



HAL
open science

Polynomially Over-Parameterized Convolutional Neural Networks Contain Structured Strong Winning Lottery Tickets

Arthur Carvalho Walraven da Cunha, Francesco d'Amore, Emanuele Natale

► **To cite this version:**

Arthur Carvalho Walraven da Cunha, Francesco d'Amore, Emanuele Natale. Polynomially Over-Parameterized Convolutional Neural Networks Contain Structured Strong Winning Lottery Tickets. 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Dec 2023, New Orleans, United States. hal-04143024v2

HAL Id: hal-04143024

<https://hal.science/hal-04143024v2>

Submitted on 16 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Polynomially Over-Parameterized Convolutional Neural Networks Contain Structured Strong Winning Lottery Tickets

Arthur da Cunha^{1,2}, Francesco d’Amore^{3,4}, and Emanuele Natale¹

¹Université Côte d’Azur, Inria, CNRS, I3S, France

²Aarhus University, Denmark

³Aalto University, Finland

⁴Bocconi University, Italy

November 16, 2023

Abstract

The Strong Lottery Ticket Hypothesis (SLTH) states that randomly-initialised neural networks likely contain subnetworks that perform well without any training. Although unstructured pruning has been extensively studied in this context, its structured counterpart, which can deliver significant computational and memory efficiency gains, has been largely unexplored. One of the main reasons for this gap is the limitations of the underlying mathematical tools used in formal analyses of the SLTH. In this paper, we overcome these limitations: we leverage recent advances in the multidimensional generalisation of the Random Subset-Sum Problem and obtain a variant that admits the stochastic dependencies that arise when addressing structured pruning in the SLTH. We apply this result to prove, for a wide class of random Convolutional Neural Networks, the existence of structured subnetworks that can approximate any sufficiently smaller network.

This result provides the first sub-exponential bound around the SLTH for structured pruning, opening up new avenues for further research on the hypothesis and contributing to the understanding of the role of over-parameterization in deep learning.

1 Introduction

Much of the success of deep learning techniques relies on extreme over-parameterization [ZBH⁺17, ZBH⁺21, NBA⁺18, BGMS18, DZPS19, KMH⁺20]. While such excess of parameters has allowed neural networks to become the state of the art in many tasks, the associated computational cost limits both the progress of those techniques and their deployment in real-world applications. This limitation motivated the development of methods for reducing the number of parameters of neural networks; both in the past [Ree93] and in the present [BOFG20, HAB⁺21].

Although pruning methods have traditionally targeted reducing the size of networks for inference purposes, recent works have indicated that they can also be used to reduce parameter counts during training or even at initialisation without sacrificing model accuracy. In particular, [FC19] proposed the *Lottery Ticket Hypothesis (LTH)*, which conjectures that randomly-initialised networks contain sparse subnetworks that can be trained and reach the performance of the fully-trained original network. Empirical investigations on the LTH [ZLLY19, RWK⁺20, WZX⁺20] pointed towards an even more striking phenomenon: the existence of subnetworks that perform well without any training. This conjecture was named the *Strong Lottery Ticket Hypothesis (SLTH)* by [PRN⁺20].

While the SLTH has been proved for many different classes of neural networks (see Section 2), those works are restricted to unstructured pruning, where the subnetworks are obtained by freely removing individual parameters from the original network. However, this lack of structure can significantly reduce the gains that sparsity can bring, both in terms of memory and computational efficiency. The possibility of removing

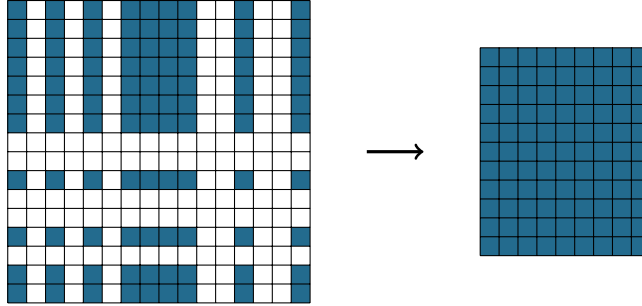


Figure 1: Illustration of neuron pruning. The left side shows the effect of pruning of neurons in the weight-matrix of a fully-connected layer. The rows in white correspond to neurons pruned in the associated layer while the columns in white represent the effect of removing neurons from the previous layers. On the right, we allude to the possibility of collapsing the pruned matrix into a smaller, dense one.

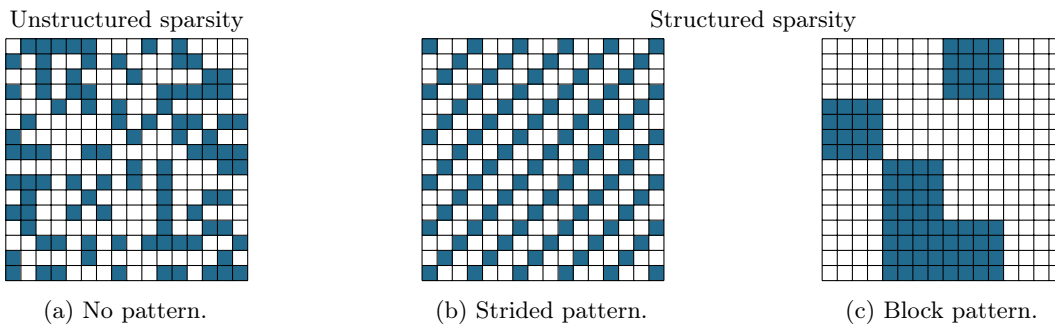


Figure 2: Examples of different pruning patterns.

parameters at arbitrary points of the network implies the need to store the indices of the remaining non-zero parameters, which can become a significant overhead with its own research challenges [PN73]. Moreover, the theoretical computational gains of unstructured sparsity can also be difficult to realise in standard hardware, which is optimised for dense operations. Most notably, the irregularity of the memory access patterns can lead to both data and instruction cache misses, significantly reducing the performance of the pruned network.

The limitations of parameter-level pruning have motivated extensive research on *structured pruning*, which constrain the sparsity patterns to reduce the complexity of parameter indexation and, more generally, to make the processing of the pruned network more efficient. A simple example of structured pruning is *neuron pruning* of fully-connected layers: deletions in the weight matrix are constrained to the level of whole rows/columns. As illustrated by Figure 1, pruning under this constraint produces a smaller network that is still dense, directly reducing the computational costs without any need for extra memory to store indices. Similarly, deleting entire filters in Convolutional Neural Networks (CNNs) [PW15] or “heads” in attention-based architectures [MLN19] also produces direct reductions in computational costs.

It is important to note that structured pruning is a restriction of unstructured pruning so, theoretically, the former is bound to perform at most as well as the latter. For example, by deleting whole neurons one can remove about 70% of the weights in dense networks without significantly affecting accuracy, while through unstructured pruning one can usually reach 95% sparsity without accuracy loss [AS16, LSZ⁺19]. In practice, however, the computational advantage of structured pruning can offset this difference: a suitably structured sparse network can be more efficient than an even sparser one that lacks structure. This trade-off between sparsity and actual efficiency has motivated the study of less coarse sparsity patterns since weaker structural constraints such as strided sparsity [AHS17] (Figure 2b) or block sparsity [Sis21] (Figure 2c) are already sufficient to deliver the bulk of the computational gains that structured can offer.

Despite its benefits, structured pruning has received little attention in the context of the SLTH. In fact, the work that first proved a version of the SLTH, [MYSSS20], has remained the only one to study this scenario,

to the best of our knowledge. Moreover, it brings a negative result: the authors prove that removing neurons from a randomly-initialised shallow neural network (a single hidden layer) is equivalent to the random features model¹ (e.g., [RR07, RR08]), which cannot efficiently approximate even a single ReLU neuron [YS19]. As most proofs around the SLTH involve pruning a shallow random network to build the desired approximations, those results show that the SLTH can be challenging to tackle when restricted to neuron pruning.

In addition, we believe that a factor hindering progress toward structured versions of the SLTH more generally is a limitation of a result underlying almost all of the theoretical works on the SLTH: a theorem by Lueker on the *Random Subset-Sum Problem (RSSP)*.

Theorem 1 ([Lue98, DCdG⁺23]). *Let X_1, \dots, X_n be independent uniform random variables over $[-1, 1]$, and let $\varepsilon \in (0, 1/3)$. There exists a universal constant $C > 0$ such that, if $n \geq C \log(1/\varepsilon)$, then, with probability at least $1 - \varepsilon$, for all $z \in [-1, 1]$ there exists $S_z \subseteq [n]$ for which*

$$\left| z - \sum_{i \in S_z} X_i \right| \leq \varepsilon.$$

In general terms, the theorem states that given a rather small number of random variables, there is a high probability that any target value within an interval of interest can be approximated by a sum of a subset of the random variables. An important remark is that even though Theorem 1 is stated in terms of uniform random variables, it is not hard to extend it to a wide class of distributions.²

While Theorem 1 closely matches the setup of the SLTH, it only concerns individual random variables and directly applying it to entire random structures, as needed when considering structured pruning, would require an exponential number of random variables. The recent works [BDHT22, BdCC⁺22] reduced this gap by proving multidimensional versions of Theorem 1. Still, the intricate manipulation of the network parameters in proofs around the SLTH imposes restrictions that are not covered by those results.

Contributions

In this work, we overcome those obstacles and prove that random networks in a wide class of CNNs are likely to contain structured subnetworks that approximate any sufficiently smaller CNN. To the best of our knowledge, our results provide the first sub-exponential bounds around the SLTH for structured pruning of deep neural networks of any kind. More precisely,

- We prove a multidimensional version Theorem 1 that is robust to some dependencies between coordinates, which is crucial for structured pruning (Theorem 5);
- We leverage this result and, by combining two types of structured sparsity (block and neuron/filter sparsity), we show that, with high probability and for a wide class of architectures, polynomially over-parameterized random networks can be pruned in a structured manner to approximate any target network (Theorem 2);
- Our results cover CNNs, which generalise fully-connected networks as well as many layer types commonly used in modern architectures, such as pooling and normalisation layers;
- Additionally, our pruning scheme focuses on filter pruning, which, like neuron pruning, allows for a direct reduction of the size and computational cost relative to the original CNN.

2 Related Work

SLTH Put roughly, research on the SLTH revolves around the following question:

Question. Given an error margin $\varepsilon > 0$ and a target neural network f_{target} , how large must an architecture f_{random} be to ensure that, with high probability on the sampling of its parameters, one can prune f_{random} to obtain a subnetwork that approximates f_{target} up to output error ε ?

¹In the random features model, one can only train the last layer of the network.

²Distributions whose probability density function f satisfies $f(x) \geq b$ for all $x \in [-a, a]$, for some constants $a, b > 0$ (see [Lue98, Corollary 3.3]).

[MYSSS20] first proved that, for dense networks with ReLU activations, it was sufficient for f_{random} to be twice as deep and polynomially wider than f_{target} . [OHR20] showed that the width overhead could be greatly reduced by sampling parameters from a hyperbolic distribution. [PRN⁺20] improved the original result for a wide class of weight distribution, requiring only a logarithmic width overhead, which the authors proved to be asymptotically optimal. [dCNV22] generalised those results with optimal bounds to CNNs with non-negative inputs, which [Bur22a] extended to general inputs and to residual architectures. [Bur22a] also reduced the depth overhead to a single extra layer and provided results that include a whole class of activation functions. [Bur22b] obtained similar improvements to dense architectures. [FB21] modified many of the previous arguments to take into consideration networks with non-zero biases. [FTGB22] further generalised previous results on CNNs to general equivariant networks. [DK21] obtained similar SLTH results for binary dense neural networks within polynomial depth and width overhead, which [SRSP22] improved to logarithmic overhead.

Structured pruning Works on structured pruning date back to the early days of the field of neural network sparsification with works such as [MS88] and [MS89]. Since then, a vast literature has been built around the topic, particularly for the pruning of CNNs. For a survey of structured pruning in general, we refer the reader to the associated sections of [HAB⁺21], and to [HX23] for a survey on structured pruning of CNNs.

RSSP [PRN⁺20] introduced the use of theoretical results on the RSSP in arguments around the SLTH, namely [Lue98, Corollary 3.3]. The work of [DCdG⁺23] provides an alternative, simpler proof of this result. [BDHT22] and [BdCC⁺22] proved multidimensional versions of the theorem. Theorem 5 diverges from those results in that it supports some dependencies between the entries of random vectors.

3 Preliminaries and contribution

Given $n \in \mathbb{N}$, we denote the set $\{1, \dots, n\}$ by $[n]$. The symbol $*$ represents the convolution operation, \odot represents the element-wise (Hadamard) product, and ϕ represents the ReLU activation function. The notation $\|\cdot\|_1$ refers to the sum of the absolute values of each entry in a tensor. Similarly, $\|\cdot\|_2$ refers to the square root of the sum of the squares of each entry in a tensor. $\|\cdot\|_{\max}$ denotes the maximum norm: the maximum among the absolute value of each entry. Sometimes we represent a tensor $X \in \mathbb{R}^{d_1 \times \dots \times d_n}$ by the notation $X = (X_{i_1, \dots, i_n})_{i_1 \in [d_1], \dots, i_n \in [d_n]}$. We denote the normal probability distribution with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$. We write $U \sim \mathcal{N}^{d_1 \times \dots \times d_n}$ to denote that U is a random tensor of size $d_1 \times \dots \times d_n$ with entries independent and identically distributed (i.i.d.), each following $\mathcal{N}(0, 1)$. We refer to such random tensors as *normal tensors*. Finally, we refer to the axis of a 4-D tensor as *rows*, *columns*, *channels*, and *kernels* (a.k.a. filters), in this order.

For the sake of simplicity, we assume CNNs to be of the form $N: [-1, 1]^{D \times D \times c_0} \rightarrow \mathbb{R}^{D \times D \times c_\ell}$ given by

$$N(X) = K^\ell * \phi(K^{\ell-1} * \dots * \phi(K^1 * X)),$$

where $K^i \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$ for $i \in [\ell]$, and the convolutions have no bias and are suitably padded with zeros. Moreover, when the kernels $K^{(i)}$ are normal tensors, we say that N is a *random CNN*.

Our main result is the following.

Theorem 2 (Structured SLTH). *Let $D, c_0, \ell \in \mathbb{N}$, and $\varepsilon \in \mathbb{R}_{>0}$. For $i \in [\ell]$, let $d_i, c_i, n_i \in \mathbb{N}$. Let \mathcal{F} be the class of functions from $[-1, 1]^{D \times D \times c_0}$ to $\mathbb{R}^{D \times D \times c_\ell}$ such that, for each $f \in \mathcal{F}$*

$$f(X) = K^{(\ell)} * \phi(K^{(\ell-1)} * \dots * \phi(K^{(1)} * X)), \quad (1)$$

where, for $i \in [\ell]$, $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$ and $\|K^{(i)}\|_1 \leq 1$.

Let also $N_0: [-1, 1]^{D \times D \times c_0} \rightarrow \mathbb{R}^{D \times D \times c_\ell}$ be a 2ℓ -layered random CNN given by

$$N_0(X) = L^{(2\ell)} * \phi(L^{(2\ell-1)} * \dots * \phi(L^{(1)} * X)), \quad (2)$$

where for $i \in [\ell]$ the kernels $L^{(2i-1)}$ and $L^{(2i)}$ are normal tensors of shape $1 \times 1 \times c_{i-1} \times 2n_i c_{i-1}$ and $d_i \times d_i \times 2n_i c_{i-1} \times c_i$, respectively.

Finally, let \mathcal{G} be the class of subnetworks that can be obtained by pruning contiguous blocks of parameters and removing entire filters from N_0 .

Then, there exists a universal constant $C > 0$, such that if, for $i \in [\ell]$,

$$n_i \geq C d_i^{13} c_i^6 \log^3 \frac{d_i^2 c_i c_{i-1} \ell}{\varepsilon},$$

then, with probability at least $1 - \varepsilon$, we have that, for all $f \in \mathcal{F}$,

$$\sup_{X \in [-1, 1]^{D \times D \times c_0}} \min_{g \in \mathcal{G}} \|f(X) - g(X)\|_{\max} \leq \varepsilon.$$

The filter removals ensured by Theorem 2 take place at layers $1, 3, \dots, 2\ell - 1$ and imply the removal of the corresponding channels in the next layer. The overall modification yields a CNN with kernels $\tilde{L}^{(1)}, \dots, \tilde{L}^{(2\ell)}$ such that, for $i \in [\ell]$, the kernels $\tilde{L}^{(2i-1)}$ and $\tilde{L}^{(2i)}$ have shape $1 \times 1 \times c_{i-1} \times 2c_{i-1}m_i$ and $d_i \times d_i \times 2c_{i-1}m_i \times c_i$, respectively, where $m_i = \sqrt{n_i / (C_1 d_i \log(1/\varepsilon))}$ for a universal constant C_1 . Moreover, our proof ensures that the kernels $\tilde{L}^{(2i-1)}$ can be required to have a specific type of block sparsity: they can be structured as if pruned by $2m_i$ -channel-blocked masks, defined as follows.

Definition 3 (n -channel-blocked mask). Given a positive integer n , a binary tensor $S \in \{0, 1\}^{d \times d \times c \times cn}$ is called n -channel-blocked if and only if

$$S_{i,j,k,l} = \begin{cases} 1 & \text{if } \lceil \frac{l}{n} \rceil = k, \\ 0 & \text{otherwise,} \end{cases}$$

for all $i, j \in [d]$, $k \in [c]$, and $l \in [cn]$.

We remark that, from a broader perspective, the central aspect of Theorem 2 is that the lower bound on the size of the random CNN depends only on the kernel sizes of the CNNs being approximated.

In subsection 4.2 we discuss the proof of Theorem 2. It requires handling subset-sum problems on multiple random variables at once (random vectors). Furthermore, the inherent parameter-sharing of CNNs creates a specific type of stochastic dependency between coordinates of the random vectors, which we capture with the following definition.

Definition 4 (NSN vector). A d -dimensional random vector Y follows a *normally-scaled normal* (NSN) distribution if, for each $i \in [d]$, $Y_i = Z \cdot Z_i$ where Z, Z_1, \dots, Z_d are i.i.d. random variables following a standard normal distribution.

A key technical contribution of ours is a Multidimensional Random Subset Sum (MRSS) result that supports NSN vectors. In subsection 4.1 we discuss the proof of the next theorem, which follows a strategy similar to that of [BDHT22, Lemmas 1, 15].

Theorem 5 (Normally-scaled MRSS). *Let $0 < \varepsilon \leq 1/4$, and let d, k , and n be positive integers such that $n \geq k^2$ and $k \geq C d^3 \log \frac{d}{\varepsilon}$ for some universal constant $C \in \mathbb{R}_{>0}$. Furthermore, let X_1, \dots, X_n be d -dimensional i.i.d. NSN random vectors. For any $\vec{z} \in \mathbb{R}^d$ with $\|\vec{z}\|_1 \leq \sqrt{k}$, there exists with constant probability a subset $S \subseteq [n]$ of size k such that $\|(\sum_{i \in S} X_i) - \vec{z}\|_{\max} \leq \varepsilon$.*

While it is possible to naively apply Theorem 1 to obtain a version of Theorem 2, doing so leads to an exponential lower bound on the required number of random vectors.

4 Analysis

In this section, after proving our MRSS result (Theorem 5), we discuss how to use it to obtain our main result on structured pruning (Theorem 2). Full proofs are deferred to Appendix B.

4.1 Multidimensional Random Subset Sum for normally-scaled normal vectors

Notation. Given a set S and a positive integer n , we denote by $\binom{S}{n}$ the family of all subsets of S containing exactly n elements of S . Given $\varepsilon \in \mathbb{R}_{>0}$, we define the interval $I_\varepsilon(z_i) = [z_i - \varepsilon, z_i + \varepsilon]$ and the multi-interval $I_\varepsilon(\vec{z}) = [\vec{z} - \varepsilon \mathbf{1}, \vec{z} + \varepsilon \mathbf{1}]$, where $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^d$. Moreover, for any event \mathcal{E} , we denote its complementary event by $\bar{\mathcal{E}}$.

In this subsection, we estimate the probability that a set of n random vectors contains a subset that sums up to a value that is ε -close to a given target. The following definition formalises this notion.

Definition 6 (Subset-sum number). Given (possibly random) vectors X_1, \dots, X_n and a vector \vec{z} , we define the ε -subset-sum number of X_1, \dots, X_n for \vec{z} as

$$T_{X_1, \dots, X_n}^k(\vec{z}) = \sum_{S \in \binom{[n]}{k}} \mathbf{1}_{\mathcal{E}_S^{(\vec{z})}},$$

where $\mathcal{E}_S^{(\vec{z})}$ denotes the event $\|(\sum_{i \in S} X_i) - \vec{z}\|_{\max} \leq \varepsilon$. We write simply $T_{n,k}$ when X_1, \dots, X_n and \vec{z} are clear from the context.

To prove Theorem 5 we use the second moment method to provide a lower bound on the probability that the subset-sum number $T_{n,k}$ is strictly positive, which implies that at least one subset of the random vectors suitably approximates \vec{z} . Hence, we seek a lower bound on $\mathbb{E}[T_{n,k}]^2 / \mathbb{E}[T_{n,k}]^2$.

Our first lemma provides a lower bound on the probability that a sum of NSN vectors is ε -close to a target vector, through which one can infer a lower bound on $\mathbb{E}[T_{n,k}]$.

Lemma 7 (Sum of NSN vectors). *Let $k \in \mathbb{N}$, $\varepsilon \in (0, \frac{1}{4})$, $\vec{z} \in \mathbb{R}^d$ such that $\|\vec{z}\|_1 \leq \sqrt{k}$ and $k \geq 16$. Furthermore, let X_1, \dots, X_k be d -dimensional i.i.d. NSN random vectors with $d \leq k$, and let $c_d = \min\{\frac{1}{d^2}, \frac{1}{16}\}$. It holds that*

$$\Pr\left(\sum_{i=1}^k X_i \in I_\varepsilon(\vec{z})\right) \geq \frac{1}{16} \left(\frac{2\varepsilon}{\sqrt{\pi(1 + 2\sqrt{c_d} + 2c_d)k}} \right)^d.$$

Overview of the proof. The main technical difficulty lies in the fact that the random vectors X_1, \dots, X_k are NSN vectors. Bounds can be easily derived for the case where the X_i are i.i.d. normal random vectors by observing that the sum of normal random variables is also normal.

For $i \in [k]$, each entry of X_i can be written as $Z_i \cdot Z_{i,j}$ where Z_i and $Z_{i,j}$ are i.i.d. normal random variables. Conditional on Z_1, \dots, Z_k , the d entries of $X = \sum_{i=1}^k X_i$ are independent and distributed as $\mathcal{N}(0, \sum_{i=1}^k Z_i^2)$. By noticing that Z_i^2 is a chi-squared random variable and employing standard concentration inequalities (Lemma 17 in Appendix A) combined with the law of total probability, we can proceed as if the entries of X were normal, up to some correction factors. \square

Bounding $\mathbb{E}[T_{n,k}^2]$ requires handling stochastic dependencies. Thus, we estimate the joint probability that two subsets of k elements of X_1, \dots, X_n sum ε -close to the same target, taking into account that the intersection of the subsets might not be empty. The next lemma provides an upper bound on this joint probability that depends only on the size of the symmetric difference between the two subsets.

Lemma 8 (Sum of NSN vectors). *Let $k, j \in \mathbb{N}_0$ with $1 \leq j \leq k$. Furthermore, let X_1, \dots, X_{k+j} be i.i.d. d -dimensional NSN random vectors with $k \geq Cd^3 \log \frac{d}{\varepsilon}$. Let $c_d = \min\{\frac{1}{d^2}, \frac{1}{16}\}$, $A = \sum_{i=1}^j X_i$, $B = \sum_{i=j+1}^k X_i$, and $C = \sum_{i=k+1}^{k+j} X_i$.³ Then, it holds that*

$$\Pr(A + B \in I_\varepsilon(\vec{z}), B + C \in I_\varepsilon(\vec{z})) \leq 3 \left(\frac{4\varepsilon^2}{\pi(1 - 2\sqrt{c_d})j} \right)^d.$$

Overview of the proof. Let $n = k + j$. We exploit once more the fact that, for all $i \in [n]$, each entry X_i can be written as $Z_i \cdot Z_{i,j}$ where Z_i and $Z_{i,j}$ are i.i.d. normal random variables. Conditional on Z_1, \dots, Z_{k+j} ,

³We adopt the convention that $\sum_{i=1}^0 X_i = 0$.

the d entries of A , B , and C are independent and distributed as $\mathcal{N}(0, \sum_{i=1}^j Z_i^2)$, $\mathcal{N}(0, \sum_{i=j+1}^k Z_i^2)$, and $\mathcal{N}(0, \sum_{i=k+1}^{k+j} Z_i^2)$, respectively. Hence, by the concentration inequalities for the sum of chi-squared random variables (Lemma 17 in Appendix A) and by the law of total probability, we can focus on the term

$$\Pr(A_i + B_i \in I_\varepsilon(z_i), B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n),$$

where A_i , B_i , and C_i indicate the i -th entries of A , B , and C , respectively.

Another concentration argument for normal random variables (Lemma 14 in Appendix A), allow us to show that

$$\begin{aligned} & \Pr(A_i + B_i \in I_\varepsilon(z_i), B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n) \\ &= \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(z_i - B_i), C_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i)] \\ &= \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i) \Pr(C_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i)] \\ &\leq \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n, B_i) \Pr(C_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n, B_i)] \\ &= \Pr(A_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n) \Pr(C_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n). \end{aligned}$$

Thus, we have reduced our argument to the estimation of probabilities of independent normal random variables being close to zero. \square

The following lemma provides an explicit expression for the variance of the ε -subset-sum number.

Lemma 9 (Second moment of $T_{n,k}$). *Let k, n be positive integers. Let S_0, S_1, \dots, S_k be subsets of $[n]$ such that $|S_0 \cap S_j| = k - j$ for $j = 0, 1, \dots, k$. Let S, S' be two random variables yielding two subsets of $[n]$ drawn independently and uniformly at random. Let X_1, \dots, X_n be d -dimensional i.i.d. NSN random vectors. For any $\varepsilon > 0$ and $\vec{z} \in \mathbb{R}^d$, the second moment of the ε -subset sum number is*

$$\mathbb{E}[T_{n,k}^2] = \binom{n}{k}^2 \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}),$$

where $\mathcal{E}_S^{(\vec{z})}$ denotes the event $\|(\sum_{i \in S} X_i) - \vec{z}\|_{\max} \leq \varepsilon$.

Proof. Let S, S' two random variables yielding two elements of $\binom{[n]}{k}$ drawn independently and uniformly at random. By the definition of $T_{n,k}$, we have that

$$\begin{aligned} \mathbb{E}[T_{n,k}^2] &= \mathbb{E}\left[\left(\sum_{S \in \binom{[n]}{k}} \mathbf{1}_{\mathcal{E}_S^{(\vec{z})}}\right) \left(\sum_{S' \in \binom{[n]}{k}} \mathbf{1}_{\mathcal{E}_{S'}^{(\vec{z})}}\right)\right] \\ &= \mathbb{E}\left[\sum_{S, S' \in \binom{[n]}{k}} \mathbf{1}_{\mathcal{E}_S^{(\vec{z})}} \mathbf{1}_{\mathcal{E}_{S'}^{(\vec{z})}}\right] \\ &= \sum_{S, S' \in \binom{[n]}{k}} \Pr(\mathcal{E}_S^{(\vec{z})} \cap \mathcal{E}_{S'}^{(\vec{z})}) \\ &= \sum_{S, S' \in \binom{[n]}{k}} \Pr(\mathcal{E}_S^{(\vec{z})} \cap \mathcal{E}_{S'}^{(\vec{z})} \mid S = S, S' = S') \Pr(S = S, S' = S') \\ &= \binom{n}{k}^2 \sum_{j=0}^k \Pr(\mathcal{E}_S^{(\vec{z})} \cap \mathcal{E}_{S'}^{(\vec{z})} \mid |S \cap S'| = k - j) \Pr(|S \cap S'| = k - j), \end{aligned}$$

as $\Pr(\mathcal{E}_S^{(\vec{z})} \cap \mathcal{E}_{S'}^{(\vec{z})})$ depends only on the size of $S \cap S'$. \square

Overview of the proof of Theorem 5

We use the second moment method (Lemma 15 in Appendix A) on the ε -subset-sum number $T_{n,k}$ of X_1, \dots, X_n . Thus, we want to lower bound the right-hand side of

$$\Pr(T > 0) \geq \frac{\mathbb{E}[T_{n,k}]^2}{\mathbb{E}[T_{n,k}^2]}.$$

Equivalently, we can provide an upper bound on the inverse, $\frac{\mathbb{E}[T_{n,k}^2]}{\mathbb{E}[T_{n,k}]^2}$. By Lemma 9,

$$\mathbb{E}[T_{n,k}^2] = \binom{n}{k}^2 \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})})$$

where S, S', S_i and $\mathcal{E}_S^{(\vec{z})}$ are defined as in the statement of the lemma. Observe also that

$$\mathbb{E}[T_{n,k}] = \sum_{S \in \binom{[n]}{k}} \mathbb{E}[\mathbf{1}_{\mathcal{E}_S^{(\vec{z})}}] = \sum_{S \in \binom{[n]}{k}} \Pr(\mathcal{E}_S^{(\vec{z})}) = \binom{n}{k} \Pr(\mathcal{E}_{S_0}^{(\vec{z})}).$$

By using the two above observations, we have

$$\begin{aligned} \frac{\mathbb{E}[T_{n,k}]^2}{\mathbb{E}[T_{n,k}^2]} &= \frac{\binom{n}{k}^2}{\mathbb{E}[T_{n,k}^2]} \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}) \\ &= \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \frac{\Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})})}{\Pr(\mathcal{E}_{S_0}^{(\vec{z})})^2}. \end{aligned}$$

Lemma 7 provides a lower bound on the term $\Pr(\mathcal{E}_{S_0}^{(\vec{z})})$ while Lemma 8 gives an upper bound on the term $\Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})})$.

In the full proof, we then show that $\Pr(|S \cap S'| \geq k/d)$ can be bounded using the Chernoff bound (Lemma 16 in Appendix A) even if we do not deal directly with Binomial random variables. This allows us to discard the indices j for which $\Pr(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})})$ is large, which leads to the result after some technical manipulations.

4.2 Proving SLTH for structured pruning

To prove Theorem 2, we first show how to obtain the same approximation result for a single-layer CNN. Then, we iteratively apply the same argument for all layers of a larger CNN and show that the approximation error stays small.

We define the *positive* and *negative* parts of a tensor.

Definition 10. Given a tensor $X \in \mathbb{R}^{d_1 \times \dots \times d_n}$, the *positive* and *negative* parts of X are respectively defined as $X_i^+ = X_i \cdot \mathbf{1}_{X_i > 0}$ and $X_i^- = -X_i \cdot \mathbf{1}_{X_i < 0}$, where $\vec{i} \in [d_1] \times \dots \times [d_n]$ points at a generic entry of X .

Approximating a single-layer CNN

We first present a preliminary lemma that shows how to prune a single-layer convolution $\phi(V * X)$ in a way that dispenses us from dealing with the ReLU ϕ .

Lemma 11. Let $D, d, c, n \in \mathbb{N}$ be positive integers, $V \in \mathbb{R}^{1 \times 1 \times c \times 2nc}$, and $X \in \mathbb{R}^{D \times D \times c}$. Let $S_1 \in \{0, 1\}^{\text{size}(V)}$ be a $2n$ -channel-blocked mask. There exists a mask $S_2 \in \{0, 1\}^{\text{size}(V)}$ which only removes filters such that, for each $(i, j, k) \in [D] \times [D] \times [2nc]$, if $\tilde{S} = S_1 \odot S_2$, then

$$\left(\phi \left((V \odot \tilde{S}) * X \right) \right)_{i,j,k} = \left((V \odot \tilde{S})^+ * X^+ + (V \odot \tilde{S})^- * X^- \right)_{i,j,k}.$$

Overview of the proof. $S_2 \in \{0, 1\}^{\text{size}(V)}$ is such that $\tilde{V} = V \odot \tilde{S} = (V \odot S_1) \odot S_2$ contains only non-negative edges going from each input channel t to the output channels $2(t-1)n+1, \dots, (2t-1)n$, and only non-positive edges going from each input channel t to the output channels $(2t-1)n+1, \dots, 2tn$. Notice that, after applying the $2n$ -channel-blocked mask S_1 , the only possible non-zero entry of the k -th filter is located at its $t = \lceil \frac{k}{2n} \rceil$ -th channel. Hence, for each $k \in [2nc]$, we keep filter k if the entry of its t -th channel is non-negative and $2t-1 = \lceil \frac{k}{n} \rceil$, or it is non-positive and $2t = \lceil \frac{k}{n} \rceil$. \square

We approximate a single convolution $K * X$ by pruning a polynomially larger neural network of the form $U * \phi(V * X)$ exploiting only a channel-blocked mask and filter removal: this is achieved using the MRSS result (Theorem 5).

Lemma 12 (Kernel pruning). *Let $D, d, c_0, c_1, n \in \mathbb{N}$ be positive integers, $\varepsilon \in (0, \frac{1}{4})$, $M \in \mathbb{R}_{>0}$, and $C \in \mathbb{R}_{>0}$ be a universal constant with*

$$n \geq Cd^{13}c_1^6 \log^3 \frac{d^2c_1c_0}{\varepsilon}.$$

*Let $U \sim \mathcal{N}^{d \times d \times 2nc_0 \times c_1}$, $V \sim \mathcal{N}^{1 \times 1 \times c_0 \times 2nc_0}$ and $S \in \{0, 1\}^{\text{size}(V)}$, with S being a $2n$ -channel-blocked mask. We define $N_0(X) = U * \phi(V * X)$ where $X \in \mathbb{R}^{D \times D \times c_0}$, and its pruned version $N_0^{(S)}(X) = U * \phi((V \odot S) * X)$. With probability $1 - \varepsilon$, for all $K \in \mathbb{R}^{d \times d \times c_0 \times c_1}$ with $\|K\|_1 \leq 1$, it is possible to remove filters from $N_0^{(S)}$ to obtain a CNN $\tilde{N}_0^{(S)}$ for which*

$$\sup_{X: \|X\|_{\max} \leq M} \left\| K * X - \tilde{N}_0^{(S)}(X) \right\|_{\max} < \varepsilon M.$$

Overview of the proof. Exploiting Lemma 11, for each $(r, s, t_1) \in [d] \times [d] \times [c_1]$, one can show that

$$\begin{aligned} \left(U * \phi \left((V \odot \tilde{S}) * X \right) \right)_{r,s,t_1} &= \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^+ \right) \cdot X_{r-i+1, s-j+1, t_0}^+ \\ &+ \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^- \right) \cdot X_{r-i+1, s-j+1, t_0}^-, \end{aligned}$$

where $\tilde{S} = S_1 \odot S_2$, with S_1 being a $2n$ -channel-blocked mask and S_2 being a mask that removes filters. Through a Chernoff bound, we show that $\tilde{V}_{1,1,t_0,:}^+$ has at least $n/3$ non-zero entries. Up to reshaping the tensor as a one-dimensional vector, we observe that $U_{:, :, k, :} \cdot \tilde{V}_{1,1,t_0,k}^+$ is an NSN vector (Lemma 18 in Appendix A). Hence, we can apply a boosted version of the MRSS result (Corollary 19 in Appendix A) and show that, with high probability, for all target filters K with $\|K\|_1 \leq 1$, we can prune all but roughly $\sqrt{n/(C_1 d \log \frac{1}{\varepsilon})}$ positive entries of $\tilde{V}_{1,1,t_0,k}^+$, with C_1 being a universal constant, such that $\sum_{k \in [nc_0]} U_{:, :, k, :} \cdot \tilde{V}_{1,1,t_0,:}^+$ approximates the channels $K_{:, :, t_0, :}$ up to error $\varepsilon/(2d^2c_0c_1)$. The same holds for $\sum_{k \in [nc_0]} U_{:, :, k, :} \cdot \tilde{V}_{1,1,t_0,:}^-$. This pruning can be achieved by applying a third mask S_3 that only removes filters. Through some non-trivial calculations and by applying the Tensor Convolution Inequality (Lemma 20 in Appendix A), one can combine the above results to get the thesis. Notice that the overall pruning can be represented by the mask $S_1 \odot (S_2 \odot S_3)$, where $S_2 \odot S_3$ is a mask that only removes filters, and $S = S_1$ is a $2n$ -channel-blocked mask. \square

Remark 13. From the proofs of Lemmas 12 and 11, we can see that the overall modification yields a pruned CNN $\hat{U} * \phi(\hat{V} * X)$ with $\hat{V} \in \mathbb{R}^{1 \times 1 \times c_0 \times 2mc_0}$ and $\hat{U} \in \mathbb{R}^{d \times d \times 2mc_0 \times c_1}$, where $m = \sqrt{n/(C_1 d \log \frac{1}{\varepsilon})}$ for a universal constant C_1 . Moreover, the kernel \hat{V} is obtained through a 3-stage pruning process: First, we apply a $2n$ -channel-blocked mask S_1 . Second, we remove filters based on entries' signs through a mask S_2 . Third, we remove filters according to the MRSS result through a mask S_3 . Masks S_2 and S_3 can be combined into a single mask $S_2 \odot S_3$ that only removes filters: overall the pruning process consists of a $2n$ -channel-blocked mask and filter removal, which justifies the statement of Theorem 2.

Overview of the proof of Theorem 2

We iteratively apply Lemma 12 to each layer while carefully controlling the approximation error via tools such as the Lipschitz property of ReLU and the Tensor Convolution Inequality (Lemma 20). More precisely, we show that (i) the approximation error does not increase too much at each layer; and (ii) all layer approximations can be combined to approximate the entire target network. Notice that Lemma 12 guarantees that, with high probability, we can approximate all possible target filters K (with $\|K\|_1 \leq 1$) from the same 2-layered network. Hence, we get the result for the supremum over all possible choices of target filters.

5 Limitations and future work

In previous works [dCNV22, Bur22a] the assumption that the kernel of every second layer has shape $1 \times 1 \times \dots$ is only an artifact of the proof since one can readily prune entries of an arbitrarily shaped tensor to enforce the desired shape. In our case, however, the concept of structured pruning can be quite broad, and such reshaping via pruning might not fit some sparsity patterns, depending on the context. The hypothesis on the shape can be a relevant limitation for such use cases. The constructions proposed by [Bur22a, Bur22b] appear as a promising direction to overcome this limitation, with the added benefit of reducing the depth overhead.

The convolution operation commonly employed in CNNs can be cumbersome at many points of our analysis. Exploring different concepts of convolution can be an interesting path for future work as it could lead to tidier proofs and more general results. For instance, employing a 3D convolution would spare a factor c in Theorem 2.

Another limitation of our results is the restriction to ReLU as the activation function. Many previous works on the SLTH exploit the fact that ReLU satisfies the identity $x = \phi(x) - \phi(-x)$. [Bur22a] leveraged that to obtain an SLTH result for CNNs with activation functions f for which $f(x) - f(-x) \approx x$ around the origin. Our analysis, on the other hand, does not rely on such property, so adapting the approach of [Bur22a] to our setting is not straightforward.

Finally, we remark that the assumption of normally distributed weights might be relaxed. [BDHT22] provided an MRSSP result for independent random variables whose distribution converges “fast enough” to a Gaussian one.⁴ We believe our arguments can serve well as baselines to generalise our results to support random weights distributed as such.

Acknowledgments

Francesco d’Amore is supported by the MUR Fare2020 Project "PARECoDi" CUP J43C22000970001.

References

- [AHS17] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. Structured pruning of deep convolutional neural networks. *ACM J. Emerg. Technol. Comput. Syst.*, 13(3):32:1–32:18, 2017. 2
- [AS16] Jose M. Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2262–2270, 2016. 2
- [BdCC⁺22] Luca Becchetti, Arthur Carvalho Walraven da Cunha, Andrea Clementi, Francesco d’ Amore, Hicham Lesfari, Emanuele Natale, and Luca Trevisan. On the Multidimensional Random Subset Sum Problem. report, Inria & Université Cote d’Azur, CNRS, I3S, Sophia Antipolis, France ; Sapienza Università di Roma, Rome, Italy ; Università Bocconi, Milan, Italy ; Università di Roma Tor Vergata, Rome, Italy, 7 2022. 3, 4

⁴The required convergence rate is higher than that ensured by the Berry-Esseen theorem.

- [BDHT22] Sander Borst, Daniel Dadush, Sophie Huiberts, and Samarth Tiwari. On the integrality gap of binary integer programs with Gaussian data. *Mathematical Programming*, 2022. [3](#), [4](#), [5](#), [10](#)
- [BGMS18] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [1](#)
- [BOFG20] Davis W. Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John V. Guttag. What is the state of neural network pruning? In Inderjit S. Dhillon, Dimitris S. Papailiopoulos, and Vivienne Sze, editors, *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020. [1](#)
- [Bur22a] Rebekka Burkholz. Convolutional and residual networks provably contain lottery tickets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2414–2433. PMLR, 2022. [4](#), [10](#)
- [Bur22b] Rebekka Burkholz. Most activation functions can win the lottery without excessive depth. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18707–18720. Curran Associates, Inc., 2022. [4](#), [10](#)
- [DCdG⁺23] Arthur Carvalho Walraven Da Cunha, Francesco d’Amore, Frédéric Giroire, Hicham Lesfari, Emanuele Natale, and Laurent Viennot. Revisiting the Random Subset Sum Problem. In Inge Li Gørtz, Martin Farach-Colton, Simon J. Puglisi, and Grzegorz Herman, editors, *31st Annual European Symposium on Algorithms (ESA 2023)*, volume 274 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 37:1–37:11, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. [3](#), [4](#)
- [dCNV22] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *International Conference on Learning Representations*, 2022. [4](#), [10](#)
- [DK21] James Diffenderfer and Bhavya Kailkhura. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [4](#)
- [Doe20] Benjamin Doerr. Probabilistic Tools for the Analysis of Randomized Optimization Heuristics. arXiv:1801.06733, 2020. [21](#)
- [DP09] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. [14](#)
- [DZPS19] Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [1](#)
- [FB21] Jonas Fischer and Rebekka Burkholz. Towards strong pruning for lottery tickets with non-zero biases. *CoRR*, abs/2110.11150, 2021. [4](#)
- [FC19] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [1](#)
- [FTGB22] Damien Ferbach, Christos Tsirigotis, Gauthier Gidel, and Avishek Joey Bose. A general framework for proving the equivariant strong lottery ticket hypothesis. *CoRR*, abs/2206.04270, 2022. [4](#)

- [HAB⁺21] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.*, 22:241:1–241:124, 2021. [1](#), [4](#)
- [HX23] Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *CoRR*, abs/2303.00566, 2023. [4](#)
- [KMH⁺20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. [1](#)
- [LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. [14](#)
- [LSZ⁺19] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [2](#)
- [Lue98] George S. Lueker. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures and Algorithms*, 12:51–62, 1998. [3](#), [4](#)
- [MLN19] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024, 2019. [2](#)
- [MS88] Michael Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 1, [NIPS Conference, Denver, Colorado, USA, 1988]*, pages 107–115. Morgan Kaufmann, 1988. [4](#)
- [MS89] Michael C Mozer and Paul Smolensky. Using relevance to reduce network size automatically. *Connection Science*, 1(1):3–16, 1989. [4](#)
- [MYSSSS20] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020. [2](#), [4](#)
- [NBA⁺18] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [1](#)
- [OHR20] Laurent Orseau, Marcus Hutter, and Omar Rivasplata. Logarithmic Pruning is All You Need. In *Advances in Neural Information Processing Systems*, volume 33, pages 2925–2934. Curran Associates, Inc., 2020. [4](#)
- [PN73] Udo W. Pooch and Al Nieder. A survey of indexing techniques for sparse matrices. *ACM Comput. Surv.*, 5(2):109–133, 1973. [2](#)
- [PRN⁺20] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris S. Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [1](#), [4](#)
- [PW15] Adam Polyak and Lior Wolf. Channel-level acceleration of deep face representations. *IEEE Access*, 3:2163–2175, 2015. [2](#)

- [Ree93] Russell Reed. Pruning algorithms—a survey. *IEEE Trans. Neural Networks*, 4(5):740–747, 1993. [1](#)
- [RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184. Curran Associates, Inc., 2007. [3](#)
- [RR08] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th annual allerton conference on communication, control, and computing*, pages 555–561. IEEE, 2008. [3](#)
- [RWK⁺20] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What’s hidden in a randomly weighted neural network? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11890–11899. Computer Vision Foundation / IEEE, 2020. [1](#)
- [Sis21] Arlene Elizabeth Siswanto. *Block sparsity and weight initialization in neural network pruning*. PhD thesis, Massachusetts Institute of Technology, 2021. [2](#)
- [SRSP22] Kartik Sreenivasan, Shashank Rajput, Jy-yong Sohn, and Dimitris S. Papailiopoulos. Finding nearly everything within random binary networks. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 3531–3541. PMLR, 2022. [4](#)
- [WZX⁺20] Yulong Wang, Xiaolu Zhang, Lingxi Xie, Jun Zhou, Hang Su, Bo Zhang, and Xiaolin Hu. Pruning from scratch. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12273–12280. AAAI Press, 2020. [1](#)
- [YS19] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6594–6604, 2019. [3](#)
- [ZBH⁺17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [1](#)
- [ZBH⁺21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. [1](#)
- [ZLLY19] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3592–3602, 2019. [1](#)

Appendix

A Technical tools

A.1 Concentration inequalities

Lemma 14 (Most-probable normal interval). *Let X follow a zero-mean normal distribution with variance ϕ^2 . For any $z, \varepsilon \in \mathbb{R}$*

$$\Pr(X \in [z - \varepsilon, z + \varepsilon]) \leq \Pr(X \in [-\varepsilon, \varepsilon]).$$

Proof. Let $\varphi(x)$ denote the probability density function of X . Then,

$$\Pr(X \in [-\varepsilon, \varepsilon]) - \Pr(X \in [z - \varepsilon, z + \varepsilon]) = \int_{-\varepsilon}^{\varepsilon} \varphi(x) dx - \int_{z-\varepsilon}^{z+\varepsilon} \varphi(x) dx.$$

If $z - \varepsilon \geq \varepsilon$ or $z + \varepsilon \leq -\varepsilon$, the thesis is trivial as $\varphi(|x|)$ decreases in x . W.l.o.g., suppose z is positive and $z - \varepsilon < \varepsilon$. Then, $-\varepsilon < z - \varepsilon < \varepsilon < z + \varepsilon$. It follows that

$$\begin{aligned} \int_{-\varepsilon}^{\varepsilon} \varphi(x) dx - \int_{z-\varepsilon}^{z+\varepsilon} \varphi(x) dx &= \int_{-\varepsilon}^{z-\varepsilon} \varphi(x) dx - \int_{\varepsilon}^{z+\varepsilon} \varphi(x) dx \\ &= \int_{-\varepsilon}^{z-\varepsilon} \varphi(x) - \varphi(x + 2\varepsilon) dx \end{aligned}$$

which is non-negative as $\varphi(x) \geq \varphi(x + 2\varepsilon)$ for $x \geq -\varepsilon$. □

Lemma 15 (Second moment method). *If Z is a non-negative random variable then*

$$\Pr(Z > 0) \geq \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

Lemma 16 (Chernoff-Hoeffding bounds [DP09]). *Let X_1, X_2, \dots, X_n be independent random variables such that $\Pr(0 \leq X_i \leq 1) = 1$ for all $i \in [n]$. Let $X = \sum_{i=1}^n X_i$ and $\mathbb{E}[X] = \mu$. Then, for any $\delta \in (0, 1)$ the following holds:*

1. *if $\mu \leq \mu_+$, then $\Pr(X \geq (1 + \delta)\mu_+) \leq \exp\left(-\frac{\delta^2 \mu_+}{3}\right)$;*
2. *if $0 \leq \mu_- \leq \mu$, then $\Pr(X \leq (1 - \delta)\mu_-) \leq \exp\left(-\frac{\delta^2 \mu_-}{2}\right)$.*

Lemma 17 (Corollary of [LM00, Lemma 1]). *Let $X \sim \chi_d^2$ be a chi-squared random variable with d degrees of freedom. For any $t > 0$, it holds that*

1. $\Pr\left(X \geq d + 2\sqrt{dt} + 2t\right) \leq \exp(-t)$;
2. $\Pr\left(X \leq d - 2\sqrt{dt}\right) \leq \exp(-t)$.

A.2 Supporting results

Lemma 18 (NSN with positive scalar). *If a d -dimensional random vector Y is such that, for each $i \in [d]$, $Y_i = \tilde{Z} \cdot \tilde{Z}_i$, where $\tilde{Z}_1, \dots, \tilde{Z}_n$ are identically distributed random variables following a standard normal distribution, \tilde{Z} is a half-normal distribution,⁵ and $\tilde{Z}, \tilde{Z}_1, \dots, \tilde{Z}_n$ are independent, then Y follows an NSN distribution.*

⁵I.e. $\tilde{Z} = |Z|$ where Z is a standard normal distribution.

Proof. By Definition 4, Y is NSN if, for each $i \in [d]$, $Y_i = Z \cdot Z_i$ where Z, Z_1, \dots, Z_n are i.i.d. random variables following a standard normal distribution. If $\tilde{Z} = |Z|$, we can rewrite $\tilde{Z}_i = \text{sign}(Z) \text{sign}(Z_i) |Z_i|$ for each $i = 1, \dots, n$, where Z, Z_1, \dots, Z_n are i.i.d. standard normal random variables, as $\text{sign}(Z) \text{sign}(Z_i)$ is independent of $\text{sign}(Z)$ and of $\text{sign}(Z) \text{sign}(Z_j)$ for $i \neq j$. Then,

$$\begin{aligned} Y_i &= \tilde{Z} \cdot \tilde{Z}_i \\ &= |Z| \cdot \text{sign}(Z) \text{sign}(Z_i) |Z_i| \\ &= \text{sign}(Z) |Z| \cdot \text{sign}(Z_i) |Z_i| \\ &= Z \cdot Z_i, \end{aligned}$$

implying the thesis. \square

Corollary 19 (of Theorem 5). *Let d, k , and n be positive integers with $n \geq C_1 d k^2 \log(\frac{1}{\varepsilon})$ and $k \geq C_2 d^3 \log \frac{d}{\varepsilon}$ for some universal constants $C_1, C_2 \in \mathbb{R}_{>0}$. Let X_1, \dots, X_n be d -dimensional i.i.d. NSN random vectors. For any $0 < \varepsilon \leq \frac{1}{4}$ it holds*

$$\Pr \left(\forall \vec{z} \in \mathbb{R}^d : \|\vec{z}\|_1 \leq 1, \exists S : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} \leq \varepsilon \right) \geq 1 - \varepsilon.$$

Proof. Observe that the set $[-1, 1]^d$ can be partitioned in $1/\varepsilon^d$ many infinity norm balls of radius ε . Let $s = \lceil C_1 d \log(\frac{1}{\varepsilon}) \rceil$ and let us partition the n vectors X_1, \dots, X_n in s disjoint sets G_1, \dots, G_s of at least k^2 vectors each. By Theorem 5, there is a constant $c \in (0, 1)$ such that for each group G_i ($i \in [s]$)

$$\Pr \left(\exists S \subset G_i : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} \leq \varepsilon \right) \geq c. \quad (3)$$

It follows that

$$\begin{aligned} &\Pr \left(\forall \vec{z} \in \mathbb{R}^d : \|\vec{z}\|_1 \leq 1, \exists S : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} \leq \varepsilon \right) \\ &\geq \Pr \left(\forall \vec{z} \in \mathbb{R}^d : \|\vec{z}\|_1 \leq 1, \exists i \in [s], \exists S \subset G_i : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} \leq \varepsilon \right) \\ &= 1 - \Pr \left(\exists \vec{z} \in \mathbb{R}^d : \|\vec{z}\|_1 \leq 1, \forall i \in [s], \forall S \subset G_i : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} > \varepsilon \right) \\ &\geq 1 - \frac{1}{\varepsilon^d} \Pr \left(\forall i \in [s], \forall S \subset G_i : |S| = k, \left\| \left(\sum_{i \in S} X_i \right) - \vec{z} \right\|_{\max} > \varepsilon \right) \\ &\geq 1 - \frac{1}{\varepsilon^d} (1 - c)^{\lceil C_1 d \log(\frac{1}{\varepsilon}) \rceil}, \end{aligned}$$

where the latter inequality comes from Eq. 3, the independence of the variables across different G_i , and the union. By choosing C_1 large enough,

$$1 - \varepsilon^{-d} (1 - c)^{\lceil C_1 d \log(\frac{1}{\varepsilon}) \rceil} = 1 - 2^{d \log \frac{1}{\varepsilon} - \lceil C_1 d \log(\frac{1}{\varepsilon}) \rceil \log \frac{1}{1-c}} \geq 1 - \varepsilon.$$

\square

Lemma 20 (Tensor Convolution Inequality). *Given real tensors K and X of respective sizes $d \times d' \times c_0 \times c_1$ and $D \times D' \times c_0$, it holds*

$$\|K * X\|_{\max} \leq \|K\|_1 \cdot \|X\|_{\max}.$$

Proof. We have

$$\begin{aligned}
& \|K * X\|_{\max} \\
& \leq \max_{i,j \in [D], \ell \in [c_1]} \sum_{i',j' \in [d], k \in [c]} |K_{i',j',k,\ell} X_{i-i'+1, j-j'+1, k}| \\
& \leq \max_{i,j \in [D], \ell \in [c_1]} \left(\sum_{i',j' \in [d], k \in [c]} |K_{i',j',k,\ell}| \right) \|X\|_{\max} \\
& \leq \max_{i,j \in [D], \ell \in [c_1]} \|K\|_1 \cdot \|X\|_{\max} \\
& = \|K\|_1 \cdot \|X\|_{\max}.
\end{aligned}$$

□

B Omitted proofs

B.1 Multidimensional Random Subset Sum for normally-scaled normal vectors

Proof of Lemma 7. By Definition 4, the j -th entry of each vector X_i is $(X_i)_j = Z_i \cdot Z_{i,j}$ where each Z_i and $Z_{i,j}$ are i.i.d. random variables following a standard normal distribution. Let $\mathcal{E}^{(\dagger)}$ be the event that $k(1 - 2\sqrt{c_d}) \leq \sum_{i=1}^k Z_i^2 \leq k(1 + 2\sqrt{c_d} + 2c_d)$, and denote $X = \sum_{i=1}^k X_i$. By the law of total probability, it holds that

$$\Pr(X \in I_\varepsilon(\vec{z})) = \mathbb{E}_{Z_1, \dots, Z_n} [\Pr(X \in I_\varepsilon(\vec{z}) \mid Z_1, \dots, Z_k)].$$

As, conditional on Z_1, \dots, Z_k , the d entries of X are independent, it follows that

$$\begin{aligned}
& \mathbb{E}_{Z_1, \dots, Z_n} [\Pr(X \in I_\varepsilon(\vec{z}) \mid Z_1, \dots, Z_k)] \\
& = \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{j=1}^d \Pr((X)_j \in I_\varepsilon(z_j) \mid Z_1, \dots, Z_k) \right] \\
& \geq \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{j=1}^d \Pr((X)_j \in I_\varepsilon(z_j) \mid Z_1, \dots, Z_k, \mathcal{E}^{(\dagger)}) \right] \Pr(\mathcal{E}^{(\dagger)}), \tag{4}
\end{aligned}$$

where the inequality in Eq. 4 holds by applying again the law of total probability.

Conditional on Z_1, \dots, Z_k , we have that $(X)_j \sim \mathcal{N}(0, \sum_{i=1}^k Z_i^2)$. Hence,

$$\begin{aligned}
& \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{j=1}^d \Pr((X)_j \in I_\varepsilon(z_j) \mid Z_1, \dots, Z_k, \mathcal{E}^{(\dagger)}) \right] \Pr(\mathcal{E}^{(\dagger)}) \\
& \geq \mathbb{E}_{Z_1, \dots, Z_k} \left[\prod_{j=1}^d \left(\frac{2\varepsilon}{\sqrt{\pi \left(\sum_{i=1}^k Z_i^2 \right)}} \exp \left(-\frac{(|z_j| + \varepsilon)^2}{2 \sum_{i=1}^k Z_i^2} \right) \right) \middle| Z_1, \dots, Z_k, \mathcal{E}^{(\dagger)} \right] \\
& \cdot \Pr(\mathcal{E}^{(\dagger)})
\end{aligned}$$

Notice that the term $\sum_i Z_i^2$ is a sum of chi-square random variables, for which there are known concentration bounds (Lemma 17). By definition of $\mathcal{E}^{(\dagger)}$ and by applying Lemma 17 to estimate the term $\Pr(\mathcal{E}^{(\dagger)})$, we get that

$$\begin{aligned}
& \mathbb{E}_{Z_1, \dots, Z_k} \left[\prod_{j=1}^d \left(\frac{2\varepsilon}{\sqrt{\pi \left(\sum_{i=1}^k Z_i^2 \right)}} \exp \left(-\frac{(|z_i| + \varepsilon)^2}{2 \sum_{i=1}^k Z_i^2} \right) \right) \middle| Z_1, \dots, Z_k, \mathcal{E}(\dagger) \right] \\
& \cdot \Pr \left(\mathcal{E}(\dagger) \right) \\
& \geq \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{\sum_i |z_i|^2 + 2\varepsilon \sum_i |z_i| + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) \Pr \left(\mathcal{E}(\dagger) \right) \\
& = \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{\|\vec{z}\|_2^2 + 2\varepsilon \|\vec{z}\|_1 + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) \Pr \left(\mathcal{E}(\dagger) \right) \\
& \geq \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{\|\vec{z}\|_2^2 + 2\varepsilon \|\vec{z}\|_1 + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) (1 - 2e^{-c_d k}).
\end{aligned}$$

As $c_d k \geq 1$ by hypotheses, $1 - 2e^{-c_d k} \geq 1/4$. Then,

$$\begin{aligned}
& \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{\|\vec{z}\|_2^2 + 2\varepsilon \|\vec{z}\|_1 + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) (1 - 2e^{-c_d k}) \\
& \geq \frac{1}{4} \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{\|\vec{z}\|_2^2 + 2\varepsilon \|\vec{z}\|_1 + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) \\
& \geq \frac{1}{4} \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{k + 2\varepsilon\sqrt{k} + d\varepsilon^2}{2(1 - 2\sqrt{c_d}) k} \right) \tag{5}
\end{aligned}$$

$$= \frac{1}{4} \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{1 + \frac{2\varepsilon}{\sqrt{k}} + \frac{d\varepsilon^2}{k}}{2(1 - 2\sqrt{c_d})} \right)$$

$$\geq \frac{1}{4} \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d \exp \left(-\frac{1 + \frac{1}{8} + \frac{1}{16}}{2(1 - 2\sqrt{c_d})} \right) \tag{6}$$

$$\geq \frac{1}{16} \left(\frac{2\varepsilon}{\sqrt{\pi (1 + 2\sqrt{c_d} + 2c_d) k}} \right)^d, \tag{7}$$

where we have used that $\|\vec{z}\|_2 \leq \|\vec{z}\|_1 \leq \sqrt{k}$ in Ineq. 5, that $k \geq 16$, $k \geq d$, that $\varepsilon < 1/4$ in Ineq. 6, and that

$$\exp \left(-\frac{1 + \frac{1}{8} + \frac{1}{16}}{2(1 - 2\sqrt{c_d})} \right) \geq \exp \left(-\frac{1 + \frac{1}{8} + \frac{1}{16}}{2(1 - 2\sqrt{\frac{1}{16}})} \right) \geq \frac{1}{16}$$

in Ineq. 7. □

Proof of Lemma 8. Let $n = k + j$. Since the X_i s are NSN random vectors, for each $i \in [n]$ and $j \in [d]$ we can write the j -th entry of X_i as $(X_i)_j = Z_i \cdot Z_{i,j}$ where the variables in $\{Z_i\}_{i \in [n]}$ and in $\{Z_{i,j}\}_{i \in [n], j \in [d]}$ are

i.i.d. random variables following a standard normal distribution. By the law of total probability, we have

$$\begin{aligned}
& \Pr(A + B \in I_\varepsilon(\vec{z}), B + C \in I_\varepsilon(\vec{z})) \\
&= \mathbb{E}_{Z_1, \dots, Z_n} [\Pr(A + B \in I_\varepsilon(\vec{z}), B + C \in I_\varepsilon(\vec{z}) \mid Z_1, \dots, Z_n)] \\
&= \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{i=1}^d \Pr(A_i + B_i \in I_\varepsilon(z_i), B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n) \right], \tag{8}
\end{aligned}$$

where the latter equality holds by independence.

Then,

$$\begin{aligned}
& \Pr(A_i + B_i \in I_\varepsilon(z_i), B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n) \\
&= \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(z_i - B_i), C_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i)] \\
&= \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i) \Pr(C_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i)],
\end{aligned}$$

where the latter inequality holds by independence of A_i and C_i . By Lemma 14, it holds that

$$\begin{aligned}
& \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i) \Pr(C_i \in I_\varepsilon(z_i - B_i) \mid Z_1, \dots, Z_n, B_i)] \\
&\leq \mathbb{E}_{B_i} [\Pr(A_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n, B_i) \Pr(C_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n, B_i)] \\
&= \Pr(A_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n) \Pr(C_i \in I_\varepsilon(0) \mid Z_1, \dots, Z_n) \\
&\leq \frac{2\varepsilon}{\sqrt{\pi \left(\sum_{r=1}^j Z_r^2 \right)}} \cdot \frac{2\varepsilon}{\sqrt{\pi \left(\sum_{r=k+1}^{k+j} Z_r^2 \right)}}, \tag{9}
\end{aligned}$$

where the latter inequality comes from the fact that, conditional on Z_1, \dots, Z_n , we have that $A_i \sim \mathcal{N}(0, \sum_{r=1}^j Z_r^2)$, $B_i \sim \mathcal{N}(0, \sum_{r=j+1}^k Z_r^2)$, and $C_i \sim \mathcal{N}(0, \sum_{r=k+1}^{k+j} Z_r^2)$ for each $i \in [d]$.

We now proceed similarly to the proof of Lemma 7. We denote the event that $(1 - 2\sqrt{cd})j \leq \sum_{i=1}^j Z_i^2, \sum_{i=k+1}^{k+j} Z_i^2$ by $\mathcal{E}^{(\downarrow)}$. Then, by Eq. 8 and the law of total probability, we have that

$$\begin{aligned}
& \Pr(A + B \in I_\varepsilon(\vec{z}), B + C \in I_\varepsilon(\vec{z})) \\
&= \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{i=1}^d \Pr(A_i + B_i \in I_\varepsilon(z_i), B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n) \right] \\
&\leq \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{i=1}^d \Pr(A_i + B_i, B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n, \mathcal{E}^{(\downarrow)}) \right] + \Pr(\overline{\mathcal{E}^{(\downarrow)}}).
\end{aligned}$$

Eq. 9 implies that

$$\begin{aligned}
& \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{i=1}^d \Pr(A_i + B_i, B_i + C_i \in I_\varepsilon(z_i) \mid Z_1, \dots, Z_n, \mathcal{E}^{(\downarrow)}) \right] + \Pr(\overline{\mathcal{E}^{(\downarrow)}}) \\
&\leq \mathbb{E}_{Z_1, \dots, Z_n} \left[\prod_{i=1}^d \frac{2\varepsilon}{\sqrt{\pi \left(\sum_{r=1}^j Z_r^2 \right)}} \cdot \frac{2\varepsilon}{\sqrt{\pi \left(\sum_{r=k+1}^{k+j} Z_r^2 \right)}} \Bigg| \mathcal{E}^{(\downarrow)} \right] + \Pr(\overline{\mathcal{E}^{(\downarrow)}}) \\
&= \mathbb{E}_{Z_1, \dots, Z_n} \left[\left(\frac{4\varepsilon^2}{\pi \sqrt{\left(\sum_{r=1}^j Z_r^2 \right) \left(\sum_{r=k+1}^{k+j} Z_r^2 \right)}} \right)^d \Bigg| \mathcal{E}^{(\downarrow)} \right] + \Pr(\overline{\mathcal{E}^{(\downarrow)}}).
\end{aligned}$$

By independence of $\sum_{r=1}^j Z_r^2$ and $\sum_{r=k+1}^{k+j} Z_r^2$ and by Lemma 17, we obtain that

$$\begin{aligned}
& \mathbb{E}_{Z_1, \dots, Z_n} \left[\left(\frac{4\varepsilon^2}{\pi \sqrt{\left(\sum_{i=1}^j Z_i^2 \right) \left(\sum_{i=k+1}^{k+j} Z_i^2 \right)}} \right)^d \middle| \mathcal{E}^{(\downarrow)} \right] + \Pr \left(\overline{\mathcal{E}^{(\downarrow)}} \right) \\
& \leq \left(\frac{4\varepsilon^2}{\pi j (1 - 2\sqrt{c_d})} \right)^d + \Pr \left(\overline{\mathcal{E}^{(\downarrow)}} \right) \\
& \leq \left(\frac{4\varepsilon^2}{\pi j (1 - 2\sqrt{c_d})} \right)^d + 2 \exp(-c_d j) \\
& = \exp \left(-d \log \frac{\pi j (1 - 2\sqrt{c_d})}{4\varepsilon^2} \right) + 2 \exp(-c_d k) \\
& \leq 3 \left(\frac{4\varepsilon^2}{\pi (1 - 2\sqrt{c_d}) k} \right)^d.
\end{aligned}$$

Finally, for a large enough constant C , the hypothesis on k implies that $k \geq 2 \frac{d}{c_d} \log \frac{\pi k (1 - 2\sqrt{c_d})}{4\varepsilon^2}$. Hence,

$$\exp \left(-d \log \frac{\pi j (1 - 2\sqrt{c_d})}{4\varepsilon^2} \right) + 2 \exp(-c_d k) \leq 3 \left(\frac{4\varepsilon^2}{\pi (1 - 2\sqrt{c_d}) k} \right)^d.$$

□

Proof of Theorem 5. We use the second moment method (Lemma 15) on the ε -subset-sum number $T_{n,k}$ of X_1, \dots, X_n . Thus, we aim to provide a lower bound on the right-hand side of

$$\Pr(T > 0) \geq \frac{\mathbb{E}[T_{n,k}]^2}{\mathbb{E}[T_{n,k}^2]}. \tag{10}$$

Equivalently, we can provide an upper bound on the inverse $\frac{\mathbb{E}[T_{n,k}^2]}{\mathbb{E}[T_{n,k}]^2}$. By Lemma 9

$$\mathbb{E}[T_{n,k}^2] = \binom{n}{k}^2 \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})} \right) \tag{11}$$

where S, S', S_i and $\mathcal{E}_S^{(\bar{z})}$ are defined as in the statement of the lemma. Observe also that

$$\mathbb{E}[T_{n,k}] = \sum_{S \in \binom{[n]}{k}} \mathbb{E} \left[\mathbf{1}_{\mathcal{E}_S^{(\bar{z})}} \right] = \sum_{S \in \binom{[n]}{k}} \Pr \left(\mathcal{E}_S^{(\bar{z})} \right) = \binom{n}{k} \Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \right). \tag{12}$$

By Eq.s 11 and 12, we have

$$\begin{aligned}
\frac{\mathbb{E}[T_{n,k}^2]}{\mathbb{E}[T_{n,k}]^2} &= \frac{\binom{n}{k}^2}{\mathbb{E}[T_{n,k}]^2} \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})} \right) \\
&= \sum_{j=0}^k \Pr(|S \cap S'| = k - j) \frac{\Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})} \right)}{\Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \right)^2}.
\end{aligned} \tag{13}$$

As for the denominator of the second term in Eq. 13, by Lemma 7 we have

$$\Pr \left(\mathcal{E}_{S_0}^{(\bar{z})} \right)^2 \geq \frac{1}{256} \left(\frac{4\varepsilon^2}{\pi (1 + 2\sqrt{c_d} + 2c_d) k} \right)^d. \tag{14}$$

As for the numerator of the second term in Eq. 13, Lemma 8 implies that we have

$$\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right) \leq 3 \left(\frac{4\varepsilon^2}{\pi(1-2\sqrt{c_d})j} \right)^d. \quad (15)$$

By plugging Eq. 15 and Eq. 14 in Eq. 13, for $j \geq k(1 - \frac{1}{d})$ and $d > 1$ we can upper bound the factor $\frac{\Pr(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})})}{\Pr(\mathcal{E}_{S_0}^{(\bar{z})})^2}$ of the summation as follows:

$$\begin{aligned} \frac{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)^2} &\leq \frac{3 \left(\frac{4\varepsilon^2}{\pi(1-2\sqrt{c_d})j} \right)^d}{\frac{1}{256} \left(\frac{4\varepsilon^2}{\pi(1+2\sqrt{c_d}+2c_d)k} \right)^d} \\ &= 768 \left(\frac{(1+2\sqrt{c_d}+2c_d)k}{(1-2\sqrt{c_d})j} \right)^d. \end{aligned}$$

As $j \geq k(1 - \frac{1}{d})$ with $d > 1$, then

$$\begin{aligned} 768 \left(\frac{(1+2\sqrt{c_d}+2c_d)k}{(1-2\sqrt{c_d})j} \right)^d &\leq 768 \left(\frac{(1+2\sqrt{c_d}+2c_d)}{(1-2\sqrt{c_d})(1-\frac{1}{d})} \right)^d \\ &\leq 768 \left(\frac{2+7\sqrt{c_d}}{2-7\sqrt{c_d}} \right)^d \\ &\leq 270801, \end{aligned} \quad (16)$$

because $\left(\frac{(1+2\sqrt{c_d}+2c_d)}{(1-2\sqrt{c_d})(1-\frac{1}{d})} \right)^d$ is maximised or $d = 4$. Let $C' = 270801$. If $d > 1$, by plugging Eq. 16 in Eq. 13, we have that

$$\begin{aligned} \frac{\mathbb{E}[T_{n,k}^2]}{\mathbb{E}[T_{n,k}]^2} &= \sum_{j=0}^{\lceil k-\frac{k}{d} \rceil - 1} \Pr(|S \cap S'| = k-j) \frac{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)^2} \\ &\quad + \sum_{j=\lceil k-\frac{k}{d} \rceil}^k \Pr(|S \cap S'| = k-j) \frac{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)^2} \\ &\leq \sum_{j=0}^{\lceil k-\frac{k}{d} \rceil - 1} \Pr(|S \cap S'| = k-j) \frac{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)^2} + C'. \end{aligned}$$

As $\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right) \leq \Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)$, then

$$\begin{aligned} &\sum_{j=0}^{\lceil k-\frac{k}{d} \rceil - 1} \Pr(|S \cap S'| = k-j) \frac{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})} \cap \mathcal{E}_{S_j}^{(\bar{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)^2} + C' \\ &\leq \frac{1}{\Pr\left(\mathcal{E}_{S_0}^{(\bar{z})}\right)} \sum_{j=0}^{\lceil k-\frac{k}{d} \rceil - 1} \Pr(|S \cap S'| = k-j) + C' \end{aligned}$$

$$\leq \frac{\Pr(|S \cap S'| > \frac{k}{d})}{\Pr(\mathcal{E}_{S_0}^{(\tilde{z})})} + C'. \quad (17)$$

Notice that, if $d = 1$, the same bound holds as $\Pr(|S \cap S'| > \frac{k}{d}) = 0$. We now observe that, by the law of total probability

$$\Pr\left(|S \cap S'| \geq \frac{k}{d}\right) = \sum_{\tilde{S} \in \binom{[n]}{k}} \Pr(S = \tilde{S}) \Pr\left(|S \cap S'| \geq \frac{k}{d} \mid S = \tilde{S}\right). \quad (18)$$

Conditional on $S = \tilde{S}$, $|S \cap S'|$ is a hypergeometric random variable with

$$\mathbb{E}\left[|S \cap S'| \mid S = \tilde{S}\right] = \sum_{i \in \tilde{S}} \Pr(i \in S') = k \Pr(1 \in S) = \frac{k^2}{n}.$$

Since $n \geq k^2$, then $\frac{k^2}{n} \leq 1$. Hence, since Chernoff bounds holds for the hypergeometric distribution [Doe20, Theorem 1.10.25]

$$\begin{aligned} \Pr\left(|S \cap S'| \geq \frac{k}{d} \mid S = \tilde{S}\right) &\geq \Pr\left(|S \cap S'| \geq \frac{k^2}{n} + \left(\frac{k}{d} - 1\right) \mid S = \tilde{S}\right) \\ &\leq \exp\left(-2 \frac{\left(\frac{k}{d} - 1\right)^2}{k}\right) \\ &\leq \exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2\right). \end{aligned} \quad (19)$$

Substituting Eq. 19 in Eq. 18 we get

$$\Pr\left(|S \cap S'| \geq \frac{k}{d}\right) \leq \exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2\right). \quad (20)$$

We can now keep bounding from above $\frac{\mathbb{E}[T_{n,k}^2]}{\mathbb{E}[T_{n,k}]^2}$ by plugging Eq. 20 in Eq. 17:

$$\frac{\Pr(|S \cap S'| \geq \frac{k}{d})}{\Pr(\mathcal{E}_{S_0}^{(\tilde{z})})} + C' \leq \frac{\exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2\right)}{\Pr(\mathcal{E}_{S_0}^{(\tilde{z})})} + C'. \quad (21)$$

By Lemma 7, and since $1 + 2\sqrt{cd} + 2c_d \leq 2$, we have

$$\begin{aligned} &\frac{\exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2\right)}{\Pr(\mathcal{E}_{S_0}^{(\tilde{z})})} + C' \\ &\leq \frac{16 \exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2\right)}{\left(\frac{2\varepsilon}{\sqrt{\pi 2k}}\right)^d} + C' \\ &= 16 \exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2 + d \log\left(\frac{\sqrt{\pi 2k}}{2\varepsilon}\right)\right) + C'. \end{aligned} \quad (22)$$

By the hypothesis, since $k \geq Cd^3 \log \frac{d}{\varepsilon}$ for a large enough C , it holds that

$$16 \exp\left(-2 \frac{k}{d^2} \left(1 - \frac{d}{k}\right)^2 + d \log\left(\frac{\sqrt{\pi 2k}}{2\varepsilon}\right)\right) + C'$$

$$\begin{aligned}
&\leq C' + 16 \exp\left(-d \log \frac{d}{\varepsilon}\right) \\
&< C' + \frac{16}{e},
\end{aligned} \tag{23}$$

where the latter inequality holds as $\varepsilon \leq 1/4$.

Plugging the inverse of the expression in Eq. 23 in Eq. 10 we obtain the thesis. \square

B.2 Kernel Pruning

Proof of Lemma 11. $S_2 \in \{0, 1\}^{\text{size}(V)}$ is such that $\tilde{V} = V \odot \tilde{S} = (V \odot S_1) \odot S_2$ contains only non-negative edges going from each input channel t to the output channels $2(t-1)n+1, \dots, (2t-1)n$, and only non-positive⁶ edges going from each input channel t to the output channels $(2t-1)n+1, \dots, 2tn$, while all remaining edges are set to zero. Let us define some convenient notations before proceeding with the proofs. By $[n, m]$ we denote the set $\{n, n+1, \dots, m\}$ for each pair of integers $n \leq m \in \mathbb{N}$. In formulas, we obtain a tensor \tilde{V} such that, for each $(t, k) \in [c], \times [2nc]$:

$$(V \odot \tilde{S})_{1,1,t,k} = \begin{cases} V_{1,1,t,k} \cdot \mathbf{1}_{V_{1,1,t,k} > 0} & \text{if } k \in [(2t-2)n+1, (2t-1)n], \\ V_{1,1,t,k} \cdot \mathbf{1}_{V_{1,1,t,k} < 0} & \text{if } k \in [(2t-1)n+1, 2tn], \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

To simplify the notation, we define the following indicator functions: for any $(t, k) \in [c] \times [2nc]$,

$$\begin{aligned}
\mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} &= 1 \text{ iff } k \in [(2t-2)n+1, (2t-1)n], \text{ and} \\
\mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]} &= 1 \text{ iff } k \in [(2t-1)n+1, 2tn].
\end{aligned} \tag{25}$$

For each $(i, j, k) \in [D] \times [D] \times [2nc]$, applying Eq. 24 and using Definition 10, it then holds

$$\begin{aligned}
&\left(\phi\left(\left(V \odot \tilde{S}\right) * X\right)\right)_{i,j,k} \\
&= \phi\left(\sum_{t=1}^{c_0} \tilde{V}_{1,1,t,k} X_{i,j,t}\right) \\
&= \phi\left(\sum_{t=1}^{c_0} \left(V_{1,1,t,k} X_{i,j,t} \cdot \mathbf{1}_{V_{1,1,t,k} > 0} \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} \right. \right. \\
&\quad \left. \left. + V_{1,1,t,k} X_{i,j,t} \cdot \mathbf{1}_{V_{1,1,t,k} < 0} \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]}\right)\right) \\
&= \phi\left(\sum_{t=1}^{c_0} \left(V_{1,1,t,k}^+ X_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} - V_{1,1,t,k}^- X_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]}\right)\right) \\
&= \phi\left(\sum_{t=1}^{c_0} \left(V_{1,1,t,k}^+ (X_{i,j,t}^+ - X_{i,j,t}^-) \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} \right. \right. \\
&\quad \left. \left. + V_{1,1,t,k}^- (X_{i,j,t}^- - X_{i,j,t}^+) \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]}\right)\right).
\end{aligned} \tag{26}$$

Observe that only one term survives in the summation in Eq. 26, as there exists only one $t \in [c_0]$ such that $k \in [(2t-2)n+1, 2tn]$, say t^* . Moreover, out of the four additive terms in the expression

$$V_{1,1,t^*,k}^+ (X_{i,j,t^*}^+ - X_{i,j,t^*}^-) \mathbf{1}_{\frac{k}{2n} \in (t^*-1, t^*-\frac{1}{2}]} + V_{1,1,t^*,k}^- (X_{i,j,t^*}^- - X_{i,j,t^*}^+) \mathbf{1}_{\frac{k}{2n} \in (t^*-\frac{1}{2}, t^*]},$$

at most one is non-zero, due to Definition 10. The ReLU cancels out negative ones, implying that Eq. 26 can be rewritten without the ReLU as a sum of only non-negative terms (out of which, at most one is non-zero)

⁶We consider 0 to be both non-negative and non-positive.

as follows

$$\begin{aligned}
& \phi \left(\sum_{t=1}^{c_0} \left(V_{1,1,t,k}^+ (X_{i,j,t}^+ - X_{i,j,t}^-) \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} \right. \right. \\
& \quad \left. \left. + V_{1,1,t,k}^- (X_{i,j,t}^- - X_{i,j,t}^+) \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]} \right) \right) \\
&= \sum_{t=1}^{c_0} \left(V_{1,1,t,k}^+ X_{i,j,t}^+ \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} + V_{1,1,t,k}^- X_{i,j,t}^- \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]} \right). \tag{27}
\end{aligned}$$

Finally, by Eq. 24 and Eq. 25, $\tilde{V}_{1,1,t,k}^+ = 0$ if $\frac{k}{2n} \notin (t-1, t-\frac{1}{2}]$, and $\tilde{V}_{1,1,t,k}^- = 0$ if $\frac{k}{2n} \in (t-\frac{1}{2}, t]$, which means that in Eq. 27 we can ignore the indicator functions and further simplify the expression as

$$\begin{aligned}
& \sum_{t=1}^{c_0} \left(V_{1,1,t,k}^+ X_{i,j,t}^+ \mathbf{1}_{\frac{k}{2n} \in (t-1, t-\frac{1}{2}]} + V_{1,1,t,k}^- X_{i,j,t}^- \mathbf{1}_{\frac{k}{2n} \in (t-\frac{1}{2}, t]} \right) \\
&= \sum_{t=1}^{c_0} \left(\tilde{V}_{1,1,t,k}^+ X_{i,j,t}^+ + \tilde{V}_{1,1,t,k}^- X_{i,j,t}^- \right) \\
&= \left(\sum_{t=1}^{c_0} \tilde{V}_{1,1,t,k}^+ X_{i,j,t}^+ + \sum_{t=1}^{c_0} \tilde{V}_{1,1,t,k}^- X_{i,j,t}^- \right) \\
&= \left(\tilde{V}^+ * X^+ + \tilde{V}^- * X^- \right)_{i,j,k}.
\end{aligned}$$

□

Proof of Lemma 12. Adopting the same definitions as in Lemma 11 (and Eq. 24), for each $(r, s, t_1) \in [d] \times [d] \times [c_1]$ we have, by Lemma 11,

$$\begin{aligned}
& \left(U * \phi \left((V \odot \tilde{S}) * X \right) \right)_{r,s,t_1} \\
&= \left(U * \left((\tilde{V}^+ * X^+) + (\tilde{V}^- * X^-) \right) \right)_{r,s,t_1} \\
&= \sum_{i,j \in [d], k \in [2nc_0]} U_{i,j,k,t_1} \cdot \left((\tilde{V}^+ * X^+) + (\tilde{V}^- * X^-) \right)_{r-i+1, s-j+1, k} \\
&= \sum_{i,j \in [d], k \in [2nc_0]} U_{i,j,k,t_1} \cdot \sum_{t_0 \in [c_0]} \left(\tilde{V}_{1,1,t_0,k}^+ \cdot X_{r-i+1, s-j+1, t_0}^+ \right. \\
& \quad \left. + \tilde{V}_{1,1,t_0,k}^- \cdot X_{r-i+1, s-j+1, t_0}^- \right) \\
&= \sum_{t_0 \in [c_0]} \sum_{i,j \in [d], k \in [2nc_0]} \left(U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^+ \right) \cdot X_{r-i+1, s-j+1, t_0}^+ \\
& \quad + \sum_{t_0 \in [c_0]} \sum_{i,j \in [d], k \in [2nc_0]} \left(U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^- \right) \cdot X_{r-i+1, s-j+1, t_0}^- \\
&= \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [2nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^+ \right) \cdot X_{r-i+1, s-j+1, t_0}^+ \\
& \quad + \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [2nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^- \right) \cdot X_{r-i+1, s-j+1, t_0}^-.
\end{aligned}$$

We remind the reader that $\tilde{S} = S_1 \odot S_2$, with S_1 being a $2n$ -channel-blocked mask and S_2 being a mask that removes filters. Define $L_{i,j,t_0,t_1}^+ = \sum_{k \in [nc]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^+$ and, similarly, $L_{i,j,t_0,t_1}^- = \sum_{k \in [nc]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^-$.

Then,

$$\begin{aligned}
& \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^+ \right) \cdot X_{r-i+1,s-j+1,t_0}^+ \\
& + \sum_{i,j \in [d], t_0 \in [c_0]} \left(\sum_{k \in [nc_0]} U_{i,j,k,t_1} \cdot \tilde{V}_{1,1,t_0,k}^- \right) \cdot X_{r-i+1,s-j+1,t_0}^- \\
= & \sum_{i,j \in [d], t_0 \in [c_0]} L_{i,j,t_0,t_1}^+ \cdot X_{r-i+1,s-j+1,t_0}^+ + \sum_{i,j \in [d], t_0 \in [c_0]} L_{i,j,t_0,t_1}^- \cdot X_{r-i+1,s-j+1,t_0}^-.
\end{aligned}$$

We now show that, for each $t_0 \in [c_0]$, $K_{:::,t_0,:}$ can be ε -approximated by $L_{:::,t_0,:}^+$ by suitably pruning \tilde{V}^+ , i.e. by further zeroing entries of \tilde{S} , and that such pruning corresponds to solving an instance of MRSS according to Theorem 5. The same reasoning applies to K^- and L^- .

For each $t_0 \in [c_0]$, let

$$I_+^{(t_0)} = \left\{ k \in \{(2t_0 - 2)n + 1, \dots, (2t_0 - 1)n\} : \tilde{S}_{1,1,t_0,k} = 1 \right\}.$$

Observe that $I_+^{(t_0)}$ consists of the strictly positive entries of $\tilde{V}_{1,1,t_0,:}^+$.⁷ Since the entries of V follow a standard normal distribution, each entry is positive with probability $1/2$. By a standard application of Chernoff bounds (Lemma 16 in Appendix A), we then have

$$\Pr \left(\left| I_+^{(t_0)} \right| > \frac{n}{3} \right) \geq 1 - \frac{\varepsilon}{4}, \quad (28)$$

provided that the constant C in the bound on n is sufficiently large.

For each $k \in I_+^{(t_0)}$, up to reshaping the tensor as a one-dimensional vector, $U_{:::,k,:} \cdot \tilde{V}_{1,1,t_0,k}^+$ is an NSN vector (Definition 4) by Lemma 18 (Appendix A). Thus, for each $t_0 \in [c_0]$ and a sufficiently-large value of C , since we have $n \geq Cd^{13}c_1^6 \log^3 \frac{d^2c_1c_0}{\varepsilon}$, we can apply an amplified version of Theorem 5 (i.e. Corollary 19 in Appendix A with vectors of dimension d^2c_1) to show that, with probability $1 - \frac{\varepsilon}{4c_0}$, for all target filter K such that $\|K_{:::,t_0,:}\|_1 \leq 1$, there exists a way to zero the entries indexed by $I_+^{(t_0)}$ of \tilde{S} (and thus $\tilde{V}_{1,1,t_0,:}^+$), so that the pruned version of $L_{:::,t_0,:}^+ = \sum_{k \in [nc_0]} U_{:::,k,:} \cdot \tilde{V}_{1,1,t_0,k}^+$ approximates $K_{:::,t_0,:}$. In particular, there exists another binary mask $S_3^+ \in \{0, 1\}^{\text{size } \tilde{S}}$ such that $\hat{L}_{:::,t_0,:}^+ = \sum_{k \in [nc_0]} U_{:::,k,:} \cdot \hat{V}_{1,1,t_0,k}^+$ approximates $K_{:::,t_0,:}$, where $\hat{V}^+ = \tilde{V}^+ \odot S_3^+$. An analogous argument carries on for a binary mask S_3^- and $-\hat{L}_{:::,t_0,:}^-$.⁸ More formally, let

$$\begin{aligned}
\mathcal{E}_{t_0,+}^{(\text{kernel})} &= \left\{ \forall K : \|K\|_1 \leq 1, \exists S_3^+ \in \{0, 1\}^{\text{size } \tilde{S}} \left\| \hat{L}_{:::,t_0,:}^+ - K_{:::,t_0,:} \right\|_{\max} \leq \frac{\varepsilon}{2d^2c_1c_0} \right\}, \\
\mathcal{E}_{t_0,-}^{(\text{kernel})} &= \left\{ \forall K : \|K\|_1 \leq 1, \exists S_3^- \in \{0, 1\}^{\text{size } \tilde{S}} \left\| \hat{L}_{:::,t_0,:}^- + K_{:::,t_0,:} \right\|_{\max} \leq \frac{\varepsilon}{2d^2c_1c_0} \right\}, \text{ and} \\
\mathcal{E}^{(\text{kernel})} &= \left(\bigcap_{t_0 \in [c_0]} \mathcal{E}_{t_0,+}^{(\text{kernel})} \right) \cap \left(\bigcap_{t_0 \in [c_0]} \mathcal{E}_{t_0,-}^{(\text{kernel})} \right).
\end{aligned}$$

Then, by Corollary 19,

$$\begin{aligned}
\Pr \left(\mathcal{E}_{t_0,+}^{(\text{kernel})} \mid \left| I_+^{(t_0)} \right| > \frac{n}{3} \right) &\geq 1 - \frac{\varepsilon}{4c_0}, \text{ and} \\
\Pr \left(\mathcal{E}_{t_0,-}^{(\text{kernel})} \mid \left| I_-^{(t_0)} \right| > \frac{n}{3} \right) &\geq 1 - \frac{\varepsilon}{4c_0}.
\end{aligned}$$

⁷Notice that excluding zero entries implies conditioning on the event that the entry is not zero. However, such an event has zero probability and, thus, does not impact the analysis.

⁸The negative sign in front of $\hat{L}_{:::,t_0,:}^-$ does not affect the random subset sum result as each entry is independently negative or positive with the same probability.

By the union bound, we have the following:

$$\begin{aligned}
& \Pr\left(\mathcal{E}^{(\text{kernel})} \mid \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \\
&= 1 - \Pr\left(\left(\bigcup_{t_0 \in [c_0]} \bar{\mathcal{E}}_{t_0,+}^{(\text{kernel})}\right) \cup \left(\bigcup_{t_0 \in [c_0]} \bar{\mathcal{E}}_{t_0,-}^{(\text{kernel})}\right) \mid \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \\
&\geq 1 - \sum_{t_0 \in [c_0]} \left[\Pr\left(\bar{\mathcal{E}}_{t_0,+}^{(\text{kernel})} \mid \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) + \Pr\left(\bar{\mathcal{E}}_{t_0,-}^{(\text{kernel})} \mid \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right)\right] \\
&\geq 1 - 2 \sum_{t_0 \in [c_0]} \frac{\varepsilon}{4c_0} \\
&\geq 1 - \frac{\varepsilon}{2}.
\end{aligned}$$

Since $\Pr\left(\left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \geq 1 - \frac{\varepsilon}{2}$, then we can remove the conditional event obtaining

$$\begin{aligned}
\Pr\left(\mathcal{E}^{(\text{kernel})}\right) &\geq \Pr\left(\mathcal{E}^{(\text{kernel})} \mid \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \Pr\left(\left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \\
&\geq \left(1 - \frac{\varepsilon}{2}\right)^2 \\
&\geq 1 - \varepsilon.
\end{aligned} \tag{29}$$

To rewrite the latter in terms of the filter K and a mask \hat{S} , we notice that pruning $L_{:::,t_0,:}^+$ and $L_{:::,t_0,:}^-$ separately, with two binary masks, is equivalent to say that there exists a single binary mask $S_3 \in \{0, 1\}^{\text{size } \tilde{S}}$ such that, $\hat{L}_{:::,t_0,:}$ can be written as $\hat{L}_{:::,t_0,:} = \sum_{k \in [nc_0]} U_{:::,k,:} \cdot \hat{V}_{1,1,t_0,k}$, where $\hat{V} = \tilde{V} \odot S_3$. Ineq. 29 implies that, with probability $1 - \varepsilon$, for all target filters K with $\|K\|_1 \leq 1$ such S_3 exists and hence,

$$\left\|K - \hat{L}^+\right\|_{\max} + \left\|K + \hat{L}^-\right\|_{\max} \leq \frac{\varepsilon}{d^2 c_1 c_0}. \tag{30}$$

Let $\hat{S} = \tilde{S} \odot S_3$: then $\hat{S} = S_1 \odot S_2 \odot S_3$ is the combination of a $2n$ -channel-blocked mask $S = S_1$ and a mask $S_2 \odot S_3$ that only removes filters.

Notice that, with probability $1 - \varepsilon$, for all target filters K with $\|K\|_1 \leq 1$, such binary masks \hat{S} exists and the following holds:

$$\begin{aligned}
& \sup_{X: \|X\|_{\max} \leq M} \left\|K * X - N_0^{(\hat{S})}(X)\right\|_{\max} \\
&= \sup_{X: \|X\|_{\max} \leq M} \left\|K * X - U * \phi\left((V \odot \hat{S}) * X\right)\right\|_{\max} \\
&= \sup_{X: \|X\|_{\max} \leq M} \left\|K * X - U * \phi\left((V \odot \tilde{S} \odot S_3) * X\right)\right\|_{\max} \\
&= \sup_{X: \|X\|_{\max} \leq M} \left\|K * (X^+ - X^-) - U * \left(\left(\hat{V}^+ * X^+\right) + \left(\hat{V}^- * X^-\right)\right)\right\|_{\max},
\end{aligned}$$

where the latter holds by Lemma 11.⁹ Then, by the distributive property of the convolution and the triangle inequality,

$$\begin{aligned}
& \sup_{X: \|X\|_{\max} \leq M} \left\|K * (X^+ - X^-) - U * \left(\left(\hat{V}^+ * X^+\right) + \left(\hat{V}^- * X^-\right)\right)\right\|_{\max} \\
&= \sup_{X: \|X\|_{\max} \leq M} \left\|K * X^+ - U * \left(\hat{V}^+ * X^+\right) - K * X^- - U * \left(\hat{V}^- * X^-\right)\right\|_{\max}
\end{aligned}$$

⁹The presence of S_3 does not influence the proof of Lemma 11.

$$\begin{aligned} &\leq \sup_{X: \|X\|_{\max} \leq M} \left\| K * X^+ - U * (\hat{V}^+ * X^+) \right\|_{\max} \\ &\quad + \sup_{X: \|X\|_{\max} \leq M} \left\| K * X^- + U * (\hat{V}^- * X^-) \right\|_{\max}. \end{aligned}$$

One can now apply the Tensor Convolution Inequality (Lemma 20) and obtain

$$\begin{aligned} &\sup_{X: \|X\|_{\max} \leq M} \left\| K * X^+ - U * (\hat{V}^+ * X^+) \right\|_{\max} \\ &\quad + \sup_{X: \|X\|_{\max} \leq M} \left\| K * X^- + U * (\hat{V}^- * X^-) \right\|_{\max} \\ &\leq \sup_{X: \|X\|_{\max} \leq M} \|X^+\|_{\max} \cdot \|K - U * \hat{V}^+\|_1 \\ &\quad + \sup_{X: \|X\|_{\max} \leq M} \|X^-\|_{\max} \cdot \|K + U * \hat{V}^-\|_1 \\ &= M \cdot \|K - U * \hat{V}^+\|_1 + M \cdot \|K + U * \hat{V}^-\|_1. \end{aligned}$$

Now, observing that the number of entries of the two tensors in the expression above is $d^2 c_1 c_0$, and using Ineq. 30, we get that

$$\begin{aligned} &M \cdot \|K - U * \hat{V}^+\|_1 + M \cdot \|K + U * \hat{V}^-\|_1 \\ &\leq d^2 c_1 c_0 \left(\|K - U * \hat{V}^+\|_{\max} + \|K + U * \hat{V}^-\|_{\max} \right) \\ &\leq d^2 c_1 c_0 M \frac{\varepsilon}{d^2 c_1 c_0} \\ &= \varepsilon M. \end{aligned}$$

proving the thesis. \square

Proof of Theorem 2. To bound the error propagation across layers, we define the layers' outputs

$$\begin{aligned} X^{(0)} &= X, \\ X^{(i)} &= \phi \left(K^{(i)} * X^{(i-1)} \right) \quad \text{for } 1 \leq i \leq \ell. \end{aligned} \tag{31}$$

Notice that $X^{(\ell)}$ is the output of the target function, i.e., $f(X) = X^{(\ell)}$.

For brevity's sake, given masks $S^{(2i-1)} \in \{0, 1\}^{\text{size}(L^{(2i-1)})}$ for each $i = 1, \dots, \ell$, let us denote

$$\tilde{L}^{(2i-1)} = L^{(2i-1)} \odot S^{(2i-1)}. \tag{32}$$

We now show that the random network approximates the target one after being pruned only at layers $L^{(2i-1)}$ for all $i = 1, \dots, \ell$, by masks $S^{(2i-1)}$ which are the composition of a $2n_{2i-1}$ -channel-blocked mask and a mask that only removes filters. The key idea is to iteratively applying Lemma 12. Since the ReLU function is 1-Lipschitz, for all i it holds

$$\begin{aligned} &\left\| \phi \left(K^{(i)} * X^{(i-1)} \right) - \phi \left(L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * X^{(i-1)} \right) \right) \right\|_{\max} \\ &\leq \left\| K^{(i)} * X^{(i-1)} - L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * X^{(i-1)} \right) \right\|_{\max}. \end{aligned} \tag{33}$$

The key step of the proof is that, for each layer i , since $n_i \geq C d^{13} c_i^6 \log^3 \frac{d^2 c_i c_{i-1} \ell}{\varepsilon}$ for a suitable constant C , we can apply Lemma 12 to get that, with probability at least $1 - \frac{\varepsilon}{2\ell}$, for all i and for all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$ it holds

$$\left\| K^{(i)} * X^{(i-1)} - L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * X^{(i-1)} \right) \right\|_{\max} < \frac{\varepsilon}{2\ell} \cdot \|X^{(i-1)}\|_{\max}. \tag{34}$$

Hence, combining Eq. 33 and Eq. 34 we get that, with probability at least $1 - \frac{\varepsilon}{2\ell}$, for all i and all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$ with $\|K^{(i)}\|_1 \leq 1$,

$$\begin{aligned} & \left\| \phi \left(K^{(i)} * X^{(i-1)} \right) - \phi \left(L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * X^{(i-1)} \right) \right) \right\|_{\max} \\ & < \frac{\varepsilon}{2\ell} \cdot \left\| X^{(i-1)} \right\|_{\max}. \end{aligned} \quad (35)$$

By a union bound, with probability at least $1 - \varepsilon$, we get that Eq. 35 holds for all layers i and all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$.

Analogously, we can define the pruned layers' outputs

$$\begin{aligned} \tilde{X}^{(0)} &= X, \\ \tilde{X}^{(i)} &= \phi \left(L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * \tilde{X}^{(i-1)} \right) \right) \quad \text{for } 1 \leq i \leq \ell. \end{aligned} \quad (36)$$

Notice that $\tilde{X}^{(\ell)}$ is the output of the pruned network, i.e. $N_0^{(S^{(1)}, \dots, S^{(2\ell)})}(X) = \tilde{X}^{(\ell)}$.

By the same reasoning employed to derive Eq. 34 and Eq. 35 we have that, with probability $1 - \varepsilon$, for all layers i and all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$, the output of all pruned layers satisfies

$$\begin{aligned} & \left\| \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) - \phi \left(L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * \tilde{X}^{(i-1)} \right) \right) \right\|_{\max} \\ & < \frac{\varepsilon}{2\ell} \cdot \left\| \tilde{X}^{(i-1)} \right\|_{\max}. \end{aligned} \quad (37)$$

Moreover, for each $1 \leq i \leq \ell - 1$, by the triangle inequality and by Eq. 37,

$$\begin{aligned} \left\| \tilde{X}^{(i)} \right\|_{\max} &= \left\| \tilde{X}^{(i)} - \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) + \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) \right\|_{\max} \\ &\leq \left\| \tilde{X}^{(i)} - \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) \right\|_{\max} + \left\| \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) \right\|_{\max} \\ &\leq \frac{\varepsilon}{2\ell} \cdot \left\| \tilde{X}^{(i-1)} \right\|_{\max} + \left\| \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) \right\|_{\max}. \end{aligned}$$

By the Lipschitz property of ϕ and Lemma 20

$$\begin{aligned} & \frac{\varepsilon}{2\ell} \cdot \left\| \tilde{X}^{(i-1)} \right\|_{\max} + \left\| \phi \left(K^{(i)} * \tilde{X}^{(i-1)} \right) \right\|_{\max} \\ & \leq \frac{\varepsilon}{2\ell} \cdot \left\| \tilde{X}^{(i-1)} \right\|_{\max} + \left\| K^{(i)} * \tilde{X}^{(i-1)} \right\|_{\max} \\ & \leq \frac{\varepsilon}{2\ell} \cdot \left\| \tilde{X}^{(i-1)} \right\|_{\max} + \left\| K^{(i)} \right\|_1 \left\| \tilde{X}^{(i-1)} \right\|_{\max} \\ & = \left\| \tilde{X}^{(i-1)} \right\|_{\max} \left(1 + \frac{\varepsilon}{2\ell} \right). \end{aligned}$$

By unrolling the recurrence, we get that, with probability $1 - \varepsilon$, for all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$,

$$\left\| \tilde{X}^{(i)} \right\|_{\max} \leq \left\| \tilde{X}^{(0)} \right\|_{\max} \left(1 + \frac{\varepsilon}{2\ell} \right)^i. \quad (38)$$

Thus, combining Eq. 37 and Eq. 38, with probability $1 - \varepsilon$ we get that, for each $i \in [\ell]$ and for all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$,

$$\begin{aligned} & \left\| K^{(i)} * \tilde{X}^{(i-1)} - L^{(2i)} * \phi \left(\tilde{L}^{(2i-1)} * \tilde{X}^{(i-1)} \right) \right\|_{\max} \\ & < \frac{\varepsilon}{2\ell} \cdot \left(1 + \frac{\varepsilon}{2\ell} \right)^{i-1} \left\| \tilde{X}^{(0)} \right\|_{\max}. \end{aligned} \quad (39)$$

We then see that with probability $1 - \varepsilon$, for $1 \leq i \leq \ell$ and for all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$, by Eq. 31 and Eq. 36, and by the triangle inequality,

$$\left\| X^{(\ell)} - \tilde{X}^{(\ell)} \right\|_{\max}$$

$$\begin{aligned}
&= \left\| \phi \left(K^{(\ell)} * X^{(\ell-1)} \right) - \phi \left(L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right) \right\|_{\max} \\
&\leq \left\| \phi \left(K^{(\ell)} * X^{(\ell-1)} \right) - \phi \left(K^{(\ell)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\quad + \left\| \phi \left(K^{(\ell)} * \tilde{X}^{(\ell-1)} \right) - \phi \left(L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right) \right\|_{\max}.
\end{aligned}$$

Again by the 1-Lipschitz property of the ReLU activation function, and by the distributive property of the convolution operation,

$$\begin{aligned}
&\left\| \phi \left(K^{(\ell)} * X^{(\ell-1)} \right) - \phi \left(K^{(\ell)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&+ \left\| \phi \left(K^{(\ell)} * \tilde{X}^{(\ell-1)} \right) - \phi \left(L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right) \right\|_{\max} \\
&\leq \left\| K^{(\ell)} * X^{(\ell-1)} - K^{(\ell)} * \tilde{X}^{(\ell-1)} \right\|_{\max} \\
&\quad + \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&= \left\| K^{(\ell)} * \left(X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\quad + \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max}.
\end{aligned}$$

Lemma 20 and the hypothesis $\|K^{(\ell)}\|_1 \leq 1$ imply that

$$\begin{aligned}
&\left\| K^{(\ell)} * \left(X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&+ \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\leq \left\| K^{(\ell)} \right\|_1 \cdot \left\| \left(X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\quad + \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\leq \left\| \left(X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\quad + \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max}.
\end{aligned}$$

Now, we first apply Eq. 39 and then we unroll the recurrence for all layers (as, with probability $1 - \varepsilon$, Eq. 39 holds for all layers and for all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$), obtaining

$$\begin{aligned}
&\left\| \left(X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&+ \left\| K^{(\ell)} * \tilde{X}^{(\ell-1)} - L^{(2\ell)} * \phi \left(\tilde{L}^{(2\ell-1)} * \tilde{X}^{(\ell-1)} \right) \right\|_{\max} \\
&\leq \left\| X^{(\ell-1)} - \tilde{X}^{(\ell-1)} \right\|_{\max} + \frac{\varepsilon}{2\ell} \cdot \left(1 + \frac{\varepsilon}{2\ell} \right)^{\ell-1} \\
&\leq \sum_{j=1}^{\ell} \frac{\varepsilon}{2\ell} \cdot \left(1 + \frac{\varepsilon}{2\ell} \right)^{j-1}.
\end{aligned}$$

By summing the geometric series and observing that $\varepsilon < 1$, we conclude that

$$\begin{aligned}
\sum_{j=1}^{\ell} \frac{\varepsilon}{2\ell} \cdot \left(1 + \frac{\varepsilon}{2\ell} \right)^{j-1} &= \left(1 + \frac{\varepsilon}{2\ell} \right)^{\ell} - 1 \\
&\leq e^{\frac{\varepsilon}{2}} - 1 \\
&\leq \varepsilon.
\end{aligned}$$

Hence, with probability $1 - \varepsilon$, for all $X \in [-1, 1]^{D \times D \times c_0}$ and all choices of target filters $K^{(i)} \in \mathbb{R}^{d_i \times d_i \times c_{i-1} \times c_i}$, for all $\ell \in [c]$ it holds that

$$\left\| X^{(\ell)} - \tilde{X}^{(\ell)} \right\|_{\max} \leq \varepsilon,$$

yielding the thesis. □