OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Strong Lottery Ticket Hypothesis with ε–perturbation

Zheyang Xiong* ZHEYANG.XIONG@RICE.EDU
Fangshuo Liao* FANGSHUO.LIAO@RICE.EDU
Anastasios Kyrillidis ANASTASIOS@RICE.EDU
Rice University, Houston, Texas

Abstract

The strong Lottery Ticket Hypothesis (LTH) [18, 25] claims that there exists a subnetwork in a sufficiently large, randomly initialized neural network that approximates some target neural network without the need of training. This work extends the theoretical guarantee of the strong LTH literature to a scenario more similar to the original LTH, by generalizing the weight change achieved in the pre-training step to some perturbation around the initialization. In particular, we focus on the following open questions: *By allowing an ε-scale perturbation on the random initial weights, can we reduce the over-parameterization requirement for the candidate network in the strong LTH? Furthermore, does the weight change by SGD coincide with a good set of such perturbation?*

We answer the first question by first extending the theoretical result on subset sum [14] to a scenario that allows perturbation on the candidates, and forms the conjecture of the perturbed strong LTH by applying our generalized theoretical result to a neural network formulation. To answer the second question, we show via experiments that, when a larger perturbation is allowed, the required over-parameterization of the strong LTH decreases, and the final accuracy after pruning increases.

1. Introduction

Pruning techniques for over-parameterized neural networks have drawn growing attention in recent years [1, 10–13, 19–21, 26]. Amongst them, the *Lottery Ticket Hypothesis* (LTH) [8, 9] claims the existence of a small (sparse) subnetwork within a large (dense) neural network such that, when trained in isolation, achieves comparable or even better performance than the original dense network. Such subnetworks can be identified by pre-training the dense network and pruning it based on the magnitude of the learned weights [8]. Currently, to the best of our knowledge, the LTH lacks of any rigorous theoretical guarantees that justify superior performance of the subnetwork, especially under the pretraining-based pruning; yet, it has been proven to be effective in practice.

The *Strong Lottery Ticket Hypothesis* [18, 25] leverages a different pruning scheme: given a target dense neural network, and a randomly initialized, *sufficiently over-parameterized* candidate network, there exists a subnetwork in the latter that approximates the former arbitrarily well without the need of training. While we require a significant over-parameterization in the randomly initialized network, the strong LTH enjoys extensive theoretical guarantees [15–17]. Yet, the same theory hardly applies to the LTH, as strong LTH assumes that the candidate weights pruned are fixed at initialization. *That LTH pruning is based on weights modified by pre-training motivates us to analyze the approximation behavior that emerges beyond the randomness in the candidate weights.*

* Equal Contribution

Further study on the pre-training process of LTH shows that the lottery ticket emerges in the early stage of training [23]. This implies that converging to a small training loss is not necessarily the intent of pre-training in the LTH procedure; *in other words, achieving small loss does not necessarily explain how and why pre-training helps pruning in LTH.* Instead, one could hypothesize that the pre-training –based on loss minimization– could *guide* the weight perturbation to a direction that facilitate the pruning. Based on this hypothesis, our work extends the theoretical guarantee of strong LTH to a scenario more similar to the original LTH, by generalizing the weight change in the pre-training step to some perturbation around the initialization. Our central question is as follows:

> *"By allowing an ε-scale perturbation on the random initial weights, can we reduce the over-parameterization requirement for the candidate network in the strong LTH? Furthermore, does the weight change by SGD coincide with a good set of such perturbation?"*

To be more specific, let $f_{\boldsymbol{\theta}}$ be a neural network with a set of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. Formally, an ε-perturbation is a mapping $\mathcal{P} : \mathbb{R}^d \to \mathbb{R}^d$ such that the maximum entrywise perturbation is bounded in absolute value by ε, i.e., $\|\mathcal{P}(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_\infty \leq \varepsilon$. We point out that this definition of perturbation generalizes to two existing scenarios: when $\varepsilon = 0$ –i.e., we allow no weight perturbation– the question above reduces to the original strong LTH. When $\varepsilon = \infty$ –i.e., we allow arbitrarily large weight perturbation– the required over-parameterization for the candidate network is at most the size of the target network. In this case, one would use gradient based optimizers such as SGD to find such weight perturbation, but often without the need of pruning. Yet, both cases cover only one aspect in pruning and perturbation.

In this paper, we study the inter-dependence of the two aspects above by treating ε as a variable. In particular, we show that a larger perturbation scale ε, which corresponds to a larger amount of training, would alleviate the over-parameterization requirement, while keeping the accuracy of the pruned neural network the same. Our contributions can be summarized as below:

- We consider a generalized version of the subset sum problem where each candidate in the summation is allowed perturbation bounded by a fixed scale ε. We extend the analysis of the subset sum [14] to our generalized version, and show that when a larger perturbation is allowed, the required size of the candidate set can be reduced. We empirically validate our theoretical result on the perturbed subset sum problem.
- Applying the theoretical result above to neural networks, we conjecture that, when an ε-scale perturbation is allowed, the strong LTH on randomly initialized neural network requires less over-parameterization to achieve a specific approximation error.
- On neural networks, we empirically show that $i)$ the perturbation that alleviates the overparameterization requirement of the strong LTH can be obtained by projected SGD on the initialized weights; and, $ii)$under fixed overparameterization, neural networks with a larger freedom over the level of perturbation achieves a higher accuracy after pruning. This result establish the connection between the amount of pre-training and the accuracy of the pruned network.

2. Related Works

Lottery Ticket Hypothesis. Several works attempt to explain the LTH theoretically. [7] empirically study the behavior of gradient flow in the pruned network. [24] assumes that the optimal mask is given, and proves that the pruned network achieves faster convergence and better generalization when trained from initialization. [22] provides a theoretical guarantee of the loss in a

pruning-after-training fashion. However, these works differ from ours in the following: i) they usually consider neuron pruning on small neural network archictures (e.g., [22] focuses on a two-layer neural network with smooth activations), while our work considers weight pruning of a deep ReLU neural network; ii) they consider minimizing the loss on a specific dataset, and require an over-parameterization that scales quadratically with the number of samples, while we approximate a target network with a fixed architecture in terms of the function norm, and require an over-parameterization that scales with the width of the target network.

Strong Lottery Ticket Hypothesis. The strong LTH originates from the empirical observation that, by fixing the weights at initialization and learning the mask the weights, one can identify subnetworks that achieve comparable accuracy to the dense one with learned weights [25]. [18] improved this idea by proposing the `edge-popup` algorithm to efficiently learn the mask. [15] first proved such hypothesis under the assumption that the dense network's size scales polynomially with the target network's width and depth. Leveraging the advantage of weight decomposition and theoretical results on the subset sum problem [14], [16] and [17] improved the over-parameterization requirement to a logarithm factor times the size of the target network. Later work explores different variations of the strong LTH: [6] and [2] show that the strong LTH holds in the case of convolutional neural networks. [3] extends the strong LTH proof to a universial family of functions. Finally, [4] further reduces the over-parameterization requirement by employing the iterative randomization.

3. Notations and Setup

Notation. For a vector $\mathbf{a}$, we use $\|\mathbf{a}\|_2$ to denote its ℓ_2 (Euclidean) norm, and $\|\mathbf{a}\|_\infty$ to denote its ℓ_∞ norm. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_{\max} = \max_{ij} |A_{ij}|$ to denote its max norm. Moreover, $\texttt{Unif}\,(I)$ denotes the uniform distribution on the interval I, and $\texttt{Geom}\,(\cdot)$ and $\texttt{Bin}\,(\cdot,\cdot)$ denotes the geometric and binomial distributions, respectively. We use $\sigma\,(a) = \max\{0, a\}$ to denote the ReLU activation. We use $\min\{\cdot\}$ refers to the entrywise minimum and abs$(\cdot)$ refers to the entrywise absolute value.
Setup. Similar to [17], our focus is to approximate an L-layer, ReLU activated target multi-layer perceptron (MLP) $f(\mathbf{x})$ by pruning a $2L$-layer, ReLU activated candidate MLP $g_{\boldsymbol{\theta}}(\mathbf{x})$. For a input vector $\mathbf{x} \in \mathbb{R}^{d_0}$, we assume $f(\mathbf{x})$ and $g_{\boldsymbol{\theta}}(\mathbf{x})$ are represented by:

$$f(\mathbf{x}) = \mathbf{W}^L \sigma\left(\mathbf{W}^{L-1} \ldots \sigma\left(\mathbf{W}^1 \mathbf{x}\right)\right) g(\mathbf{x}) = \mathbf{U}^{2L} \sigma\left(\mathbf{U}^{2L-1} \ldots \sigma\left(\mathbf{U}^1 \mathbf{x}\right)\right)$$

We consider the pruning of $g_{\boldsymbol{\theta}}(\mathbf{x})$ with masks for the weights $\mathcal{M} = \{\mathbf{M}^\ell\}_{\ell=1}^{2L}$, denoted as $g_{\mathcal{M},\boldsymbol{\theta}}(\mathbf{x})$:

$$g_{\mathcal{M},\boldsymbol{\theta}}(\mathbf{x}) = \left(\mathbf{M}^{2L} \odot \mathbf{U}^{2L}\right) \sigma\left(\left(\mathbf{M}^{2L-1} \odot \mathbf{U}^{2L-1}\right) \ldots \sigma\left(\left(\mathbf{M}^1 \odot \mathbf{U}^1\right) \mathbf{x}\right)\right)$$

For a set of weights $\boldsymbol{\theta}$, we consider its ε-perturbation $\mathcal{P}(\boldsymbol{\theta}) = \{\mathcal{P}\left(\mathbf{U}^\ell\right)\}_{\ell=1}^{2L}$, with $\mathcal{P}$ applied to each $\mathbf{U}^\ell$ such that $\|\mathcal{P}(\mathbf{U}^\ell) - \mathbf{U}^\ell\|_{\max} = \max_{ij} |\mathcal{P}(\mathbf{U}^\ell)_{ij} - \mathbf{U}^\ell_{ij}| \leq \varepsilon$. Our focus in this paper is on the approximation error defined as:

$$\mathcal{L}(f, g) = \min_{\mathcal{M},\mathcal{P}} \sup_{\mathbf{x}\in\mathcal{B}} \left\| f(\mathbf{x}) - g_{\mathcal{M},\mathcal{P}(\theta)}\left(\mathbf{x}\right)\right\|_2. \tag{1}$$

4. Subset Sum with ε-Perturbation

For each layer in the target network, [17] constructed a two-layer subnetwork with block structure, such that each block approximates a single entry in the weight matrix of the target network. In

particular, they obtain a logarithmic-scale over-parameterization by formulating the approximation as a subset sum problem [5, 14] Given a candidate set of values $\{x_i\}_{i=1}^n$ of size n and a target value z, the solution to the subset sum problem finds the best approximation of z using the sum of a subset of $\{x_i\}_{i=1}^n$. From an optimization perspective, the optimal approximation error $\eta^\star$ is the solution to the following problem $\eta^\star = \min_{\boldsymbol{\delta}\in\{0,1\}^n} |\sum_{i=1}^n \delta_i x_i - z|$ where $\delta_i \in \{0,1\}$ is the indicator variable on whether x_i is selected in the sum to approximate z.

From the perspective of strong LTH, we can treat z as the weight entry in the target network that we wish to approximate, and $\{x_i\}_{i=1}^n$ as the weights in the candidate network we will prune. Here, $\delta_i = 1$ means that the i-th weight is kept, while $\delta_i = 0$ means that the i-th weight is pruned. [14] shows that, with high probability over the randomness of $\mathbf{x}_i \sim \texttt{Unif}([-1,1])$, a candidate set with size of the order $n = \Omega\left(\log \eta^{-1}\right)$ is enough to guarantee that $\eta^* \leq \eta$ for all $z \in [-1/2, 1/2]$.

As an extension to the strong LTH, our setup incorporates an ε-perturbation on the weights of the random neural network. This calls for the attention of extending the subset sum problem to a version with ε-perturbation. In particular, we consider the following joint minimization problem:

$$\eta^\star = \min_{\boldsymbol{\delta}\in\{0,1\}^n, \mathbf{y}\in[-\varepsilon,\varepsilon]^n} \left| \sum_{i=1}^n \delta_i \left(x_i + y_i\right) - z \right|. \tag{2}$$

We denote the values that lead to the optimal approximation error as $\boldsymbol{\delta}^\star$ and $\mathbf{y}^\star$, respectively. In words, the above problem aims to select values from the set $\{x_i\}_{i=1}^n$ such that, after potential entrywise perturbation by some tunable $y_i \in [-\varepsilon, \varepsilon]$, the summation of the selected and perturbed $\sum_{i=1}^n \delta_i \left(x_i + y_i\right)$ will approximate z.

A central technical difficulty in our work is to extend the result of [14] to incorporate ε-perturbation. Intuitively, as the perturbation scale ε becomes larger, each candidate is susceptible to a larger change in order to better approximate the objective z. Intuitively, this implies that we should require a smaller size for the candidate size. We formally show this in the theorem below.

Theorem 1 *Given a candidate set $\{x_i\}_{i=1}^n$ with $x_i \sim \texttt{Unif}([-1,1])$ for all $i \in [n]$. Consider the ε-perturbed subset sum problem in Equation 2. Let the number of candidates be $n = K_1 + K_2$ with*

$$K_1 = O\left(\frac{\log \eta^{-1}}{\log\left(\frac{5}{4} + \frac{\varepsilon}{2}\right)}\right) \quad ; K_2 = O\left(1 + \frac{\log \eta^{-1}}{(1+\varepsilon)}\right)$$

Then with probability at least $1 - \exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) - \exp\left(-\frac{K_1}{18}\right) - \exp\left(-\max\{\varepsilon, \eta\} K_1\right)$ we have that all $z \in [-1/2, 1/2]$ has a 2η-approximation.

Sketch of proof: The proof of Theorem (1) is provided in Appendix (1). Compared with the proof of [14], we included the ε-perturbation when constructing the recurrence of the size of the target range that can be approximated. We sketch the proof below:

1. We start by defining an indicator function $f_{k,\eta}(z)$ corresponding to the event that z has an η-approximation by the first k candidates. However, this recursively defined sequence is hard to control as it involves the $f_{k,\eta+\varepsilon}(z)$. We further study the behavior of the set where $f_{k,\eta+\varepsilon}(z) = 1$, and, by introducing the notion of ε-extensionm, we construct another sequence of indicator functions $\{\hat{f}_k\}_{k=1}^n$ that lower bound $f_{k,\eta}$ yet shows the advantage of large ε.

2. As in [14], we define p_k to be the fraction of z on the interval $[-1/2, 1/2]$ such that $\hat{f}_k = 1$. Differently, we show that the expectation of $p_{k+1} - p_k$ is lower bounded by $1/2(1-p_k)(p_k+\varepsilon)$. This demonstrates the expected growth p_{k+1} enjoys from p_k. Noticeably, this growth is larger when ε is larger. Note that this property implies a lower bound on the *expectation* of p_n.

3. We first show the lower bound on K_1 such that $p_{K_1} \geq 1/4$ with a high probability. We do this by partitioning the interval $[0, 1/4]$ into sub-intervals that represents geometric grown. We then show that the sum of the number of steps of growth that escapes these intervals can be represented as a binomial random variable, which can be bounded by applying Hoeffding's inequality. Next, starting from $p_k \geq 1/4$, we lower-bound the summation of $Z_{k+1} = \frac{p_{k+1}-p_k}{p_k(1-p_k)}$ using Azuma's inequality. To relate the summation of Z_{k+1} to the growth of p_k, we define a function $\psi(p)$ such that $\psi(p_{k+1}) - \psi(p_k) \geq Z_{k+1}$. In this way, we arrive at a lower bound on $\psi(p_{K_1+K_2}) - \psi(1/4)$. Enforcing a lower bound on $p_{K_1+K_2}$ gives a lower bound on K_2.

Remark. Notice that the lower bound on the candidate set n depends on K_1 and K_2, where K_1 scales inversely with $\log(5+\varepsilon)$ and K_2 scales inversely with $1+2\varepsilon$. This implies that n decreases monotonically as ε increases. We utilize this result to analyze the approximation error in Equation 1.

5. Strong Lottery Ticket Hypothesis with ε Perturbation

Next, we extend this idea to the approximation of a deep neural network. The idea is to approximate each layer in the target network using a two-layer ReLU MLP . Each entry in the weight matrix of the target network is approximated by a subnetwork of the MLP as in Lemma (8). Therefore, a concatenation of these MLPs gives an approximation of the target network.

Theorem 2 *Consider approximating f with g as defined above. Assume that assumption (2) holds. Also, assume that for $1 \leq \ell \leq L$,*

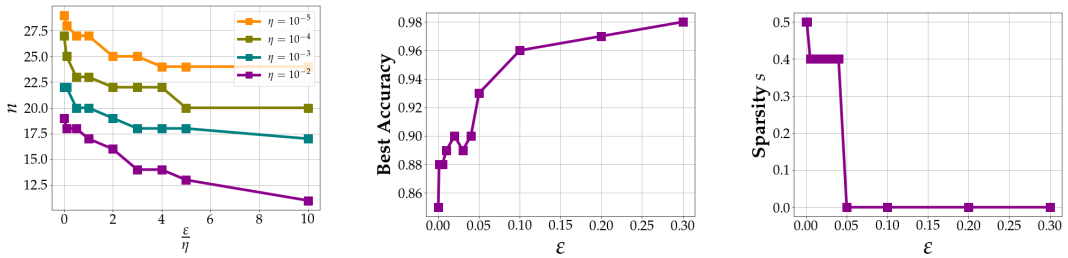
$$K_1 = C_1 d_{\ell-1} \left(\log\left(\frac{d_{\ell-1} d_\ell L}{\eta}\right) / \log\left(5/4 + \varepsilon/2\right) \right); \quad K_2 = C_2 d_{\ell-1} \left(\log\left(\frac{d_{\ell-1} d_\ell L}{\eta}\right) / 1+\varepsilon \right)$$
$$dim(\mathbf{U}^{2\ell}) = d_\ell \times (K_1 + K_2), \quad dim(\mathbf{U}^{2\ell-1}) = (K_1 + K_2) \times d_{\ell-1}.$$

Then with probability at least $1 - d_1 d_2 L \left(\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) + \exp\left(-\frac{K_1}{18}\right) + \exp\left(-\max\{\varepsilon, \eta\} K_1\right) \right)$, we have $\mathcal{L}(f, g) < \eta$, where $\mathcal{L}(f, g)$ is defined in equation (1).

6. Experiments

6.1. Approximating Neural Nets with SubsetSum and ε Perturbation

We study the effect of weight perturbation on the required overparametrization by approximating a two-layer, 500 hidden node target. Each weight was approximated using a subset sum of n randomly initialized candidates where each candidate was allowed to perturbed by at most ε. In particular, for some given ε, η, we say n satisfies the over-parametrization requirement if η^* in Equation 2 satisfies $\eta^* \leq \eta$. We randomly generate 10 sets of x_i's, and e record the minimum n such that 8 of such sets make n satisfying the over-parametrization requirement. We vary η from $1e-2$ to $1e-4$ and choose ε such that ε/η varies between 0 and 10. We are interested how such n changes as η and ε/η change. From Figure 1(*a*)subfigure, we can observe that for fixed η, n decreases as ε/η increases.



(*a*) (*b*) (*c*)

Figure 1: (a). Change of the required size of the candidate set (n) v.s. relative perturbation scale (ε/η); (b). change of accuracy on the pruned network v.s. the perturbation scale; (c). change of the optimal sparsity in pruning v.s. the perturbation scale.

More specifically, as ε increases, more changes in ε are required to make a decrease in the minimum over-parameterization requirement n, which coincides with Theorem 1.

6.2. Perturbation Using SGD

We hypothesize that *weight perturbation using SGD coincide with the desirable perturbation $\boldsymbol{\delta}^*$ in equation (2)*. We propose a two-stage algorithm to validate this hypothesis. With a given perturbation scale ε, we first train an over-parameterized neural network using projected SGD to convergence. Note that by applying the projected SGD we guarantee that each value of the neural network weight stays in an ε-neighborhood of initialization. Then, we run `edge-popup` for a range of pruning (sparsity) levels. We then consider the best accuracy amongst all pruning levels to be the optimal approximation scale. Details referred to Algorithm 1.

We train a four-layer MLP with width 500 on MNIST. We use Algorithm 1 to train the network: we use a learning rate of 0.03 for projected SGD epochs; for pruning we use `edge-popup` with a learning rate of 0.1. All weights in the network are initialized from `Unif` $([-1/2, 1/2])$, and ε ranges from 0 to 0.4. The results are shown in Table 1 and Figure 1(*c*)subfigure,1(*b*)subfigure. Note that as the perturbation scale ϵ increase, the optimal approximation error decreases. Also, as ϵ increase, the pruning level that achieves the best approximation error decreases.

Sparsity	Perturbation Scale ε									
s	0	10^{-3}	$5\cdot10^{-3}$	10^{-2}	$2\cdot10^{-2}$	$3\cdot10^{-2}$	$4\cdot10^{-2}$	$5\cdot10^{-2}$	10^{-1}	$2\cdot10^{-1}$
0	0.12	0.14	0.25	0.42	0.68	0.84	**0.90**	**0.93**	**0.96**	**0.97**
0.1	0.49	0.48	0.65	0.70	0.78	0.82	0.87	0.87	0.94	0.97
0.2	0.75	0.76	0.77	0.79	0.84	0.86	0.88	0.87	0.93	0.96
0.3	0.83	0.82	0.82	0.82	0.88	0.88	0.86	0.90	0.92	0.94
0.4	0.82	0.86	**0.88**	**0.89**	**0.90**	**0.89**	0.90	0.90	0.88	0.91
0.5	**0.85**	**0.88**	0.86	0.89	0.87	0.88	0.89	0.89	0.90	0.89
0.6	0.83	0.87	0.87	0.83	0.86	0.88	0.87	0.88	0.87	0.85
0.7	0.81	0.85	0.84	0.83	0.86	0.82	0.81	0.81	0.79	0.74
0.8	0.73	0.71	0.71	0.75	0.77	0.75	0.73	0.68	0.77	0.55

Table 1: Test accuracy for different pruning level s and perturbation scale ε. For each different perturbation scale (each column), the highest accuracy is marked bold.

REFERENCES

[1] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning?, 2020. URL https://arxiv.org/abs/2003.03033.

[2] Rebekka Burkholz. Convolutional and residual networks provably contain lottery tickets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2414–2433. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/burkholz22a.html.

[3] Rebekka Burkholz, Nilanjana Laha, Rajarshi Mukherjee, and Alkis Gotovos. On the existence of universal lottery tickets. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SYB4WrJql1n.

[4] Daiki Chijiwa, Shin'ya Yamaguchi, Yasutoshi Ida, Kenji Umakoshi, and Tomohiro INOUE. Pruning randomly initialized neural networks with iterative randomization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=QCPY2eMXYs.

[5] Arthur da Cunha, Francesco d'Amore, Frédéric Giroire, Hicham Lesfari, Emanuele Natale, and Laurent Viennot. Revisiting the random subset sum problem, 2022. URL https://arxiv.org/abs/2204.13929.

[6] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the lottery ticket hypothesis for convolutional neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Vjki79-619-.

[7] Utku Evci, Yani Ioannou, Cem Keskin, and Yann Dauphin. Gradient flow in sparse neural networks and how lottery tickets win. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6577–6586, Jun. 2022. doi: 10.1609/aaai.v36i6.20611. URL https://ojs.aaai.org/index.php/AAAI/article/view/20611.

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

[9] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis, 2019. URL https://arxiv.org/abs/1903.01611.

[10] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks, 2015. URL https://arxiv.org/abs/1506.02626.

[11] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks, 2017. URL https://arxiv.org/abs/1707.06168.

[12] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: Single-shot network pruning based on connection sensitivity, 2018. URL https://arxiv.org/abs/1810.02340.

[13] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets, 2016. URL https://arxiv.org/abs/1608.08710.

[14] George S. Lueker. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures & Algorithms*, 12(1):51–62, 1998. doi: https://doi.org/10.1002/(SICI)1098-2418(199801)12:1⟨51::AID-RSA3⟩3.0.CO;2-S. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2418%28199801%2912%3A1%3C51%3A%3AAID-RSA3%3E3.0.CO%3B2-S.

[15] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6682–6691. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/malach20a.html.

[16] Laurent Orseau, Marcus Hutter, and Omar Rivasplata. Logarithmic pruning is all you need. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2925–2934. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1e9491470749d5b0e361ce4f0b24d037-Paper.pdf.

[17] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. *Optimal Lottery Tickets via SUBSETSUM: Logarithmic over-Parameterization is Sufficient*. Curran Associates Inc., Red Hook, NY, USA, 2020. ISBN 9781713829546.

[18] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network?, 2019. URL https://arxiv.org/abs/1911.13299.

[19] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgsACVKPH.

[20] Yulong Wang, Xiaolu Zhang, Lingxi Xie, Jun Zhou, Hang Su, Bo Zhang, and Xiaolin Hu. Pruning from scratch, 2019. URL https://arxiv.org/abs/1909.12579.

[21] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks, 2016. URL https://arxiv.org/abs/1608.03665.

[22] Cameron R. Wolfe, Qihan Wang, Junhyung Lyle Kim, and Anastasios Kyrillidis. How much pre-training is enough to discover a good subnetwork?, 2021. URL https://arxiv.org/abs/2108.00259.

[23] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Towards more efficient training of deep networks, 2019. URL https://arxiv.org/abs/1909.11957.

[24] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2707–2720. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/15f99f2165aa8c86c9dface16fefd281-Paper.pdf.

[25] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1113d7a76ffceca1bb350bfe145467c6-Paper.pdf.

[26] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression, 2017. URL https://arxiv.org/abs/1710.01878.

Appendix A. Using Projected Gradient Descent to Find ε-Perturbation

Algorithm 1: PGD+StrongLTH

Data: Perturbation scale ε, neural network loss $\mathcal{L}$, initial weight $\mathbf{W}_0$, learning rate $\{\alpha_t\}_{t=0}^{T-1}$
Result: Optimal loss ℓ^*, mask $\mathbf{M}^*$ and sparsity level s

$\Delta\mathbf{W} \leftarrow 0$;
foreach $t \in \{0, \dots, T-1\}$ **do**
  $\Delta\mathbf{W} \leftarrow \texttt{sign}(\Delta\mathbf{W} - \alpha_t \nabla\mathcal{L}(\mathbf{W}_t)) \cdot \min\{\texttt{abs}(\Delta\mathbf{W} - \alpha_t\nabla\mathcal{L}(\mathbf{W}_t)), \varepsilon\}$;
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_0 + \Delta\mathbf{W}$;
end
$\ell^* \leftarrow \infty$, $\mathcal{M}^* \leftarrow$ None;
foreach *pruning level* $s \in \{0.1, 0.2, \dots, 0.9\}$ **do**
  $\ell, \mathcal{M} \leftarrow \texttt{Edge-Popup}(\mathcal{L}, \mathbf{W}_T, s)$;
  if $\ell \leq \ell^*$ **then**
    $\ell^* \leftarrow \ell, \mathcal{M}^* \leftarrow \mathcal{M}$
  end
end
return ℓ^*, $\mathbf{M}^*$, s

Appendix B. Proof of Theorem 1

The subset sum problem considers finding $\boldsymbol{s} \in \{0,1\}^n$ that minimizes $\ell(z, \boldsymbol{s}) = |z - \sum_{i=1}^n s_i x_i|$ for a given z and given x_i's. Previous work finds that, with $n = \Omega(\log 1/\eta)$, it holds with high probability that there exists $\boldsymbol{s} \in \{0,1\}^n$ such that $\ell(z, \boldsymbol{s}) \leq \eta$. Alternatively, this problem can be started as finding the smallest n such that $\eta^* \leq \eta$ with

$$\eta^* = \min_{\boldsymbol{s} \in \{0,1\}^n} \ell(z, \boldsymbol{s})$$

In our case, we would like to give the freedom of each x_i to be perturbed for a small degree ε. In particular, we extend the definition of ℓ to

$$\ell(z, \boldsymbol{s}, \mathbf{y}) = \left| z - \sum_{i=1}^n s_i (x_i + y_i) \right|$$

and seeks condition of n such that $\eta^* \leq \eta$ with

$$\eta^* = \min_{\boldsymbol{s} \in \{0,1\}^n, \mathbf{y} \in [-\varepsilon, \varepsilon]^n} \ell(z, \boldsymbol{s}, \mathbf{y}) \tag{3}$$

If this condition is met for a fixed z, we say that such z has an η approximation.

Assumption 1 *Let the candidate values* $x_i \sim \texttt{Unif}([-1, 1])$ *for all* $i \in [n]$*, and the target value* $z \in [-1/2, 1/2]$*. Let* $0 \leq \varepsilon \leq \eta \leq 1$ *be given.*

Notice that if $\varepsilon > 1$, then by Hoeffding's inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} x_i\right| \geq \frac{n\varepsilon}{2}\right) \leq \exp\left(-\frac{n\varepsilon^2}{2}\right)$$

When $|\sum_{i=1}^{n} x_i| \leq \frac{n\varepsilon}{2}$ holds, we have that $|\sum_{i=1}^{n}(x_i + y_i)|$ can be anything in $[-n\varepsilon/2, n\varepsilon/2]$ by varing y_i. Therefore, as long as $n = \frac{1}{\varepsilon^2}\log s^{-1}$, it holds with probability at least $1 - s$ that $\eta^* = 0$ for all $z \in [-1/2, 1/2]$. Thus, our focus is on the case of $\varepsilon \leq 1$. Under this assumption, we attempts to prove the following theorem

Theorem 1 *Let the number of candidates satisfy $n = K_1 + K_2$ with*

$$K_1 = O\left(\frac{\log \eta^{-1}}{\log\left(\frac{5}{4} + \frac{\varepsilon}{2}\right)}\right) \quad ; K_2 = O\left(1 + \frac{\log \eta^{-1}}{(1+\varepsilon)}\right)$$

Then with probability at least $1 - \exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) - \exp\left(-\frac{K_1}{18}\right) - \exp\left(-\max\{\varepsilon, \eta\}K_1\right)$ we have that all $z \in [-1/2, 1/2]$ has a 2η-approximation.

We define the indicator function for the existence of $\hat{\eta}$-approximation within the first k candidate.

$$f_{k,\hat{\eta}}(z) = \mathbb{I}\left\{\exists s \in \{0,1\}^k, y \in [-\varepsilon, \varepsilon]^k \text{ s.t. } \left|\sum_{i=1}^{k} s_i(x_i + y_i) - z\right| \leq \hat{\eta}\right\}$$

This indicator function has the following recurrence

$$f_{0,\eta} = \mathbb{I}\{|z| \leq \eta\}; \quad f_{k+1,\eta} = f_{k,\eta}(z) + (1 - f_{k,\eta}(z))\, f_{k,\eta+\varepsilon}(z - x_{k+1})$$

Define the following random variable (depending on $\{x_k\}_{i=1}^{k}$)

$$p_k = \int_{-1/2}^{1/2} f_{k,\eta}(z)dz$$

This random variable denotes the portion of $z \in [-1/2, 1/2]$ that can be approximated within η error.

Definition 1 *For a candidate set $\{x_i\}_{i=1}^{n}$, and some $k \in \{0\} \cup [n]$, define its (k, η)-feasible set as*

$$\mathcal{F}_{k,\eta} = \left\{z \in [-1/2, 1/2] : \exists s \in \{0,1\}^k \text{ s.t. } \left|\sum_{i=1}^{k} s_i x_i - z\right| \leq \hat{\eta}\right\}$$

By definition, $\mathcal{F}_{k,\eta}$ is the union of finitely many mutually disjoint closed intervals on $[-1/2, 1/2]$. Let μ denote the Lebesgue measure on $\mathbb{R}$. Consider the following definition of ε-extension of a set

Definition 2 *Let $I \subset [-1/2, 1/2]$ be a closed interval. A set S is called an ε-extension of I, denoted $S \in \Xi_\varepsilon(I)$ if*

1. *$S \subseteq [-1/2, 1/2] \setminus I$*
2. *for all $s \in S$, we have that $\min_{a \in I} |s - a| \leq \varepsilon$*

3. $\mu(S) = \min\{\varepsilon, 1 - \mu(I)\}$

By definition, for each $I \subset [-1/2, 1/2]$, there is at least one ε-extension of I, since we can choose

$$S = [-1/2, 1/2] \cup \begin{cases} [-1/2, \inf I) \cup (\sup I, \sup I - \inf I + \varepsilon - 1/2] & \text{if } \inf I \leq \varepsilon - \frac{1}{2} \\ [\inf I - \varepsilon, \inf I) & \text{otherwise} \end{cases}$$

Let $\mathcal{F} = \cup_{j=1}^m I_j$ be a finite union of closed intervals. A set S is called an ε-extension of $\mathcal{F}$, denoted by $S \in \Xi_\varepsilon(\mathcal{F})$ if

1. $S \subseteq \cup_{j=1}^m \cup_{\xi_j \in \Xi_\varepsilon(I_j)} \xi_j \setminus \mathcal{F}$
2. $\mu(S) = \min\{\varepsilon, 1 - \mu(\mathcal{F})\}$

By lemma 2, there is at least one ε-extension of $\mathcal{F}$.

Lemma 2 *There is at least one ε-extension for each $\mathcal{F}$ of the form $\mathcal{F} = \cup_{j=1}^m I_j$.*

Proof *Suppose there is no ε-extension of some $\mathcal{F} = \cup_{j=1}^m I_j$. Consider two cases:*

***Case 1**: $\mu(\mathcal{F}) \geq 1 - \varepsilon$. Since $\mathcal{F}$ has no ε-extension, there must be a subset A of $[-1/2, 1/2]$ with nonzero Lebesgue measure such that every element in A is at least ε away from $\mathcal{F}$. This is, however, a contradiction, since by $\mu(\mathcal{F}) \geq 1 - \varepsilon$, every point in $[-1/2, 1/2]$ must be within ε distance of $\mathcal{F}$.*

***Case 2**: $\mu(\mathcal{F}) \leq 1 - \varepsilon$. Let $S = \cup_{j=1}^m \cup_{\xi_j \in \Xi_\varepsilon(I_j)} \xi_j$. Since $\mathcal{F}$ has no ε-extension, we must have $\mu(S) < 1$. This implies that there exist $a \in [-1/2, 1/2]$ such that $a \notin S$. Thus $\inf_{a' \in \mathcal{F}} |a - a'| \geq f$. Let such a' be given, then if $a > a'$, $(a', a' + \varepsilon]$ is an ε-extension, and if $a < a'$, $[a' - \varepsilon, a')$ is an ε-extension. This is a contradiction.* ■

Let $\{S_k\}_{k=0}^n$ be given such that $S_k \in \Xi_\varepsilon(\mathcal{F}_{k,\eta})$. Moreover, let $g_k(z) = \mathbb{I}\{z \in S_k\}$. We define another recurrence of indicator function

$$\hat{f}_{k+1}(z) = \hat{f}_k(z) + \left(1 - \hat{f}_k(z)\right)\left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1})\right); \quad \hat{f}_0(z) = f_{0,\eta}(z)$$

Lemma 3 *The sequence $\left\{\hat{f}_k\right\}_{k=0}^n$ satisfies $\hat{f}_k(z) \leq f_{k,\eta}(z)$ for all $z \in [-1/2, 1/2]$.*

Proof *We show this by induction. For $k = 0$, we have $\hat{f}_0(z) = f_{0,\eta}(z)$ by definition. Assume $\hat{f}_k(z) \leq f_{k,\eta}(z)$, we would like to show $\hat{f}_{k+1}(z) \leq f_{k+1,\eta}(z)$. To do this, we first notice that, by definition of g_k*

$$f_{k,\eta+\varepsilon}(z) = f_{k,\eta}(z) + (1 - f_{k,\eta}(z))\mathbb{I}\left\{z \in \cup_{\xi_k \in \Xi_\varepsilon(\mathcal{F}_{k,\eta})}\right\} \geq f_{k,\eta}(z) + (1 - f_{k,\eta}(z))g_k(z)$$

Moreover, if $g_k(z) = 1$, we must have $f_{k,\eta}(z) = 0$. Therefore $(1 - f_{k,\eta}(z))g_k(z) = g_k(z)$. This implies that

$$1 \geq f_{k,\eta+\varepsilon}(z) \geq f_{k,\eta}(z) + g_k(z) \geq \hat{f}_k(z) + g_k(z)$$

Using this, we have

$$\begin{aligned}
f_{k+1,\eta}(z) &= f_{k,\eta}(z) + (1 - f_{k,\eta}(z)) f_{k,\eta+\varepsilon}(z - x_{k+1}) \\
&\geq f_{k,\eta}(z) + (1 - f_{k,\eta}(z)) \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) \\
&= \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) + \left(1 - \hat{f}_k(z - x_{k+1}) - g_k(z - x_{k+1}) \right) f_{k,\eta}(z) \\
&\geq \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) + \left(1 - \hat{f}_k(z - x_{k+1}) - g_k(z - x_{k+1}) \right) \hat{f}_k(z) \\
&= \hat{f}_k(z) + \left(1 - \hat{f}_k(z) \right) \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) \\
&= \hat{f}_{k+1}(z)
\end{aligned}$$

This completes the proof. ■

Based on the definition of $\left\{ \hat{f}_k \right\}_{k=0}^{n}$, we define $\{\tilde{p}_k\}_{k=0}^{n}$ as

$$\tilde{p}_k = \int_{-1/2}^{1/2} \hat{f}_k(z) dz$$

Then by definition we have $\tilde{p}_k \leq p_k$, with $\tilde{p}_0 = p_0$. Moreover, we have

$$\begin{aligned}
\tilde{p}_{k+1} &= \int_{-1/2}^{1/2} \left(\hat{f}_k(z) + \left(1 - \hat{f}_k(z) \right) \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) \right) dz \\
&\leq \tilde{p}_k + \int_{-1/2}^{1/2} \left(\hat{f}_k(z - x_{k+1}) + g_k(z - x_{k+1}) \right) dz \\
&\leq \tilde{p}_k + \int_{-1/2}^{1/2} \left(\hat{f}_k(u) + g_k(u) \right) du \\
&= 2\tilde{p}_k + \mu(S_k) \\
&\leq 2\tilde{p}_k + \varepsilon
\end{aligned}$$

Furthermore, by definition of $\tilde{p}_{k+1}$, we have $\tilde{p}_{k+1} \leq 1$. For $\tilde{p}_k$, we can compute its expectation with respect to $\{x_i\}_{i=1}^{k}$ as

$$\begin{aligned}
\mathbb{E}\left[\tilde{p}_{k+1} \right] &= \tilde{p}_k + \frac{1}{2} \int_{-1}^{1} \int_{-1/2}^{1/2} \left(1 - \hat{f}_k(z) \right) \left(\hat{f}_k(z - x) + g_k(z - x) \right) dz dx \\
&= \tilde{p}_k + \frac{1}{2} \int_{-1/2}^{1/2} \left(1 - \hat{f}_k(z) \right) dz \int_{-1}^{1} \left(\hat{f}_k(u) + g_k(u) \right) du \\
&= \tilde{p}_k + \frac{1}{2} (1 - \tilde{p}_k)(\tilde{p}_k + \mu(S_k)) \\
&= \tilde{p}_k + \frac{1}{2} (1 - \tilde{p}_k) \min\left\{ 1, \tilde{p}_k + \varepsilon \right\}
\end{aligned}$$

B.1. Growth up to ¼

We define K_1 as below

$$K_1 = \begin{cases} \min\left\{k \geq 0 : p_k > \frac{1}{4}\right\} & \text{if } \frac{1}{4} \leq 1 - \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

To upper bound K_1, we consider $\frac{1}{4} \leq 1 - \varepsilon$.

Lemma 4 *For all $0 \leq k \leq K$, it holds that*

$$\mathbb{P}\left(\tilde{p}_{k+1} \geq \frac{5}{4}\tilde{p}_k + \frac{1}{8}\varepsilon \mid \{x_i\}_{i=1}^k\right) \geq \frac{1}{6}$$

Proof *Given that $\tilde{p}_k \leq \frac{1}{4}$, we can show that*

$$\mathbb{E}\left[\tilde{p}_{k+1}\right] = \tilde{p}_k + \frac{1}{2}(1 - \tilde{p}_k)(\tilde{p}_k + \varepsilon) \geq \frac{11}{8}\tilde{p}_k + \frac{3}{8}\varepsilon$$

Moreover, recall that we have that $\tilde{p}_{k+1} \leq 2\tilde{p}_k + \varepsilon$. Thus, we can apply the reverse Markov's inequality

$$\begin{aligned} \mathbb{P}\left(p_{k+1} \geq \frac{5}{4}\tilde{p}_k + \frac{1}{8}\varepsilon \mid \{x_i\}_{i=1}^k\right) &\geq \frac{\mathbb{E}\left[\tilde{p}_{k+1}\right] - \frac{5}{4}\tilde{p}_k - \frac{1}{8}\varepsilon}{\tilde{p}_{k+1} - \frac{5}{4}\tilde{p}_k - \frac{1}{8}\varepsilon} \\ &\geq \frac{\frac{11}{8}\tilde{p}_k + \frac{3}{8}\varepsilon - \frac{5}{4}\tilde{p}_k - \frac{1}{8}\varepsilon}{2\tilde{p}_k + \varepsilon - \frac{5}{4}\tilde{p}_k - \frac{1}{8}\varepsilon} \\ &= \frac{\frac{1}{8}\tilde{p}_k + \frac{1}{4}\varepsilon}{\frac{3}{4}\tilde{p}_k + \frac{7}{8}\varepsilon} \\ &\geq \frac{1}{6} \end{aligned}$$

■

Lemma 5 *With probability at least $1 - \exp\left(-\frac{1}{18}K_1\right)$ we have that*

$$K \leq O\left(\frac{\log \eta^{-1}}{\log\left(\frac{5}{4} + \frac{\varepsilon}{2}\right)}\right)$$

Proof *By using $\tilde{p}_k \leq \frac{1}{4}$, it thus follows from lemma (4) that*

$$\mathbb{P}\left(\tilde{p}_{k+1} \geq \tilde{p}_k\left(\frac{5}{4} + \frac{\varepsilon}{2}\right)\right) \geq \mathbb{P}\left(\tilde{p}_{k+1} \geq \frac{5}{4}\tilde{p}_k + \frac{1}{8}\varepsilon \mid \{x_i\}_{i=1}^k\right) \geq \frac{1}{6}$$

Denote $\beta = \left(\frac{5}{4} + \frac{\varepsilon}{2}\right)$. As in previous work, we define the following partition of the interval $(0, 1/4)$.

$$\begin{aligned} I_1 &= (0, \eta] \\ I_i &= \left(\beta^{i-1}\eta, \beta^i\eta\right] \\ I_{i^*} &= \left(\beta^{i^*-1}\eta, \frac{1}{4}\right] \end{aligned}$$

where i^* is the smallest integer such that $\beta^{i^*}\eta \geq \frac{1}{4}$, that is

$$i^* = \left\lceil \frac{\log 1/4\eta}{\log \beta} \right\rceil$$

For $i > 1$, let $\hat{k}_i$ be given such that $\tilde{p}_{\hat{k}_i} \geq \beta^{i-1}\eta$, let $\hat{Y}_i$ be the smallest number of steps such that $\tilde{p}_{\hat{k}_i+\hat{Y}_i} > \beta^i\eta$. Then we have that $\hat{Y}_i \leq Y_i \sim \texttt{Geom}(1/6)$, since, according to lemma (4), we have

$$\mathbb{P}\left(\tilde{p}_{k+1} \geq \beta\tilde{p}_k \mid \{x_i\}_{i=1}^k\right) \geq \frac{1}{6}$$

Therefore, for all K^*, we have that

$$\mathbb{P}\left(K \geq K^*\right) \leq \mathbb{P}\left(\sum_{i=1}^{i^*} Y_i \geq K^*\right) = \mathbb{P}\left(B_{K^*} \leq i^*\right)$$

where $B_{K^*} \sim \texttt{Bin}\left(K^*, 4/7\right)$. Given that $\mathbb{E}[B_{K^*}] = \frac{1}{6}K^*$, we can apply the Hoeffding's inequality for binomial distribution

$$\mathbb{P}\left(B_{K^*} \leq \frac{1}{6}K^* - t\right) \leq \exp\left(-\frac{2t^2}{K^*}\right)$$

choose $K^* = 12i^*$ and $t = \frac{1}{6}K^*$ gives that

$$\mathbb{P}\left(K \leq K^*\right) \geq \mathbb{P}\left(B_{K^*} \geq i^*\right) \geq 1 - \exp\left(-\frac{1}{18}K^*\right)$$

■

B.2. Growth from $1/4$ to $1 - \max\{\varepsilon, \eta\}$

Recall the recurrence

$$\mathbb{E}\left[\tilde{p}_{k+1}\right] \geq \tilde{p}_k + \frac{1}{2}(1 - \tilde{p}_k)(\tilde{p}_k + \varepsilon)$$

We define

$$Z_{k+1} = \frac{\tilde{p}_{k+1} - \tilde{p}_k(z)}{(1 - \tilde{p}_k)(\tilde{p}_k + \varepsilon)}$$

Then we have $\mathbb{E}[Z_{k+1}] \geq 1/2$. Let $Y_k = -k/2 + \sum_{i=K_1+1}^{K_1+k+1} Z_i$, then Y_k is a submartingale. We bound Z_{k+1} as follows

Lemma 6

$$0 \leq Z_{k+1} \leq \frac{2}{1+\varepsilon}$$

Proof *We notice that $\tilde{p}_k \leq \tilde{p}_{k+1} \leq \min\{2\tilde{p}_k + \varepsilon, 1\}$. Consider two cases of p_k:*

***Case 1**: $\tilde{p}_k \leq \frac{1-\varepsilon}{2}$. In this case, we have $1-\tilde{p}_k \geq \frac{1+\varepsilon}{2}$*

$$Z_{k+1} \leq \frac{2\tilde{p}_k + \varepsilon - \tilde{p}_k}{(1-\tilde{p}_k)(\tilde{p}_k+\varepsilon)} = \frac{1}{1-\tilde{p}_k} \leq \frac{2}{1+\varepsilon}$$

***Case 2**: $\tilde{p}_k \geq \frac{1-\varepsilon}{2}$. In this case, we use $\tilde{p}_{k+1} \leq 1$. Moreover, we have $\tilde{p}_k + \varepsilon \geq \frac{1+\varepsilon}{2}$:*

$$Z_{k+1} \leq \frac{1-\tilde{p}_k}{(\tilde{p}_k+\varepsilon)(1-\tilde{p}_k)} = \frac{1}{\tilde{p}_k+\varepsilon} \leq \frac{2}{1+\varepsilon}$$

■

Thus,

$$|Y_{k+1} - Y_k| = \left| -\frac{1}{2} + Z_{K_1+k+2} \right| \leq \frac{|3-\varepsilon|}{2+2\varepsilon}$$

Let $n = K_1 + K_2 + 1$. Therefore, we can apply Azuma's inequality to get that

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=K_1+1}^{n} Z_i \geq \frac{K_2}{2} - t \right) &= \mathbb{P}\left(-\frac{K_2}{2} + \sum_{i=K_1+1}^{n} Z_i \geq -t \right) \\
&= \mathbb{P}\left(Y_n - Y_0 \geq -t \right) \\
&\geq 1 - \exp\left(-\frac{2\left(1+\varepsilon\right)t^2}{K_2\left(3-\varepsilon\right)^2} \right)
\end{aligned}$$

Let $t = \frac{K_2}{4}$ gives that

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq \frac{K_2}{4} \right) \geq 1 - \exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2} \right)$$

We use the following function to track the growth of p_k, but starting from $1/4$.

$$\psi(p) = \frac{1}{1+\varepsilon}\left(\log\left(p+\varepsilon\right) - \log(1-p)\right) + \frac{16}{3}p$$

Lemma 7 *For all $p_k \geq \frac{1}{4}$, we have that*

$$\psi(\tilde{p}_{k+1}) \geq \psi(\tilde{p}_k) + Z_{k+1}$$

Proof *We first notice that*

$$\psi(\tilde{p}_{k+1}) - \psi(\tilde{p}_k) = \int_{\tilde{p}_k}^{\tilde{p}_{k+1}} \psi'(p)dp \geq \min_{p\in[\tilde{p}_k,\tilde{p}_{k+1}]} \psi'(p)(\tilde{p}_{k+1} - \tilde{p}_k)$$

It suffice to show that

$$\min_{p\in[\tilde{p}_k,\tilde{p}_{k+1}]} \geq \frac{1}{(\tilde{p}_k+\varepsilon)(1-\tilde{p}_k)}$$

The first- and second-order derivative of ψ are

$$\psi'(p) = \frac{1}{1+\varepsilon}\left(\frac{1}{p+\varepsilon} + \frac{1}{1-p}\right) + \frac{16}{3}$$
$$\psi''(p) = \frac{1}{1+\varepsilon}\left(\frac{1}{(1-p)^2} - \frac{1}{(p+\varepsilon)^2}\right)$$

Therefore, ψ' attains its minimum at $p^* = \min\left\{1, \frac{1-\varepsilon}{2}\right\}$, and ψ' decreases monotonically on $(1/4, p^*]$ and increases monotonically on $[p^*, 1]$. Notice that the function $\frac{1}{(\tilde{p}_k+\varepsilon)(1-\tilde{p}_k)}$ also decreases monotonically on $(1/4, p^*]$ and increases monotonically on $[p^*, 1]$. We consider two cases of p_k:

***Case 1**: $\tilde{p}_k \in (1/4, p^*]$. In this range the function $\frac{1}{(\tilde{p}_k+\varepsilon)(1-\tilde{p}_k)}$ decreases monotonically. Thus it achieves its maximum at $\tilde{p}_k = 1/4$ with a value of $\frac{16}{3+12\varepsilon}$. However, since $0 \le p \le 1$, we have*

$$\min_{p\in[p_k, p_{k+1}]} \psi'(p) \ge \psi'(p^*) \ge \frac{16}{3}$$

Thus

$$\min_{p\in[p_k, p_{k+1}]} \psi'(p) \ge \frac{1}{(\tilde{p}_k+\varepsilon)(1-\tilde{p}_k)}$$

***Case 2**: $\tilde{p}_k \in (p^*, 1]$. Notice that ψ increases monotonically on $(p^*, 1]$. Thus*

$$\min_{p\in[p_k, p_{k+1}]} \psi'(p) = \psi'(p_k) \ge \frac{1}{1+\varepsilon}\left(\frac{1}{p_k+\varepsilon} + \frac{1}{1-p_k}\right) = \frac{1}{(p_k+\varepsilon)(1-p_k)}$$

■

Therefore, we have

$$\psi(\tilde{p}_n) \ge \psi(\tilde{p}_{K_1}) + \sum_{i=K_1+1}^{n} Z_i$$

Plugging in the value of $\psi(\tilde{p}_n)$ and $\psi(\tilde{p}_{K_1})$, and notice that ψ increases monotonically with p

$$\begin{aligned}
-\frac{\log(1-\tilde{p}_n)}{1+\varepsilon} &= \psi(\tilde{p}_n) - \frac{\log \tilde{p}_n}{1+\varepsilon} - \frac{16}{3}\tilde{p}_n \\
&\ge \psi(\tilde{p}_{K_1}) + \sum_{i=K_1+1}^{n} Z_i - \frac{16}{3} \\
&\ge \psi\left(\frac{1}{4}\right) + \sum_{i=K_1+1}^{n} Z_i - \frac{16}{3} \\
&= -\frac{1}{1+\varepsilon}\left(\log 4 + \log\frac{3}{4}\right) - 4 + \sum_{i=K_1+1}^{n} Z_i \\
&\ge \sum_{i=K_1+1}^{n} Z_i - 4
\end{aligned}$$

Since $\sum_{i=K_1+1}^{n} Z_i \geq \frac{K_2}{4}$ with probability at least $1-\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right)$, as long as $K_2 \geq \frac{2\log\eta^{-1}}{1+\varepsilon}+4 = O(\frac{\log\eta^{-1}}{1+\varepsilon}+1)$, we have that with high probability $\sum_{i=K_1+1}^{n} Z_i \geq \frac{\log\eta^{-1}}{1+\varepsilon}+4$, which implies that

$$-\log(1-\tilde{p}_n) \geq \log\max\{\varepsilon,\eta\}^{-1} = -\log(\max\{\eta,\varepsilon\})$$

which implies that $\tilde{p}_n \geq 1-\max\{\eta,\varepsilon\}$. This shows that, with high probability, for K_1, K_2 defined above, $n = K_1+K_2+1$ candidates guarantess that each point $z \in [-1/2, 1/2]$ either has an η approximation or is $\max\{\eta,\varepsilon\}$ away from an η approximation. In the case of $\eta \geq \varepsilon$, we have that each z has a 2η approximation. Otherwise, if $\varepsilon > \eta$, we need an additional set of candidates to grown from $1-\varepsilon$ to $1-\eta$.

B.3. Growth from $1-\varepsilon$ to $1-\eta$ under $\varepsilon > \eta$

Consider another set of candidates $\{\hat{x}_i\}_{i=1}^{K_3}$. By Hoeffding's inequality, we have that

$$\mathbb{P}\left(\left|\sum_{i=1}^{K_3}\hat{x}_i\right| \geq (K_3-1)\varepsilon+\eta\right) \leq \exp\left(-\frac{((K_3-1)\varepsilon+\eta)^2}{2K_3}\right)$$

This implies that with probability at least $1-\exp\left(-\frac{((K_3-1)\varepsilon+\eta)^2}{2K_3}\right)$, for each $\hat{y} \in [\eta-\varepsilon, \varepsilon-\eta]$, there exists $\mathbf{y} \in [-\varepsilon,\varepsilon]^{K_3}$ such that $\hat{y} = \sum_{i=1}^{K'}(x_i+y_i)$. This implies that with probability at least $1-\exp\left(-\frac{((K_3-1)\varepsilon+\eta)^2}{2K_3}\right)$, for all $z \in [-1/2, 1/2]$ we have that z has an η approximation. For convenience we can choose $K_3 = K_1+1$, which results in a success probability at least $1-\exp(-\max\{\varepsilon,\eta\}K_1)$ Thus, as long as $n = 2K_1+K_2+1$ with

$$K_1 = O\left(\frac{\log\eta^{-1}}{\log\left(\frac{5}{4}+\frac{\varepsilon}{2}\right)}\right) \quad ; K_2 = O\left(1+\frac{\log\eta^{-1}}{(1+\varepsilon)}\right)$$

we have that with probability at least $1-\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right)-\exp\left(-\frac{K_1}{18}\right)-\exp(-\max\{\varepsilon,\eta\}K_1)$, each point in $[-1/2, 1/2]$ either has η approximation or lies with η distance to a point with η approximation. Therefore, each point in $[-1/2, 1/2]$ has an 2η approximation.

Appendix C. Proof of Theorem 2

Lemma 8 *Let $g : \mathbb{R} \to \mathbb{R}$ be a randomly initialized network of the form $g(x) = \mathbf{v}^\top \sigma(\mathbf{u}x)$, where $\mathbf{v}, \mathbf{u} \in \mathbb{R}^{2n}$, $n = K_1 + K_2$,*

$$K_1 \geq C_1 \left(\frac{\log((\eta^{-1}))}{\log(\frac{5}{4} + \frac{\varepsilon}{2})} \right),$$
$$K_2 \geq C_2 \left(\frac{\log((\eta^{-1})))}{1+\varepsilon} \right),$$

where $u_i = 1$ for $i \leq n$, $u_i = -1$ for $i \geq n+1$, and $v_i's$ are drawn from $\mathtt{Unif}[-1, 1]$. Then, with probability at least $1 - \delta$, there exist $\mathbf{s} \in \{0, 1\}^{2n}, \mathbf{y} \in [-\varepsilon, +\varepsilon]^{2n}$ such that

$$\sup_{x:|x|\leq 1} \left| wx - (\mathbf{v} + \mathbf{y})^\top \sigma((\mathbf{u} \odot \mathbf{s})x) \right| < \eta,$$

for all $w \in [-\frac{1}{2}, \frac{1}{2}]$ with

$$\delta = \exp\left(-\tfrac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) + \exp\left(-\tfrac{K_1}{18}\right) + \exp\left(-\max\{\varepsilon, \eta\}K_1\right)$$

Proof Note that $wx = \sigma(wx) - \sigma(-wx)$ and without loss of generality we assume $w \geq 0$. The case of $w < 0$ can be handled by changing x to $-x$. Furthermore, we decompose $\mathbf{u}, \mathbf{v}, \mathbf{y}, \mathbf{s}$ by

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix}, \mathbf{s} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where $\mathbf{u}_1 = \mathbf{1}_n, \mathbf{u}_2 = -\mathbf{1}_n, \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, $\mathbf{s}_1, \mathbf{s}_2 \in \{0, 1\}^n$, and $\mathbf{y}_1, \mathbf{y}_2 \in [-\varepsilon, \varepsilon]^n$. Then we have

$$(\mathbf{v} + \mathbf{y})^\top \sigma((\mathbf{u} \odot \mathbf{s})x) = (\mathbf{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) + (\mathbf{v}_2 + \mathbf{y}_2)^\top \sigma((\mathbf{u}_2 \odot \mathbf{s}_2)x)$$

We use the first half of the RHS to approximate $\sigma(wx)$ and use the second half of the RHS to approximate $-\sigma(-wx)$.

Approximating $\sigma(wx)$. Note that since $w \geq 0$, then for $x \leq 0$, $(\mathbf{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) = \sigma(wx) = 0$. Consider $x > 0$. By definition of $\mathbf{u}_1$, we have

$$\mathbf{v}_1^\top \sigma(\mathbf{u}_1 x) = \mathbf{v}_1^\top x = \left(\sum_{i=1}^n v_{1,i} \right) x.$$

Now consider $(\sum_{i=1}^n v_{1,i})$, Theorem 1 states that with probability at least $1 - \frac{\delta}{4}$,

$$\forall w \in \left[0, \frac{1}{2}\right], \exists \mathbf{s}_1 \in \{0, 1\}^n, \mathbf{y}_1 \in [-\varepsilon, \varepsilon]^n \text{ s.t. } \left| w - \sum_{i=1}^n s_{1,i}(v_{1,i} + y_{1,i}) \right| < \frac{\eta}{2}.$$

Since

$$(\boldsymbol{v}_1 + \mathbf{y}_1)^\top (\mathbf{s}_1 \odot \mathbf{u}_1) = \sum_{i=1}^n s_{1,i}(v_{1,i} + y_{1,i}),$$

with probability at least $1-\frac{\delta}{4}$, we have

$$\forall w \in \left[0, \frac{1}{2}\right], \exists \mathbf{s}_1 \in \{0,1\}^n, \mathbf{y}_1 \in [-\varepsilon, \varepsilon]^n \text{ s.t. } \left| w - (\boldsymbol{v}_1 + \mathbf{y}_1)^\top (\mathbf{s}_1 \odot \mathbf{u}_1) \right| < \frac{\eta}{2}.$$

Since $|x| \leq 1$, with probability at least $1-\frac{\delta}{4}$, we have

$$\forall w \in \left[0, \frac{1}{2}\right], \exists \mathbf{s}_1 \in \{0,1\}^n, \mathbf{y}_1 \in [-\varepsilon, \varepsilon]^n \text{ s.t. } \left| wx - (\boldsymbol{v}_1 + \mathbf{y}_1)^\top (\mathbf{s}_1 \odot \mathbf{u}_1)x \right| < \frac{\eta}{2}.$$

Recall that $(\mathbf{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) = \sigma(wx) = 0$ for $x \leq 0$. Also, for $x > 0$, $\sigma(wx) = wx$ and $(\boldsymbol{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) = (\boldsymbol{v}_1 + \mathbf{y}_1)^\top (\mathbf{s}_1 \odot \mathbf{u}_1)x$. Therefore, with probability at least $1-\frac{\delta}{4}$, we have

$$\forall w \in \left[0, \frac{1}{2}\right], \exists \mathbf{s}_1 \in \{0,1\}^n, \mathbf{y}_1 \in [-\varepsilon, \varepsilon]^n \text{ s.t. } \left| \sigma(wx) - (\boldsymbol{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) \right| < \frac{\eta}{2}.$$

Approximating $-\sigma(-wx)$. For $x \geq 0$, $(\mathbf{v}_2 + \mathbf{y}_2)^\top \sigma((\mathbf{u}_2 \odot \mathbf{s}_2)x) = -\sigma(-wx) = 0$. Now, consider $x < 0$. By definition of $\mathbf{u}_2$, it holds that

$$\boldsymbol{v}_2^\top \sigma(\mathbf{u}_2 x) = \left(\sum_{i=1}^{n} v_{2,i} \right) x.$$

Therefore, similarly we have that with probability at least $1-\frac{\delta}{4}$,

$$\forall w \in \left[0, \frac{1}{2}\right], \exists \mathbf{s}_2 \in \{0,1\}^n, \mathbf{y}_2 \in [-\varepsilon, \varepsilon]^n : \quad \left| -\sigma(-wx) - (\boldsymbol{v}_2 + \mathbf{y}_2)^\top \sigma((\mathbf{u}_2 \odot \mathbf{s}_2)x) \right| < \frac{\eta}{2}.$$

Hence by a union bound, with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}
&\min_{\mathbf{s},\mathbf{y}} \sup_{x:|x|\leq 1} \left| wx - (\mathbf{v} + \mathbf{y})^\top \sigma((\mathbf{u} \odot \mathbf{s})x) \right| \\
=&\min_{\mathbf{s}_1,\mathbf{y}_1,\mathbf{s}_2,\mathbf{y}_2} \sup_{x:|x|\leq 1} \left| wx - \Big((\mathbf{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) + (\mathbf{v}_2 + \mathbf{y}_2)^\top \sigma((\mathbf{u}_2 \odot \mathbf{s}_2)x) \Big) \right| \\
\leq &\min_{\mathbf{s}_1,\mathbf{y}_1} \sup_{x:|x|\leq 1} \left| \sigma(wx) - (\mathbf{v}_1 + \mathbf{y}_1)^\top \sigma((\mathbf{u}_1 \odot \mathbf{s}_1)x) \right| + \\
&\qquad \min_{\mathbf{s}_2,\mathbf{y}_2} \sup_{x:|x|\leq 1} \left| -\sigma(-wx) - (\mathbf{v}_2 + \mathbf{y}_2)^\top \sigma((\mathbf{u}_2 \odot \mathbf{s}_2)x) \right| \\
<&\eta.
\end{aligned}$$

Note that for the case $w \leq 0$, the result has the same probability and the approximation error, so by a union bound, the lemma hold with probability at least $1-\delta$. ■

Lemma 9 *Let $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be a randomly initialized network of the form $g(\mathbf{x}) = \mathbf{V}\sigma(\mathbf{U}\mathbf{x})$, where $\mathbf{V} \in \mathbb{R}^{d_2 \times 2n}, \mathbf{U} \in \mathbb{R}^{2n \times d_1}, n = K_1 + K_2$,*

$$K_1 \geq C_1 d_1 \left(\frac{\log\left(\frac{d_1 d_2}{\eta}\right)}{\log\left(\frac{5}{4} + \frac{\varepsilon}{2}\right)} \right),$$

$$K_2 \geq C_2 d_1 \left(\frac{\log\left(\frac{d_1 d_2}{\eta}\right)}{1+\varepsilon} \right),$$

where weights in $\mathbf{V}$ are drawn i.i.d. from $\mathtt{Unif}[-1,1]$, $\mathbf{U} = \begin{pmatrix} \mathbf{U}^+ \\ \mathbf{U}^- \end{pmatrix}$, with $\mathbf{U}^+$ being a matrix of all 1s and $\mathbf{U}^-$ being a matrix of all -1s . Let $\hat{g}(\mathbf{x}) = (\mathbf{S} \odot (\mathbf{V} + \mathbf{Y}))\sigma((\mathbf{B} \odot \mathbf{U})\mathbf{x})$ be the pruned network for masks $\mathbf{S} \in \{0,1\}^{d_2 \times 2n}$, $\mathbf{B} \in \{0,1\}^{2n \times d_1}$ and perturbation matrix $\mathbf{Y} \in [-\varepsilon, \varepsilon]^{2n \times d_1}$. Let the target network be $f_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}\mathbf{x}$, then with probability at least $1 - d_1 d_2 \left(\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) - \exp\left(-\frac{K_1}{18}\right) - \exp\left(-\max\{\varepsilon, \eta\} K_1\right) \right)$, there exist $\mathbf{S}, \mathbf{B}, \mathbf{Y}$ such that

$$\sup_{x : \|x\|_\infty \leq 1} \|f_{\mathbf{W}}(\mathbf{x}) - \hat{g}(\mathbf{x})\| < \eta,$$

for all $\mathbf{W}$ such that $\|\mathbf{W}\|_\infty \leq \frac{1}{2}$.

Proof Since $\mathbf{U}$ can be written as $\begin{pmatrix} \mathbf{U}^+ \\ \mathbf{U}^- \end{pmatrix}$, with $\mathbf{U}^+$ being a matrix of all 1s and $\mathbf{U}^-$ being a matrix of all -1s, we choose $\hat{\mathbf{B}}$ such that $\hat{\mathbf{B}} \odot \mathbf{U}$ is of the form

$$\hat{\mathbf{B}} \odot \mathbf{U} = \begin{pmatrix} \mathbf{u}_1^+ & 0 & \dots & 0 \\ 0 & \mathbf{u}_2^+ & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{u}_{d_1}^+ \\ \mathbf{u}_1^- & 0 & \dots & 0 \\ 0 & \mathbf{u}_2^- & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{u}_{d_1}^- \end{pmatrix}$$

where $\mathbf{u}_j^+ = \mathbf{1}$ and $\mathbf{u}_j^- = -\mathbf{1}$. Moreover, we decompose $\mathbf{S} \odot (\mathbf{V} + \mathbf{Y})$ as

$$\mathbf{S} = \begin{pmatrix} \mathbf{s}_{1,1}^{+\top} & \cdots & \mathbf{s}_{1,d_1}^{+\top} & \mathbf{s}_{1,1}^{-\top} & \cdots & \mathbf{s}_{1,d_1}^{-\top} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{s}_{d_2,1}^{+\top} & \cdots & \mathbf{s}_{d_2,d_1}^{+\top} & \mathbf{s}_{d_2,1}^{-\top} & \cdots & \mathbf{s}_{d_2,d_1}^{-\top} \end{pmatrix}$$
$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_{1,1}^{+\top} & \cdots & \mathbf{v}_{1,d_1}^{+\top} & \mathbf{v}_{1,1}^{-\top} & \cdots & \mathbf{v}_{1,d_1}^{-\top} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{v}_{d_2,1}^{+\top} & \cdots & \mathbf{v}_{d_2,d_1}^{+\top} & \mathbf{v}_{d_2,1}^{-\top} & \cdots & \mathbf{v}_{d_2,d_1}^{-\top} \end{pmatrix}$$
$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_{1,1}^{+\top} & \cdots & \mathbf{y}_{1,d_1}^{+\top} & \mathbf{y}_{1,1}^{-\top} & \cdots & \mathbf{y}_{1,d_1}^{-\top} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{y}_{d_2,1}^{+\top} & \cdots & \mathbf{y}_{d_2,d_1}^{+\top} & \mathbf{y}_{d_2,1}^{-\top} & \cdots & \mathbf{y}_{d_2,d_1}^{-\top} \end{pmatrix}$$

where each $\mathbf{s}_{i,j}^{\pm}, \mathbf{v}_{i,j}^{\pm}, \mathbf{y}_{i,j}^{\pm} \in \mathbb{R}^{n/d_1}$. Then we have

$$\left[(\mathbf{S} \odot (\mathbf{V} + \mathbf{Y}))\sigma((\hat{\mathbf{B}} \odot \mathbf{U})\mathbf{x})\right]_i = \sum_{j=1}^{d_1}((\mathbf{v}_{i,j}^+ + \mathbf{y}_{i,j}^+) \odot \mathbf{s}_{i,j}^+)^\top \sigma(\mathbf{u}_j^+ x_j) + \sum_{j=1}^{d_1}((\mathbf{v}_{i,j}^- + \mathbf{y}_{i,j}^+) \odot \mathbf{s}_{i,j}^+)^\top \sigma(\mathbf{u}_j^- x_j)$$

Letting $\mathbf{v}_{ij} = \begin{pmatrix} \mathbf{v}_{ij}^+ \\ \mathbf{v}_{ij}^- \end{pmatrix}, \mathbf{s}_{ij} = \begin{pmatrix} \mathbf{s}_{ij}^+ \\ \mathbf{s}_{ij}^- \end{pmatrix}, \mathbf{y}_{ij} = \begin{pmatrix} \mathbf{y}_{ij}^+ \\ \mathbf{y}_{ij}^- \end{pmatrix}$ and $\mathbf{y}_{ij} = \begin{pmatrix} \mathbf{y}_{ij}^+ \\ \mathbf{y}_{ij}^- \end{pmatrix}$, we then have

$$\left[(\mathbf{S} \odot (\mathbf{V} + \mathbf{Y}))\sigma((\hat{\mathbf{B}} \odot \mathbf{U})\mathbf{x})\right]_i = \sum_{j=1}^{d_1}((\mathbf{v}_{i,j} + \mathbf{y}_{i,j}) \odot \mathbf{s}_{i,j})^\top \sigma(\mathbf{u}_j x_j)$$

Now define the event

$$F_{i,j,\eta} := \left\{ \sup_{w:|w|\le\frac{1}{2}} \inf_{\substack{\mathbf{s}_i\in\{0,1\}^{2n/d_1}, \\ \mathbf{y}_{i,j}\in[-\varepsilon,\varepsilon]^{2n/d_1}}} \sup_{x:|x|\le 1} \left| wx - ((\mathbf{v}_{i,j} + \mathbf{y}_{i,j}) \odot \mathbf{s}_{i,j})^\top \sigma(\mathbf{u}_j x) \right| < \eta \right\}.$$

Define $F_\eta := \bigcap_{i=1}^{d_2} \bigcap_{j=1}^{d_1} F_{i,j,\eta}$, then

$$\mathbb{P}\left(F_{\frac{\eta}{d_1 d_2}}\right) \ge 1 - d_1 d_2 \left(\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) - \exp\left(-\frac{K_1}{18}\right) - \exp\left(-\max\{\varepsilon,\eta\}K_1\right)\right).$$

On event $F_{\frac{\eta}{d_1 d_2}}$, we have

$$\begin{aligned}
&\sup_{\|\mathbf{W}\|_\infty \le \frac{1}{2}} \inf_{\mathbf{S},\mathbf{B},\mathbf{Y}} \sup_{\|\mathbf{x}\|_\infty \le 1} \|\mathbf{W}\mathbf{x} - (\mathbf{S} \odot (\mathbf{V} + \mathbf{Y}))\sigma((\mathbf{B} \odot \mathbf{U})\mathbf{x})\| \\
\le &\sup_{\|\mathbf{W}\|_\infty \le \frac{1}{2}} \inf_{\mathbf{S},\mathbf{Y}} \sup_{\|\mathbf{x}\|_\infty \le 1} \left\|\mathbf{W}\mathbf{x} - (\mathbf{S} \odot (\mathbf{V} + \mathbf{Y}))\sigma((\hat{\mathbf{B}} \odot \mathbf{U})\mathbf{x})\right\| \\
\le &\sup_{\|\mathbf{W}\|_\infty \le \frac{1}{2}} \inf_{\mathbf{S},\mathbf{Y}} \sup_{\|\mathbf{x}\|_\infty \le 1} \sum_{i=1}^{d_2} \left| \sum_{j=1}^{d_1} w_{i,j}x_j - \sum_{j=1}^{d_1} ((\mathbf{v}_{i,j} + \mathbf{y}_{i,j}) \odot \mathbf{s}_{i,j})^\top \sigma(\mathbf{u}_j x_j) \right| \\
\le &\sup_{\|\mathbf{W}\|_\infty \le \frac{1}{2}} \inf_{\mathbf{S},\mathbf{Y}} \sup_{\|\mathbf{x}\|_\infty \le 1} \sum_{i=1}^{d_2} \sum_{j=1}^{d_1} \left| w_{i,j}x_j - ((\mathbf{v}_{i,j} + \mathbf{y}_{i,j}) \odot \mathbf{s}_{i,j})^\top \sigma(\mathbf{u}_j x_j) \right| \\
< &d_1 d_2 \frac{\eta}{d_1 d_2} \\
= &\eta.
\end{aligned}$$

■

With the help of this lemma, we are ready to prove theorem 2. Recall that our goal is to approximate an L-layer, ReLU activated target multi-layer perceptron (MLP) $f(\mathbf{x})$ by pruning a $2L$-layer, ReLU activated candidate MLP $g(\mathbf{x})$. For some input vector $\mathbf{x} \in \mathbb{R}^{d_0}$, we assume $f(\mathbf{x}) = f^L(\mathbf{x})$ has a fixed set of parameters $\{\mathbf{W}^\ell\}_{\ell=1}^L$, represented by:

$$f^\ell(\mathbf{x}) = \begin{cases} \mathbf{W}^L f^{L-1}(\mathbf{x}), & \text{if } \ell = L, \\ \sigma\left(\mathbf{W}^\ell f^{\ell-1}(\mathbf{x})\right), & \text{if } \ell \in [L-1], \\ \mathbf{x}, & \text{if } \ell = 0, \end{cases}$$

where $\mathbf{W}^\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$. Similarly, let $g(\mathbf{x}) = g^{2L}(\mathbf{x})$ with parameters $\{\mathbf{U}_\ell\}_{\ell=1}^{2L}$, represented by:

$$g^\ell(\mathbf{x}) = \begin{cases} \mathbf{U}^{2L} g^{2L-1}(\mathbf{x}), & \text{if } \ell = 2L, \\ \sigma\left(\mathbf{U}^\ell g^{\ell-1}(\mathbf{x})\right), & \text{if } \ell \in [2L-1], \\ \mathbf{x}, & \text{if } \ell = 0, \end{cases}$$

where $\mathbf{U}^\ell \in \mathbb{R}^{\hat{d}_\ell \times \hat{d}_{\ell-1}}$. In particular, g is a neural network with twice the depth of f. We consider the pruning and ε-perturbation of $g(\mathbf{x})$ with a set of masks for the weights $\mathcal{S} = \{\mathbf{S}^\ell\}_{\ell=1}^{2L}$ and perturbation matrices $\mathcal{Y} = \{\mathbf{Y}^i\}_{i=1}^L$, denoted as $g_{\mathcal{S},\mathcal{Y}}(\mathbf{x}) = g_{\mathcal{S},\mathcal{Y}}^{2L}(\mathbf{x})$:

$$g_{\mathcal{S},\mathcal{Y}}^\ell(\mathbf{x}) = \begin{cases} (\mathbf{S}^{2L} \odot (\mathbf{U}^{2L} + \mathbf{Y}^{2L}))g_{\mathcal{S},\mathcal{Y}}^{2L-1}(\mathbf{x}), & \text{if } \ell = 2L, \\ \sigma\left((\mathbf{S}^\ell \odot (\mathbf{U}^\ell + \mathbf{Y}^\ell))g_{\mathcal{S},\mathcal{Y}}^{\ell-1}(\mathbf{x})\right), & \text{if } i \in [L-1], \\ \mathbf{x}, & \text{if } \ell = 0. \end{cases}$$

Let $\mathcal{F}_\mathcal{Y}$ denote the feasible set of the perturbation $\mathcal{Y}$. Also recall our assumptions for the setup

Assumption 2 *We assume the following condition for f, g and $\mathcal{F}_\mathcal{Y}$:*

1. *For all $\ell \in \{0\} \cup [L]$, the weight matrix $\mathbf{W}^\ell$ of the target neural network f satisfies $\|\mathbf{W}^\ell\| \leq 1$ and $\left\|\mathbf{W}^\ell\right\|_\infty \leq \frac{1}{2}$.*
2. *The initialization of g satisfies $\mathbf{U}_{ij}^{2\ell} \sim \texttt{Unif}[-1,1]$, and $\mathbf{U}_{ij}^{2\ell-1} = 1$ if $i \leq \hat{d}_{2(\ell-1)}/2$ and $\mathbf{U}_{ij}^{2\ell-1} = -1$ if $i > \hat{d}_{2(\ell-1)}/2$ for all $\ell \in [L]$ and $j \in [\hat{d}_{2\ell-3}]$.*
3. *The feasible set of $\mathcal{Y}$ is defined as*

$$\mathcal{F}_{\mathcal{Y}} = \left\{\mathcal{Y} : \forall \ell \in [L], \left\|\mathbf{Y}^{2\ell-1}\right\|_{\max} = 0 \text{ and } \left\|\mathbf{Y}^{2\ell}\right\|_{\max} \leq \varepsilon\right\}.$$

We focus on the approximation error defined as:

$$\min_{\mathcal{Y}\in\mathcal{F}_{\mathcal{Y}},\mathcal{S}} \sup_{\mathbf{x}:\|\mathbf{x}\|\leq 1} \|f(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}(\mathbf{x})\|. \tag{4}$$

We state the theorem here for convenience.

Theorem 10 *Consider approximating f with g as defined above. Assume that assumption (2) holds. Also, assume that for $1 \leq \ell \leq L$,*

$$K_1 = C_1 d_{\ell-1}\left(\frac{\log\left(\frac{d_{\ell-1}d_\ell L}{\eta}\right)}{\log\left(\frac{5}{4}+\frac{\varepsilon}{2}\right)}\right); \quad K_2 = C_2 d_{\ell-1}\left(\frac{\log\left(\frac{d_{\ell-1}d_\ell L}{\eta}\right)}{1+\varepsilon}\right)$$
$$\texttt{dim}(\mathbf{U}^{2\ell}) = d_\ell \times (K_1+K_2); \quad \texttt{dim}(\mathbf{U}^{2\ell-1}) = (K_1+K_2)\times d_{\ell-1}.$$

Then with probability at least $1-2d_1d_2L\left(\exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) - \exp\left(-\frac{K_1}{18}\right) - \exp\left(-\max\{\varepsilon,\eta\}K_1\right)\right)$,

$$\min_{\mathcal{S},\mathcal{Y}} \sup_{\mathbf{x}:\|\mathbf{x}\|_\infty\leq 1} \|f(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}(\mathbf{x})\| < \eta,$$

where $g_{\mathcal{S},\mathcal{Y}}$ is a pruning & ε-perturbation of g.

Proof By Lemma 9, for ℓ-th layer, with probability $1-d_1d_2\delta$ with

$$\delta = \exp\left(-\frac{K_2(1+\varepsilon)^2}{8(3-\varepsilon)^2}\right) + \exp\left(-\frac{K_1}{18}\right) + \exp\left(-\max\{\varepsilon,\eta\}K_1\right)$$

we have

$$\sup_{\mathbf{W}^\ell\in\mathcal{F}_{\mathbf{W}^\ell}} \min_{S^{2\ell},S^{2\ell-1},\mathbf{Y}^\ell} \sup_{\mathbf{x}:\|\mathbf{x}\|\leq 1} \|\mathbf{W}^\ell\mathbf{x} - (\mathbf{S}^{2\ell}\odot\mathbf{U}^{2\ell})\sigma((\mathbf{S}^{2\ell-1}\odot(\mathbf{U}^{2\ell-1}+\mathbf{Y}^\ell))\mathbf{x})\| < \frac{\eta}{2L}. \tag{5}$$

where $\mathcal{F}_{\mathbf{W}^\ell} = \{\mathbf{W}\in\mathbb{R}^{d_\ell\times d_{\ell-1}} : \|\mathbf{W}\|\leq 1, \|\mathbf{W}\|_\infty \leq \frac{1}{2}\}$. Since ReLU is 1-Lipschitz, with same probability, we have

$$\sup_{\mathbf{W}^\ell\in\mathcal{F}_{\mathbf{W}^\ell}} \min_{S^{2\ell},S^{2\ell-1},\mathbf{Y}^\ell} \sup_{\mathbf{x}:\|\mathbf{x}\|\leq 1} \|\sigma(\mathbf{W}^\ell\mathbf{x}) - \sigma((\mathbf{S}^{2\ell}\odot\mathbf{U}^{2\ell})\sigma((\mathbf{S}^{2\ell-1}\odot(\mathbf{U}^{2\ell-1}+\mathbf{Y}^\ell))\mathbf{x}))\| < \frac{\eta}{2L}. \tag{6}$$

Then with probability at least $1-2d_1d_2L\delta$, (5) and (6) hold simultaneously for every layer $1\leq\ell\leq L$. Equation (6) implies for $1\leq\ell\leq L-1$,

$$\left\|\sigma\left(\mathbf{W}^{\ell+1}g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right) - g_{\mathcal{S},\mathcal{Y}}^{2(\ell+1)}(\mathbf{x})\right\| \leq \frac{\eta}{2L}\left\|g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\|$$

Since $\left\|\mathbf{W}^{\ell}\right\| \leq 1$ for all $\ell \in [L]$, we have that

$$\left\|g_{\mathcal{S},\mathcal{Y}}^{2(\ell+1)}(\mathbf{x})\right\| \leq \frac{\eta}{2L}\left\|g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\| + \left\|\sigma\left(\mathbf{W}^{\ell+1}g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right)\right\| \leq \left(1+\frac{\eta}{2L}\right)\left\|g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\|$$

This implies that, for all $\mathbf{x}$ such that $\|\mathbf{x}\| \leq 1$,

$$\left\|g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\| \leq \left(1+\frac{\eta}{2L}\right)^{\ell-1}\|\mathbf{x}\| \leq \left(1+\frac{\eta}{2L}\right)^{\ell-1}$$

Thus, we have that for all $\mathbf{x}$ such that $\|\mathbf{x}\| \leq 1$

$$\begin{aligned}
\left\|f^{\ell+1}(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}^{2(\ell+1)}(\mathbf{x})\right\| &= \left\|\sigma\left(\mathbf{W}^{\ell+1}f^{\ell}(\mathbf{x})\right) - g_{\mathcal{S},\mathcal{Y}}^{2(\ell+1)}(\mathbf{x})\right\| \\
&\leq \left\|\sigma\left(\mathbf{W}^{\ell+1}f^{\ell}(\mathbf{x})\right) - \sigma\left(\mathbf{W}^{\ell+1}g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right)\right\| + \\
&\qquad \left\|\sigma\left(\mathbf{W}^{\ell+1}g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right) - g_{\mathcal{S},\mathcal{Y}}^{2(\ell+1)}(\mathbf{x})\right\| \\
&\leq \left\|f^{\ell}(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\| + \frac{\eta}{2L}\left\|g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\| \\
&\leq \left\|f^{\ell}(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}^{2\ell}(\mathbf{x})\right\| + \left(1+\frac{\eta}{2L}\right)^{\ell-1}\frac{\eta}{2L}
\end{aligned}$$

Solving the recurrence thus gives

$$\begin{aligned}
\|f(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}(\mathbf{x})\| &= \left\|f^{L}(\mathbf{x}) - g_{\mathcal{S},\mathcal{Y}}^{2L}(\mathbf{x})\right\| \\
&\leq \sum_{i=1}^{L-1}\left(1+\frac{\eta}{2L}\right)^{i-1}\frac{\eta}{2L} \\
&= \frac{\eta}{2L}\frac{2L}{\eta}\left(\left(1+\frac{\eta}{2L}\right)^{L} - 1\right) \\
&< e^{\eta/2} - 1 \\
&< \eta.
\end{aligned}$$

■