})\| = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\sigma^2(\boldsymbol{\theta},\boldsymbol{x}^{(j)})} \leq CT$. And

$$\begin{aligned}\|\boldsymbol{\phi}(\boldsymbol{\theta})\|_{\text{Lip}} &= \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\frac{\|\boldsymbol{\phi}(\boldsymbol{\theta}_1)-\boldsymbol{\phi}(\boldsymbol{\theta}_2)\|}{\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|}\\ &= \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\frac{\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\sigma(\boldsymbol{\theta}_1,\boldsymbol{x}^{(j)})-\sigma(\boldsymbol{\theta}_2,\boldsymbol{x}^{(j)})\right)^2}}{\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|}\\ &\leq TC\,\|\sigma_+\|_{\text{Lip}}+\|\sigma_+\|_{\infty}.\end{aligned}$$

Thus given any $T<\infty$, all those three quantities can be bounded by some constant.

13.2. Proof of Lemma 9

The following proof follows line 1 and and line 2 of the proof of Lemma 16 of (Chaudhuri & Dasgupta, 2010).

Define $g_{\boldsymbol{\theta}_B}(\boldsymbol{\theta}) = \mathbb{I}\{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_B,r_B)\}$ and $\beta_N = \sqrt{(4/N)(d_{\text{VC}}\log 2N+\log(8/\delta))}$, where d_{VC} is the VC dimension of the function class $\mathcal{G}=\{g_{\boldsymbol{\theta}_B},\boldsymbol{\theta}_B\in\Omega\}$ and thus $d_{\text{VC}}\leq d+1$ (Dudley, 1979). Let $\mathbb{E}g_{\boldsymbol{\theta}_B} = \mathbb{P}_{\boldsymbol{\theta}\sim\rho}(\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_B,r_B))$ and $\mathbb{E}_N g_{\boldsymbol{\theta}_B} = \sum_{i=1}^{N} g_{\boldsymbol{\theta}_B}(\boldsymbol{\theta}_i)/N$. So

$$A=\{\boldsymbol{\theta}_B\,|\,\mathbb{E}g_{\boldsymbol{\theta}_B}>\beta_N^2\}$$

and we further define

$$A_2=\{\boldsymbol{\theta}_B\,|\,\mathbb{E}_N g_{\boldsymbol{\theta}_B}>0\}.$$

From theorem 15 of (Chaudhuri & Dasgupta, 2010) (which is a rephrase of the generalization bound), we know that: for any $\delta>0$, with probability at least $1-\delta$, the following holds for all $g_{\boldsymbol{\theta}_B}\in\mathcal{G}$,

$$\mathbb{E}g_{\boldsymbol{\theta}_B}-\mathbb{E}_N g_{\boldsymbol{\theta}_B}\leq\beta_N\sqrt{\mathbb{E}g_{\boldsymbol{\theta}_B}} \tag{19}$$

Notice that for any $g_{\boldsymbol{\theta}_B}$ which satisfies (19),

$$\mathbb{E}g_{\boldsymbol{\theta}_B}>\beta_N^2\Rightarrow\mathbb{E}_N g_{\boldsymbol{\theta}_B}>0$$

So this means: for any $\delta>0$, with probability at least $1-\delta$,

$$A\subseteq A_2=B$$

where the last equality follows from the following:

$$A_2=\{\boldsymbol{\theta}_B\,|\,\mathbb{E}_N g_{\boldsymbol{\theta}_B}>0\}=\{\text{there exists some }\boldsymbol{\theta}_i\text{ such that }\boldsymbol{\theta}_i\in\mathcal{B}(\boldsymbol{\theta}_B,r_B)\}=B$$

13.3. Proof of Lemma 10

Given $\epsilon>0$, we choose r_0 sufficiently small such that $Cr_0\leq\epsilon$ (here C is some constant defined in Lemma 8). For this choice of r_0, given the corresponding p_0 (defined in assumption 6), for any $\delta>0$, there exists $N(\delta)$ such that $\forall N\geq N(\delta)$, we have

$$p_0>\frac{4}{N}\left((d+1)\log(2N)+\log(8/\delta)\right):=\beta_N^2.$$

And thus from assumption 6, we have

$$\forall\boldsymbol{\theta}\in\text{supp}(\rho_T^{\infty}),\ \mathbb{P}_{\boldsymbol{\theta}'\sim\rho_T^{\infty}}\left(\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},r_0)\right)\geq p_0>\beta_N^2.$$

This implies

$$\text{supp}(\rho_T^{\infty})\subseteq A=\left\{\boldsymbol{\theta}_B\,|\,\mathbb{P}_{\boldsymbol{\theta}\sim\rho}(\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_B,r_0))>\beta_N^2\right\}$$

From Lemma 9 (set $r_B=r_0$), we know: with probability at least $1-\delta$,

$$A\subseteq B=\left\{\boldsymbol{\theta}_B\in\Omega\,\middle|\,\left\|\boldsymbol{\theta}_B-\boldsymbol{\theta}_B^N\right\|\leq r_0\right\},$$

Thus, with probability at least $1-\delta$,

$$\text{supp}(\rho_T^\infty) \subseteq B$$

and this means: with probability at least $1-\delta$, we have

$$\forall \boldsymbol{\theta} \in \text{supp}(\rho_T^\infty),\ \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^N\right\| \leq r_0.$$

The result concludes from

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}^N(\boldsymbol{\theta})\right\| \\
\leq &\sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}(\boldsymbol{\theta}^N)\right\| \\
\leq &\sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} C\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^N\right\| \\
\leq &Cr_0 \leq \epsilon.
\end{aligned}$$

Here the last inequality uses Lemma 8.

13.4. Proof of Lemma 11

In this proof, we simplify the statement that 'for any $\delta > 0$, when N is sufficiently large, event E holds with probability at least $1-\delta$' by 'when N is sufficiently large, with high probability, event E holds'.

Suppose that $\boldsymbol{\theta}_i$, $i \in [N]$ is the weight of neurons of network $f_{\rho_T^N}$. Given any $\boldsymbol{\theta} \in \text{supp}(\rho_T^\infty)$, define

$$\boldsymbol{\phi}^N(\boldsymbol{\theta}) = \mathop{\arg\min}_{\boldsymbol{\phi}(\boldsymbol{\theta}')\in\text{Vert}(\mathcal{M}_N)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}') - \boldsymbol{\phi}(\boldsymbol{\theta})\right\|.$$

Notice that the training dynamics of the network with N neurons can be characterized by

$$\begin{aligned}
\frac{\partial}{\partial t}\boldsymbol{\theta}_i(t) &= \boldsymbol{g}[\boldsymbol{\theta}_i(t), \rho_t^N], \\
\boldsymbol{\theta}_i(0) &\overset{\text{i.i.d.}}{\sim} \rho_0.
\end{aligned}$$

Here $\boldsymbol{g}[\boldsymbol{\theta}, \rho] = \mathbb{E}_{\boldsymbol{x},y\sim\mathcal{D}}\left(y - f_\rho(\boldsymbol{x})\right)\nabla_{\boldsymbol{\theta}}\sigma(\boldsymbol{\theta}, \boldsymbol{x})$. We define the following coupling dynamics:

$$\begin{aligned}
\frac{\partial}{\partial t}\bar{\boldsymbol{\theta}}_i(t) &= \boldsymbol{g}[\bar{\boldsymbol{\theta}}_i(t), \rho_t^\infty], \\
\bar{\boldsymbol{\theta}}_i(0) &= \boldsymbol{\theta}_i(0).
\end{aligned}$$

Notice that at any time t, $\bar{\boldsymbol{\theta}}_i(t)$ can be viewed as i.i.d. sample from ρ_t^∞. We define $\hat{\rho}_t^N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{\boldsymbol{\theta}}_i(t)}(\boldsymbol{\theta})$. Notice that by our definition $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i(T)$ and we also define $\bar{\boldsymbol{\theta}}_i = \bar{\boldsymbol{\theta}}_i(T)$. Using the propagation of chaos argument as Mei et al. (2019) (Proposition 2 of Appendix B.2), for any $T < \infty$, for any $\delta > 0$, we have

$$\sup_{t\in[0,T]} \max_{i\in\{1,..,N\}} \left\|\bar{\boldsymbol{\theta}}_i(t) - \boldsymbol{\theta}_i(t)\right\| \leq \frac{C}{\sqrt{N}}\left(\sqrt{\log N} + \sqrt{\log 1/\delta}\right).$$

By Lemma 10 and the bound above, when N is sufficiently large, with high probability, we have

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta})\right\| &\leq \epsilon/2 \\
\max_{i\in[N]} \left\|\bar{\boldsymbol{\theta}}_i(T) - \boldsymbol{\theta}_i(T)\right\| &\leq \frac{\epsilon}{2C},
\end{aligned}$$

where $C = \|\boldsymbol{\phi}\|_{\text{Lip}}$ and

$$\bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) = \mathop{\arg\min}_{\boldsymbol{\theta}'\in\text{supp}(\hat{\rho}_T^N)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta})\right\|.$$

We denote $\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}} \in \text{supp}(\hat{\rho}_T^N)$ such that $\bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) = \boldsymbol{\phi}(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}})$. It implies that

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}^N(\boldsymbol{\theta})\right\| &\leq \sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}\left(\boldsymbol{\theta}_{i_{\boldsymbol{\theta}}}\right)\right\| \\
&= \sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) + \bar{\boldsymbol{\phi}}^N(\boldsymbol{\theta}) - \boldsymbol{\phi}\left(\boldsymbol{\theta}_{i_{\boldsymbol{\theta}}}\right)\right\| \\
&= \sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}\left(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right) + \boldsymbol{\phi}\left(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right) - \boldsymbol{\phi}\left(\boldsymbol{\theta}_{i_{\boldsymbol{\theta}}}\right)\right\| \\
&\leq \sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}\left(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right)\right\| + \sup_{\boldsymbol{\theta}\in\text{supp}(\rho_T^\infty)} \left\|\boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\phi}\left(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right)\right\| \\
&\leq \epsilon/2 + \max_{i\in[N]} \left\|\bar{\boldsymbol{\theta}}_i(T) - \boldsymbol{\theta}_i(T)\right\| \|\boldsymbol{\phi}\|_{\text{Lip}} \\
&\leq \epsilon.
\end{aligned}$$