Convolutional neural networks contain structured strong lottery tickets

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

Convolutional neural networks contain structured strong lottery tickets

Arthur da Cunha
Université Côte d'Azur, Inria, CNRS, I3S
Sophia Antipolis, France
arthur.carvalho-walraven-da-cunha@inria.fr

Francesco d'Amore
Aalto University
Espoo, Finland
francesco.damore@aalto.fi

Emanuele Natale
Université Côte d'Azur, Inria, CNRS, I3S
Sophia Antipolis, France
emanuele.natale@inria.fr

Abstract

The Strong Lottery Ticket Hypothesis (SLTH) states that randomly-initialised neural networks contain subnetworks that can 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.

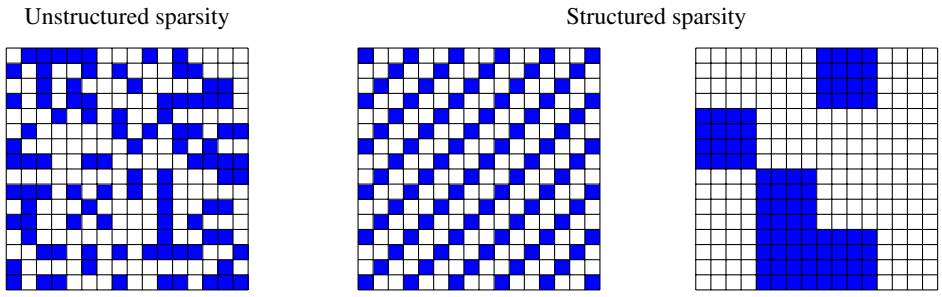
This is the first work to address the SLTH for structured pruning, opening up new avenues for further research on the hypothesis and contributing to the understanding of the role of overparameterization in deep learning.

1 Introduction

Much of the success of deep learning techniques relies on extreme overparameterization. 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 Reed [1993] and in the present Blalock et al. [2020], Hoefler et al. [2021].

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 initialization without sacrificing model accuracy. In particular, Frankle and Carbin [2019] 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 Zhou et al. [2019], Ramanujan et al. [2020], Wang et al. [2020] pointed towards an even more impressive phenomenon: the existence of subnetworks that perform well without any training. This conjecture was named the *Strong Lottery Ticket Hypothesis (SLTH)* by Pensia et al. [2020].

Preprint. Under review.



(a) No pattern. (b) Strided pattern. (c) Block pattern.

Figure 1: Examples of different pruning patterns.

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 main gains of pruning, both in terms of memory and computational efficiency. Removing 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 Pooch and Nieder [1973]. Moreover, the theoretical computational gains of unstructured sparsity can also be difficult to realize in standard hardware, which is optimized 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. With this constraint, pruning results in a smaller dense network, directly reducing the computational costs without any need for extra memory to store indices. Similarly, deleting entire filters in Convolutional Neural Networks (CNNs) Polyak and Wolf [2015] or "heads" in attention-based architectures Michel et al. [2019] 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 its performance. Through unstructured pruning, on the other hand, one can usually reach 95% sparsity without accuracy loss Alvarez and Salzmann [2016]. In practice, however, the computational advantage of structured pruning can offset this difference. This trade-off between sparsity and actual efficiency has motivated the study of less coarse sparsity patterns. Weaker structural constraints such as strided sparsity Anwar et al. [2017] (Figure 1b) or block sparsity Siswanto [2021] (Figure 1b) are already sufficient to deliver the bulk of the computational gains that structured can offer.

Despite its benefits, there have been no results on structured pruning in the context of the SLTH. We believe this gap can be attributed to the limitations of a central result underlying almost all of the theoretical works on the SLTH: a theorem by Lueker on the *Random Subset-Sum Problem (RSSP)*.

Theorem 1 ([Lueker, 1998, da Cunha et al., 2022a]). *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 as 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.[1]

While Theorem 1 closely matches the setup of the SLTH, it only concerns individual random variables, so it does not apply to entire random structures directly. The recent works Borst et al. [2022], Becchetti et al. [2022a] 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 close this gap by providing a version of Theorem 1 that allows us to prove that networks in a wide class of CNNs are likely to contain structured subnetworks that approximate any sufficiently smaller CNN in the class. To the best of our knowledge, this is the first result around the SLTH for structured pruning of 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 use this result to show that, with high probability, a rather wide class of random CNNs can be pruned (in a structured manner) to approximate any sufficiently smaller CNN in this class (Theorem 3);
- Additionally, our pruning scheme focuses on filter pruning, which, like neuron pruning, allows for a direct reduction of the size of the original CNN.

2 Related Work

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

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

Malach et al. [2020] 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}. Orseau et al. [2020] showed that the width overhead could be greatly reduced by sampling parameters from a hyperbolic distribution. Pensia et al. [2020] improved the original result for a wide class of weight distribution, requiring only a logarithmic width overhead, which they proved to be asymptotically optimal. da Cunha et al. [2022b] generalised those results with optimal bounds to CNNs with non-negative inputs, which Burkholz [2022a] extended to general inputs and to residual architectures. Burkholz [2022a] also reduced the depth overhead to a single extra layer and provided results that include a whole class of activation functions. Burkholz [2022b] obtained similar improvements to dense architectures. Fischer and Burkholz [2021] modified many of the previous arguments to take into consideration networks with non-zero biases. Ferbach et al. [2022] further generalise previous results on CNNs to general equivariant networks. Diffenderfer and Kailkhura [2021] obtained similar SLTH results for binary dense neural networks within polynomial depth and width overhead, which Sreenivasan et al. [2022] 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 Mozer and Smolensky [1988] and Mozer and Smolensky [1989]. Since then, a vast literature was 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 Hoefler et al. [2021], and to He and Xiao [2023] for a survey on structured pruning of CNNs.

RSSP Pensia et al. [2020] introduced the use of theoretical results on the RSSP in arguments around the SLTH, namely [Lueker, 1998, Corollary 3.3]. The work da Cunha et al. [2022a] provides an alternative, simpler proof of this result. Borst et al. [2022] and Becchetti et al. [2022b] prove

[1] Distributions whose probability density function f satisfies $f(x) \geq b$ for all $x \in [-a, a]$, for some constants $a, b > 0$ (see [Lueker, 1998, Corollary 3.3]).

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 $X \in \mathbb{R}^{d_1 \times \cdots \times d_n}$ 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 $U \sim \mathcal{N}^{d_1 \times \cdots \times d_n}$ to denote that U is a random tensor of size $d_1 \times \cdots \times d_n$ 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 $N \colon [-1, 1]^{D \times D \times c_0} \to \mathbb{R}^{D \times D \times c_\ell}$ given by

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

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*.

Before we proceed to our main theorem, we introduce a definition that encompasses the sparsity structure underlying our proofs.

Definition 2 (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 $l \in [cn]$.

Theorem 3 (SLTH for kernel pruning)**.** *Let D, d, c_0, c_1 and ℓ be positive integers and let ϵ and C be positive real numbers. For each $i \in [\ell]$, let $L^{(2i-1)} \sim \mathcal{N}^{1 \times 1 \times c_{i-1} \times 2c_{i-1}n_i}$ and $L^{(2i)} \sim \mathcal{N}^{d_i \times d_i \times 2c_{i-1}n_i \times c_i}$ with $n_i \geq Cd^{12}c_i^6 \log^3 \frac{d^2 c_i c_{i-1} \ell}{\epsilon}$ for some positive integers n_i and c_i. Let then $N_0 \colon [-1, 1]^{D \times D \times c_0} \to \mathbb{R}^{D \times D \times c_\ell}$ be a random CNN of the form*

$$N_0(X) = L^{(2\ell)} * \cdots \phi\left(L^{(1)} * X\right).$$

Given $2n_i$-channel-blocked masks $S^{(2i-1)} \in \{0, 1\}^{1 \times 1 \times n_i \times c_i}$ for each tensor $L^{(2i-1)}$, for $i \in [\ell]$; let

$$N_0^{\left(S^{(1)},\dots,S^{(2\ell-1)}\right)} = L^{(2\ell)} * \phi\left(\cdots\left(L^{(2)} * \phi\left(\left(S^{(1)} \odot L^{(1)}\right) * X\right)\right)\right).$$

Finally, let $\mathcal{F}$ be the class of functions $f \colon [-1, 1]^{D \times D \times c_0} \to \mathbb{R}^{D \times D \times c_\ell}$ of the form

$$f(X) = K^{(\ell)} * \cdots \phi\left(K^{(1)} * X\right),$$

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

There exists a universal value of C such that, with probability $1 - \epsilon$, for every $f \in \mathcal{F}$ it is possible to remove filters from $N_0^{\left(S^{(1)},\dots,S^{(2\ell-1)}\right)}$ to obtain a CNN $\tilde{N}_0$ for which

$$\sup_{X \in [-1,1]^{D \times D \times c_0}} \left\| f(X) - \tilde{N}_0(X) \right\|_{\max} \leq \epsilon.$$

The filter removals ensured by Theorem 3 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

$\tilde{L}^{(1)}, \ldots, \tilde{L}^{(2\ell)}$ such that, for $i \in [\ell]$, $\tilde{L}^{(2i-1)} \in \mathbb{R}^{1\times 1\times c_{i-1}\times 2c_{i-1}m_i}$ and $\tilde{L}^{(2i)} \in \mathbb{R}^{d_i\times d_i\times 2c_{i-1}m_i\times c_i}$, where $m_i = \sqrt{n_i/(C_1 \log \frac{1}{\epsilon})}$ for a universal constant C_1. Moreover, the kernels $\tilde{L}^{(2i-1)}$ are structured as if pruned by $2m_i$-channel-blocked masks.

We remark that, from a broader perspective, the central aspect of Theorem 3 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 3. 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 $Z, Z_1, \ldots, Z_d$ 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 [Borst et al., 2022, Lemmas 1, 15].

Theorem 5 (Normally-scaled MRSS)**.** *Let $0 < \epsilon \leq 1/4$, and let d, k, and n be positive integers such that $n \geq k^2$ and $k \geq Cd^3 \log \frac{d}{\epsilon}$ for some universal constant $C \in \mathbb{R}_{>0}$. Furthermore, let $X_1, \ldots, X_n$ 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 $\left\|\left(\sum_{i\in S} X_i\right) - \vec{z}\right\|_{\max} \leq \epsilon$.*

While it is possible to naïvely apply Theorem 1 to obtain a version of Theorem 3, doing so would lead 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 3). Full proofs are deferred to the supplementary material (SM).

4.1 Multidimensional Random Subset Sum for normally-scaled normal vectors

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

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 formalizes this notion.

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

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

where $\mathcal{E}_S^{(\vec{z})}$ denotes the event $\left\|\left(\sum_{i\in S} X_i\right) - \vec{z}\right\|_{\max} \leq \epsilon$. We write simply $T_{n,k}$ when $X_1, \ldots, X_n$ 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 can approximate the target value $\vec{z}$. Hence, we seek a lower bound on $\mathbb{E}[T_{n,k}]^2/\mathbb{E}[T^2_{n,k}]$.

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 $k \in \mathbb{N}$, $\epsilon \in \left(0, \frac{1}{4}\right)$, $\vec{z} \in \mathbb{R}^d$ such that $\|\vec{z}\|_1 \leq \sqrt{k}$ and $k \geq 16$. Furthermore, let $X_1, \ldots, X_k$ be d-dimensional i.i.d. NSN random vectors with $d \leq k$, and*

let $c_d = \min\left\{\frac{1}{d^2}, \frac{1}{16}\right\}$. *It holds that*

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

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

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

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

$$\begin{aligned}
&\mathbb{E}_{Z_1,\ldots,Z_n}\left[\Pr\left(X \in I_\epsilon\left(\vec{z}\right) \mid Z_1,\ldots,Z_k\right)\right] \\
&= \mathbb{E}_{Z_1,\ldots,Z_n}\left[\prod_{j=1}^{d} \Pr\left(\left(X\right)_j \in I_\epsilon\left(z_i\right) \,\middle|\, Z_1,\ldots,Z_k\right)\right] \\
&\geq \mathbb{E}_{Z_1,\ldots,Z_n}\left[\prod_{j=1}^{d} \Pr\left(\left(X\right)_j \in I_\epsilon\left(z_i\right) \,\middle|\, Z_1,\ldots,Z_k,\mathcal{E}^{(\updownarrow)}\right)\right] \Pr\left(\mathcal{E}^{(\updownarrow)}\right), \qquad (1)
\end{aligned}$$

where the inequality in Eq. 1 holds by applying again the law of total probability.

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

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

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 $\mathcal{E}^{(\updownarrow)}$ and by applying Lemma 17 to estimate the term $\Pr\left(\mathcal{E}^{(\updownarrow)}\right)$, we get that

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

$$\geq \left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{\|\vec{z}\|_2^2+2\epsilon\|\vec{z}\|_1+d\epsilon^2}{2\left(1-2\sqrt{c_d}\right)k}\right)\left(1-2e^{-c_dk}\right).$$

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

$$\begin{aligned}
&\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{\|\vec{z}\|_2^2+2\epsilon\|\vec{z}\|_1+d\epsilon^2}{2\left(1-2\sqrt{c_d}\right)k}\right)\left(1-2e^{-c_dk}\right)\\
&\geq \frac{1}{4}\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{\|\vec{z}\|_2^2+2\epsilon\|\vec{z}\|_1+d\epsilon^2}{2\left(1-2\sqrt{c_d}\right)k}\right)\\
&\geq \frac{1}{4}\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{k+2\epsilon\sqrt{k}+d\epsilon^2}{2\left(1-2\sqrt{c_d}\right)k}\right) \qquad (2)\\
&= \frac{1}{4}\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{1+\frac{2\epsilon}{\sqrt{k}}+\frac{d\epsilon^2}{k}}{2\left(1-2\sqrt{c_d}\right)}\right)\\
&\geq \frac{1}{4}\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d \exp\left(-\frac{1+\frac{1}{8}+\frac{1}{16}}{2\left(1-2\sqrt{c_d}\right)}\right) \qquad (3)\\
&\geq \frac{1}{16}\left(\frac{2\epsilon}{\sqrt{\pi\left(1+2\sqrt{c_d}+2c_d\right)k}}\right)^d, \qquad (4)
\end{aligned}$$

where we have used that $\|\vec{z}\|_2 \leq \|\vec{z}\|_1 \leq \sqrt{k}$ in Ineq. 2, that $k \geq 16$, $k \geq d$, that $\epsilon < 1/4$ in Ineq. 3, and that

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

in Ineq. 4. □

Bounding $\mathbb{E}[T_{n,k}^2]$ requires handling stochastic dependencies. Thus, we estimate the joint probability that two subsets of k elements of $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_{k+j}$ be i.i.d. d-dimensional NSN random vectors with $k \geq Cd^3 \log \frac{d}{\epsilon}$. Let $c_d = \min\left\{\frac{1}{d^2}, \frac{1}{16}\right\}$, $A = \sum_{i=1}^{j} X_i$, $B = \sum_{i=j+1}^{k} X_i$, and $C = \sum_{i=k+1}^{k+j} X_i$.[2] Then, it holds that*

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

Proof. Since the X_is 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

$$\Pr\left(A+B \in I_\epsilon\left(\vec{z}\right), B+C \in I_\epsilon\left(\vec{z}\right)\right)$$

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

$$= \mathbb{E}_{Z_1,\ldots,Z_n}\left[\Pr\left(A+B\in I_\epsilon\left(\vec{z}\right), B+C\in I_\epsilon\left(\vec{z}\right)\,|\,Z_1,\ldots,Z_n\right)\right]$$
$$= \mathbb{E}_{Z_1,\ldots,Z_n}\left[\prod_{i=1}^{d}\Pr\left(A_i+B_i\in I_\epsilon\left(z_i\right), B_i+C_i\in I_\epsilon\left(z_i\right)\,|\,Z_1,\ldots,Z_n\right)\right], \tag{5}$$

where the latter equality holds by independence.

Then,

$$\begin{aligned}
&\Pr\left(A_i+B_i\in I_\epsilon\left(z_i\right), B_i+C_i\in I_\epsilon\left(z_i\right)\,|\,Z_1,\ldots,Z_n\right)\\
&= \mathbb{E}_{B_i}\left[\Pr\left(A_i\in I_\epsilon\left(z_i-B_i\right), C_i\in I_\epsilon\left(z_i-B_i\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\right]\\
&= \mathbb{E}_{B_i}\left[\Pr\left(A_i\in I_\epsilon\left(z_i-B_i\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\Pr\left(C_i\in I_\epsilon\left(z_i-B_i\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\right],
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{B_i}\left[\Pr\left(A_i\in I_\epsilon\left(z_i-B_i\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\Pr\left(C_i\in I_\epsilon\left(z_i-B_i\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\right]\\
&\le \mathbb{E}_{B_i}\left[\Pr\left(A_i\in I_\epsilon\left(0\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\Pr\left(C_i\in I_\epsilon\left(0\right)\,|\,Z_1,\ldots,Z_n,B_i\right)\right]\\
&= \Pr\left(A_i\in I_\epsilon\left(0\right)\,|\,Z_1,\ldots,Z_n\right)\Pr\left(C_i\in I_\epsilon\left(0\right)\,|\,Z_1,\ldots,Z_n\right)\\
&\le \frac{2\epsilon}{\sqrt{\pi\left(\sum_{r=1}^{j}Z_r^2\right)}}\cdot\frac{2\epsilon}{\sqrt{\pi\left(\sum_{r=k+1}^{k+j}Z_r^2\right)}},
\end{aligned} \tag{6}$$

where the latter inequality comes from the fact that, conditioned on $Z_1,\ldots,Z_n$, 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 $\left(1-2\sqrt{c_d}\right)j\le\sum_{i=1}^{j}Z_i^2,\sum_{i=k+1}^{k+j}Z_i^2$ by $\mathcal{E}^{(\downarrow)}$. Then, by Eq. 5 and the law of total probability, we have that

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

Eq. 6 implies that

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

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

$$\mathbb{E}_{Z_1,\ldots,Z_n}\left[\left(\frac{4\epsilon^2}{\pi\sqrt{\left(\sum_{i=1}^{j}Z_i^2\right)\left(\sum_{i=k+1}^{k+j}Z_i^2\right)}}\right)^d\,\middle|\,\mathcal{E}^{(\downarrow)}\right]+\Pr\left(\overline{\mathcal{E}^{(\downarrow)}}\right)$$

$$
\begin{aligned}
&\leq \left(\frac{4\epsilon^2}{\pi j\left(1-2\sqrt{c_d}\right)}\right)^d + \Pr\left(\overline{\mathcal{E}^{(\downarrow)}}\right) \\
&\leq \left(\frac{4\epsilon^2}{\pi j\left(1-2\sqrt{c_d}\right)}\right)^d + 2\exp\left(-c_d j\right) \\
&= \exp\left(-d\log\frac{\pi j\left(1-2\sqrt{c_d}\right)}{4\epsilon^2}\right) + 2\exp\left(-c_d k\right) \\
&\leq 3\left(\frac{4\epsilon^2}{\pi\left(1-2\sqrt{c_d}\right)k}\right)^d.
\end{aligned}
$$

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

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

□

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

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

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

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

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

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

Proof of Theorem 5

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

$$\Pr(T > 0) \geq \frac{\mathbb{E}[T_{n,k}]^2}{\mathbb{E}[T_{n,k}^2]}. \tag{7}$$

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

$$\mathbb{E}\left[T_{n,k}^2\right] = \binom{n}{k}^2 \sum_{j=0}^{k} \Pr(|\mathrm{S} \cap \mathrm{S}'| = k - j) \Pr\left(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}\right) \tag{8}$$

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

$$\mathbb{E}[T_{n,k}] = \sum_{S \in \binom{[n]}{k}} \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_S^{(\vec{z})}}\right] = \sum_{S \in \binom{[n]}{k}} \Pr\left(\mathcal{E}_S^{(\vec{z})}\right) = \binom{n}{k} \Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right). \tag{9}$$

By Eq.s 8 and 9, we have

$$\begin{aligned}
\frac{\mathbb{E}\left[T_{n,k}^2\right]}{\mathbb{E}[T_{n,k}]^2} &= \frac{\binom{n}{k}^2}{\mathbb{E}[T_{n,k}]^2} \sum_{j=0}^{k} \Pr(|\mathrm{S} \cap \mathrm{S}'| = k - j) \Pr\left(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}\right) \\
&= \sum_{j=0}^{k} \Pr(|\mathrm{S} \cap \mathrm{S}'| = k - j) \frac{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)^2}.
\end{aligned} \tag{10}$$

As for the denominator of the second term in Eq. 10, by Lemma 7 we have

$$\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)^2 \geq \frac{1}{256} \left(\frac{4\epsilon^2}{\pi\left(1 + 2\sqrt{c_d} + 2c_d\right) k}\right)^d. \tag{11}$$

As for the numerator of the second term in Eq. 10, Lemma 8 implies that we have

$$\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}\right) \leq 3 \left(\frac{4\epsilon^2}{\pi\left(1 - 2\sqrt{c_d}\right) j}\right)^d. \tag{12}$$

By plugging Eq. 12 and Eq. 11 in Eq. 10, for $j \geq k\left(1 - \frac{1}{d}\right)$ and $d > 1$ we can upper bound the factor $\frac{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})} \cap \mathcal{E}_{S_j}^{(\vec{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)^2}$ of the summation as follows:

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

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

$$768 \left(\frac{\left(1 + 2\sqrt{c_d} + 2c_d\right) k}{\left(1 - 2\sqrt{c_d}\right) j}\right)^d \leq 768 \left(\frac{\left(1 + 2\sqrt{c_d} + 2c_d\right)}{\left(1 - 2\sqrt{c_d}\right)\left(1 - \frac{1}{d}\right)}\right)^d$$

$$\leq 768 \left(\frac{2+7\sqrt{c_d}}{2-7\sqrt{c_d}}\right)^d$$
$$\leq 270801, \tag{13}$$

because $\left(\frac{\left(1+2\sqrt{c_d}+2c_d\right)}{\left(1-2\sqrt{c_d}\right)\left(1-\frac{1}{d}\right)}\right)^d$ is maximized or $d=4$. Let $C'=270801$. If $d>1$, by plugging Eq. 13 in Eq. 10, we have that

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

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

$$\begin{aligned}
&\sum_{j=0}^{\lceil k-\frac{k}{d}\rceil-1} \Pr\left(|\mathrm{S}\cap\mathrm{S}'|=k-j\right)\frac{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\cap\mathcal{E}_{S_j}^{(\vec{z})}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)^2} + C' \\
&\leq \frac{1}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)} \sum_{j=0}^{\lceil k-\frac{k}{d}\rceil-1} \Pr\left(|\mathrm{S}\cap\mathrm{S}'|=k-j\right) + C' \\
&\leq \frac{\Pr\left(|\mathrm{S}\cap\mathrm{S}'|>\frac{k}{d}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)} + C'.
\end{aligned} \tag{14}$$

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

$$\Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k}{d}\right)=\sum_{\tilde{S}\in\binom{[n]}{k}}\Pr\left(\mathrm{S}=\tilde{S}\right)\Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k}{d}\,\middle|\,\mathrm{S}=\tilde{S}\right). \tag{15}$$

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

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

Since $n\geq k^2$, then $\frac{k^2}{n}\leq 1$. Hence, since Chernoff bounds holds for the hypergeometric distribution [Doerr, 2020, Theorem 1.10.25]

$$\begin{aligned}
\Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k}{d}\,\middle|\,\mathrm{S}=\tilde{S}\right) &\geq \Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k^2}{n}+\left(\frac{k}{d}-1\right)\,\middle|\,\mathrm{S}=\tilde{S}\right) \\
&\leq \exp\left(-2\frac{\left(\frac{k}{d}-1\right)^2}{k}\right) \\
&\leq \exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2\right).
\end{aligned} \tag{16}$$

Substituting Eq. 16 in Eq. 15 we get

$$\Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k}{d}\right)\leq\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2\right). \tag{17}$$

We can now keep bounding from above $\frac{\mathbb{E}\left[T_{n,k}^2\right]}{\mathbb{E}[T_{n,k}]^2}$ by plugging Eq. 17 in Eq. 14:

$$\frac{\Pr\left(|\mathrm{S}\cap\mathrm{S}'|\geq\frac{k}{d}\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)}+C'\leq\frac{\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)}+C'. \tag{18}$$

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

$$\begin{aligned}
&\frac{\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2\right)}{\Pr\left(\mathcal{E}_{S_0}^{(\vec{z})}\right)}+C'\\
&\leq\frac{16\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2\right)}{\left(\frac{2\epsilon}{\sqrt{\pi 2k}}\right)^d}+C'\\
&=16\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2+d\log\left(\frac{\sqrt{\pi 2k}}{2\epsilon}\right)\right)+C'.
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
&16\exp\left(-2\frac{k}{d^2}\left(1-\frac{d}{k}\right)^2+d\log\left(\frac{\sqrt{\pi 2k}}{2\epsilon}\right)\right)+C'\\
&\leq C'+16\exp\left(-d\log\frac{d}{\epsilon}\right)\\
&<C'+\frac{16}{e},
\end{aligned} \tag{20}$$

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

Plugging the inverse of the expression in Eq. 20 in Eq. 7 we obtain the thesis.

4.2 Proving SLTH for structured pruning

To prove Theorem 3, 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 keeps small.

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

Definition 10. Given a tensor $X\in\mathbb{R}^{d_1\times\cdots\times d_n}$, the *positive* and *negative* parts of X are respectively defined as $X^+_{\vec{i}}=X_{\vec{i}}\cdot\mathbf{1}_{X_{\vec{i}}>0}$ and $X^-_{\vec{i}}=-X_{\vec{i}}\cdot\mathbf{1}_{X_{\vec{i}}<0}$, where $\vec{i}\in[d_1]\times\cdots\times[d_n]$ 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 $\phi\left(V*X\right)$ 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 $X\in\mathbb{R}^{D\times D\times c}$. If $\tilde{S}\in\{0,1\}^{\mathrm{size}(V)}$ is a $2n$-channel blocked mask, then, for each $(i,j,k)\in[D]\times[D]\times[2nc]$,*

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

Proof. $\tilde{S} \in \{0,1\}^{\text{size}(V)}$ is such that $\tilde{V} = V \odot \tilde{S}$ contains only non-negative edges going from each input channel t to the output channels $(t-1)\,n+1, \ldots, tn$, and only non-positive[3] edges going from each input channel t to the output channels $tn+1, \ldots, 2tn$, 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 $\{n, n+1, \ldots, m\}$ for each pair of integers $n \leq m \in \mathbb{N}$. In formulas, we obtain a tensor $\tilde{V}$ such that, for each $(t, k) \in [c], \times [2nc]$:

$$\left(V \odot \tilde{S}\right)_{1,1,t,k} = \begin{cases} V_{1,1,t,k} \cdot \mathbf{1}_{V_{1,1,t,k}>0} & \text{if } k \in [(2t-2)\,n+1, (2t-1)\,n]\,, \\ V_{1,1,t,k} \cdot \mathbf{1}_{V_{1,1,t,k}<0} & \text{if } k \in [(2t-1)\,n+1, 2tn]\,, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

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

$$\begin{aligned} &\mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} = 1 \text{ iff } k \in [(2t-2)\,n+1, (2t-1)\,n]\,, \text{ and} \\ &\mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]} = 1 \text{ iff } k \in [(2t-1)\,n+1, 2tn]\,. \end{aligned} \tag{22}$$

For each $(i, j, k) \in [D] \times [D] \times [2nc]$, applying Eq. 21 and using Definition 10, it then holds

$$\begin{aligned} &\left(\phi\left(\left(V \odot \tilde{S}\right) * X\right)\right)_{i,j,k} \\ &= \phi\left(\sum_{t=1}^{c_0} \tilde{V}_{1,1,t,k} X_{i,j,t}\right) \\ &= \phi\Bigg(\sum_{t=1}^{c_0} \Big(V_{1,1,t,k} X_{i,j,t} \cdot \mathbf{1}_{V_{1,1,t,k}>0} \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} \\ &\quad + V_{1,1,t,k} X_{i,j,t} \cdot \mathbf{1}_{V_{1,1,t,k}<0} \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\Big)\Bigg) \\ &= \phi\left(\sum_{t=1}^{c_0} \left(V^+_{1,1,t,k} X_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} - V^-_{1,1,t,k} X_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\right)\right) \\ &= \phi\Bigg(\sum_{t=1}^{c_0} \Big(V^+_{1,1,t,k} (X^+_{i,j,t} - X^-_{i,j,t}) \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} \\ &\quad + V^-_{1,1,t,k} (X^-_{i,j,t} - X^+_{i,j,t}) \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\Big)\Bigg). \end{aligned} \tag{23}$$

Observe that only one term survives in the summation in Eq. 23, as there exists only one $t \in [c_0]$ such that $k \in [(2t-2)n+1, 2tn]$, say $t^\star$. Moreover, out of the four additive terms in the expression

$$V^+_{1,1,t^\star,k} (X^+_{i,j,t^\star} - X^-_{i,j,t^\star}) \mathbf{1}_{\frac{k}{2n} \in \left(t^\star-1, t^\star-\frac{1}{2}\right]} + V^-_{1,1,t^\star,k} (X^-_{i,j,t^\star} - X^+_{i,j,t^\star}) \mathbf{1}_{\frac{k}{2n} \in \left(t^\star-\frac{1}{2}, t^\star\right]},$$

at most one is non-zero, due to Definition 10. The ReLU cancels out negative ones, implying that Eq. 23 can be rewritten without the ReLU as a sum of only non-negative terms (out of which, at most one is non-zero) as follows

$$\begin{aligned} &\phi\Bigg(\sum_{t=1}^{c_0} \Big(V^+_{1,1,t,k} (X^+_{i,j,t} - X^-_{i,j,t}) \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} \\ &\quad + V^-_{1,1,t,k} (X^-_{i,j,t} - X^+_{i,j,t}) \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\Big)\Bigg) \\ &= \sum_{t=1}^{c_0} \left(V^+_{1,1,t,k} X^+_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} + V^-_{1,1,t,k} X^-_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\right). \end{aligned} \tag{24}$$

Finally, by Eq. 21 and Eq. 22, $\tilde{V}^+_{1,1,t,k} = 0$ if $\frac{k}{2n} \notin \left(t-1, t-\frac{1}{2}\right]$, and $\tilde{V}^-_{1,1,t,k} = 0$ if $\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]$, which means that in Eq. 24 we can ignore the indicator functions and further simplify the expression as

$$\sum_{t=1}^{c_0} \left(V^+_{1,1,t,k} X^+_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-1, t-\frac{1}{2}\right]} + V^-_{1,1,t,k} X^-_{i,j,t} \mathbf{1}_{\frac{k}{2n} \in \left(t-\frac{1}{2}, t\right]}\right)$$

[3] We consider 0 to be both non-negative and non-positive.

$$
\begin{aligned}
&= \sum_{t=1}^{c_0} \left(\tilde{V}^+_{1,1,t,k} X^+_{i,j,t} + \tilde{V}^-_{1,1,t,k} X^-_{i,j,t} \right) \\
&= \left(\sum_{t=1}^{c_0} \tilde{V}^+_{1,1,t,k} X^+_{i,j,t} + \sum_{t=1}^{c_0} \tilde{V}^-_{1,1,t,k} X^-_{i,j,t} \right) \\
&= \left(\tilde{V}^+ * X^+ + \tilde{V}^- * X^- \right)_{i,j,k}.
\end{aligned}
$$

□

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, $\epsilon \in \left(0, \frac{1}{4}\right), M \in \mathbb{R}_{>0}$, and $C \in \mathbb{R}_{>0}$ be a universal constant with*

$$
n \geq C d^{12} c_1^6 \log^3 \frac{d^2 c_1 c_0}{\epsilon}.
$$

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 $1 - \epsilon$, for all $K \in \mathbb{R}^{d \times d \times c_0 \times c_1}$ with $\|K_{:,:,t_0,:}\|_1 \leq 1$ for each $t_0 \in [c_0]$, it is possible to remove filters from $N_0^{(S)}$ to obtain a CNN $\tilde{N}_0^{(S)}$ for which

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

Proof. Adopting the same definitions as in Lemma 11 (and Eq. 21), for each $(r, s, t_1) \in [d] \times [d] \times [c_1]$ we have, by Lemma 11,

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

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,

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

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

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

Observe that $I^{(t_0)}_+$ consists of the strictly positive entries of $\tilde{V}^+_{1,1,t_0,:}$.[4] 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 SM A), we then have

$$\Pr\left(\left| I^{(t_0)}_+ \right| > \frac{n}{3} \right) \geq 1 - \frac{\epsilon}{4}, \tag{25}$$

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 (SM **??**). Thus, for each $t_0 \in [c_0]$ and a sufficiently-large value of C, since the target filter K is such that $\|K_{:,:,t_0,:}\|_1 \leq 1$ and we have $n \geq Cd^{12}c_1^6 \log^3 \frac{d^2c_1c_0}{\epsilon}$, then we can apply an amplified version of Theorem 5 (i.e. Corollary 19 in SM **??** with vectors of dimension d^2c_1) to show that, with probability $1 - \frac{\epsilon}{4c_0}$ 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 $\hat{S}^+ \in \{0,1\}^{\text{size } \tilde{S}}$ 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 $\hat{V}^+ = \tilde{V}^+ \odot \hat{S}^+$. An analogous argument carries on for a binary mask $\hat{S}^-$ and $-\hat{L}^-_{:,:,t_0,:}$.[5] More formally, let

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

Then, by Corollary 19,

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

[4] Notice that excluding zero entries implies conditioning on the event that the entry is not zero. However, such an event has zero probability and thus doesn't impact the analysis.

[5] The 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:

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

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

$$
\begin{aligned}
\Pr\left(\mathcal{E}^{(\text{kernel})}\right) &\geq \Pr\left(\mathcal{E}^{(\text{kernel})} \;\middle|\; \left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \Pr\left(\left|I_+^{(t_0)}\right|, \left|I_-^{(t_0)}\right| > \frac{n}{3}\right) \\
&\geq \left(1 - \frac{\epsilon}{2}\right)^2 \\
&\geq 1 - \epsilon. \qquad (26)
\end{aligned}
$$

To rewrite the latter in terms of the filter K and a mask 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 $\hat{S} \in \{0,1\}^{\text{size }\tilde{S}}$ 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 $\hat{V} = \tilde{V} \odot \hat{S}$. Ineq. 26 implies that, with probability $1-\epsilon$, such $\hat{S}$ exists and hence,

$$
\left\|K - \hat{L}^+\right\|_{\max} + \left\|K + \hat{L}^-\right\|_{\max} \leq \frac{\epsilon}{d^2 c_1 c_0}. \qquad (27)
$$

Let $S = \tilde{S} \odot \hat{S}$: S is a $2n$-channel blocked basks. Furthermore, for such an S, notice that the following holds.

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

where the latter holds by Lemma 11.[6] Then, by the distributive property of the convolution and the triangle inequality,

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

[6]The presence of $\hat{S}$ does not influence the proof of Lemma 11.

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

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

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. 27 (which holds with probability $1-\epsilon$), we get that

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

proving the thesis. □

Remark 13. From the proof of Lemma 12, 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 $m = \sqrt{n/(C_1 \log \frac{1}{\epsilon})}$ for a universal constant C_1. Moreover, the kernel $\hat{V}$ is structured as if pruned by a $2m$-channel-blocked mask.

Proof of Theorem 3

In order to bound the error propagation across layers, we define the layers' outputs

$$\begin{aligned}
&X^{(0)} = X, \\
&X^{(i)} = \phi\left(K^{(i)} * X^{(i-1)}\right) \quad \text{for } 1 \le i \le \ell.
\end{aligned} \tag{28}$$

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

For brevity's sake, given masks $S^{(1)}, \ldots, S^{(2\ell)}$, let us denote

$$\tilde{L}^{(i)} = L^{(i)} \odot S^{(i)}. \tag{29}$$

Since the ReLU function is 1-Lipschitz, for all $X^{(i-1)}$ it holds

$$\begin{aligned}
&\left\|\phi\left(K^{(i)} * X^{(i-1)}\right) - \phi\left(\tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * X^{(i-1)}\right)\right)\right\|_{\max} \\
\le &\left\|K^{(i)} * X^{(i-1)} - \tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * X^{(i-1)}\right)\right\|_{\max}.
\end{aligned} \tag{30}$$

The key step of the proof is that, for each layer i, since $n_i \ge C d^{12} c_i^6 \log^3 \frac{d^2 c_i c_{i-1} \ell}{\epsilon}$ for a suitable constant C, we can apply Lemma 12 to get that, with probability at least $1 - \frac{\epsilon}{2\ell}$, for all $X^{(i-1)} \in \mathbb{R}^{D\times D\times c_0}$ it holds

$$\left\|K^{(i)} * X^{(i-1)} - \tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * X^{(i-1)}\right)\right\|_{\max} < \frac{\epsilon}{2\ell} \cdot \left\|X^{(i-1)}\right\|_{\max}. \tag{31}$$

Hence, combining Eq. 30 and Eq. 31 we get that, with probability at least $1 - \frac{\epsilon}{2\ell}$, for all $X^{(i-1)} \in \mathbb{R}^{D\times D\times c_0}$,

$$\left\|\phi\left(K^{(i)} * X^{(i-1)}\right) - \phi\left(\tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * X^{(i-1)}\right)\right)\right\|_{\max}$$

$$< \frac{\epsilon}{2\ell} \cdot \left\| X^{(i-1)} \right\|_{\max}. \tag{32}$$

By a union bound, we get that Eq. 32 holds for all layers with probability at least $1 - \epsilon$.

Analogously, we can define the pruned layers' outputs

$$\begin{aligned} \tilde{X}^{(0)} &= X, \\ \tilde{X}^{(i)} &= \phi\left(\tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * \tilde{X}^{(i-1)}\right)\right) \quad \text{for } 1 \leq i \leq \ell. \end{aligned} \tag{33}$$

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

By the same reasoning employed to derive Eq. 31 and Eq. 32 we have that, with probability $1 - \epsilon$, the output of all pruned layers satisfies

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

Moreover, for each $1 \leq i \leq \ell - 1$, by the triangle inequality and by Eq. 34,

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

By the Lipschitz property of ϕ and Lemma 20

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

By unrolling the recurrence, we get that, with probability $1 - \epsilon$,

$$\left\| \tilde{X}^{(i)} \right\|_{\max} \leq \left\| \tilde{X}^{(0)} \right\|_{\max} \left(1 + \frac{\epsilon}{2\ell}\right)^i. \tag{35}$$

Thus, combining Eq. 34 and Eq. 35, with probability $1 - \epsilon$ we get that, for each $i \in [\ell]$,

$$\begin{aligned} &\left\| K^{(i)} * \tilde{X}^{(i-1)} - \tilde{L}^{(2i)} * \phi\left(\tilde{L}^{(2i-1)} * \tilde{X}^{(i-1)}\right) \right\|_{\max} \\ &< \frac{\epsilon}{2\ell} \cdot \left(1 + \frac{\epsilon}{2\ell}\right)^{i-1} \left\| \tilde{X}^{(0)} \right\|_{\max}. \end{aligned} \tag{36}$$

We then see that with probability $1 - \epsilon$, for $1 \leq i \leq \ell$ and all $X \in [-1, 1]^{D \times D \times c_0}$, by Eq. 28 and Eq. 33, and by the triangle inequality,

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

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

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

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

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

Now, we first apply Eq. 36 and then we unroll the recurrence for all layers (as, with probability $1-\epsilon$, Eq. 36 holds for all layers), obtaining

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

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

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

Hence, with probability $1-\epsilon$, for all $X \in [-1,1]^{D \times D \times c_0}$, for all $\ell \in [c]$ it holds that

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

yielding the thesis.

5 Limitations and future work

In previous works [da Cunha et al., 2022b, Burkholz, 2022a] the assumption that the kernel of every second layer has shape $1 \times 1 \times \ldots$ 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 Burkholz [2022a,b] 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 3.

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 $x = \phi(x) - \phi(-x)$. Burkholz [2022a] leverages 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 Burkholz [2022a] to our setting is not straightforward.

Finally, we remark that the assumption of normally distributed weights might be relaxed. Borst et al. [2022] provided an MRSSP result for independent random variables whose distribution converges "fast enough" to a Gaussian one.[7] We believe our arguments can serve well as baselines to generalise our results to support random weights distributed as such.

References

Russell Reed. Pruning algorithms-a survey. *IEEE Trans. Neural Networks*, 4(5):740–747, 1993. doi: 10.1109/72.248452. URL https://doi.org/10.1109/72.248452. 1

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. URL https://proceedings.mlsys.org/book/296.pdf. 1

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. URL http://jmlr.org/papers/v22/21-0366.html. 1, 3

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. URL https://openreview.net/forum?id=rJl-b3RcF7. 1

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. URL https://proceedings.neurips.cc/paper/2019/hash/1113d7a76ffceca1bb350bfe145467c6-Abstract.html. 1

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. doi: 10.1109/CVPR42600.2020.01191. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Ramanujan_Whats_Hidden_in_a_Randomly_Weighted_Neural_Network_CVPR_2020_paper.html. 1

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,*

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

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. URL https://aaai.org/ojs/index.php/AAAI/article/view/6910. 1

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. URL https://proceedings.neurips.cc/paper/2020/hash/1b742ae215adf18b75449c6e272fd92d-Abstract.html. 1, 3

Udo W. Pooch and Al Nieder. A survey of indexing techniques for sparse matrices. *ACM Comput. Surv.*, 5(2):109–133, 1973. doi: 10.1145/356616.356618. URL https://doi.org/10.1145/356616.356618. 2

Adam Polyak and Lior Wolf. Channel-level acceleration of deep face representations. *IEEE Access*, 3:2163–2175, 2015. doi: 10.1109/ACCESS.2015.2494536. URL https://doi.org/10.1109/ACCESS.2015.2494536. 2

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. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html. 2

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. URL https://proceedings.neurips.cc/paper/2016/hash/6e7d2da6d3953058db75714ac400b584-Abstract.html. 2

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. doi: 10.1145/3005348. URL https://doi.org/10.1145/3005348. 2

Arlene Elizabeth Siswanto. *Block sparsity and weight initialization in neural network pruning*. PhD thesis, Massachusetts Institute of Technology, 2021. 2

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. 2, 3

Arthur da Cunha, Francesco d'Amore, Frédéric Giroire, Hicham Lesfari, Emanuele Natale, and Laurent Viennot. Revisiting the Random Subset Sum problem. Research Report, Inria Sophia Antipolis - Méditerranée, Université Côte d'Azur ; Inria Paris, April 2022a. URL https://hal.archives-ouvertes.fr/hal-03654720. 2, 3

Sander Borst, Daniel Dadush, Sophie Huiberts, and Samarth Tiwari. On the integrality gap of binary integer programs with Gaussian data. *Mathematical Programming*, 2022. doi: 10.1007/s10107-022-01828-1. 3, 5, 20

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 2022a. 3

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. 3

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. 3

Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=Vjki79-619-. 3, 19

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, 2022a. URL https://proceedings.mlr.press/v162/burkholz22a.html. 3, 19, 20

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., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/76bf7786d311217077bc8bb021946cd9-Paper-Conference.pdf. 3, 20

Jonas Fischer and Rebekka Burkholz. Towards strong pruning for lottery tickets with non-zero biases. *CoRR*, abs/2110.11150, 2021. URL https://arxiv.org/abs/2110.11150. 3

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. doi: 10.48550/arXiv.2206.04270. URL https://doi.org/10.48550/arXiv.2206.04270. 3

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. URL https://openreview.net/forum?id=U_mat0b9iv. 3

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. URL https://proceedings.mlr.press/v151/sreenivasan22a.html. 3

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. 3

Michael C Mozer and Paul Smolensky. Using relevance to reduce network size automatically. *Connection Science*, 1(1):3–16, 1989. 3

Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *CoRR*, abs/2303.00566, 2023. doi: 10.48550/arXiv.2303.00566. URL https://doi.org/10.48550/arXiv.2303.00566. 3

Luca Becchetti, Arthur Carvalho Walraven da Cuhna, Andrea Clementi, Francesco d'Amore, Hicham Lesfari, Emanuele Natale, and Luca Trevisan. On the Multidimensional Random Subset Sum Problem. Research 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, July 2022b. URL https://hal.archives-ouvertes.fr/hal-03738204. 3

Benjamin Doerr. Probabilistic Tools for the Analysis of Randomized Optimization Heuristics. arXiv:1801.06733, 2020. 11

Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. ISBN 978-0-521-88427-3. URL http://www.cambridge.org/gb/knowledge/isbn/item2327542/. 24

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395. 24

Supplementary material

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 $z, \epsilon \in \mathbb{R}$*

$$\Pr\left(X \in [z-\epsilon, z+\epsilon]\right) \leq \Pr\left(X \in [-\epsilon, \epsilon]\right).$$

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

$$\Pr\left(X \in [-\epsilon, \epsilon]\right) - \Pr\left(X \in [z-\epsilon, z+\epsilon]\right) = \int_{-\epsilon}^{\epsilon} \varphi(x)\,\mathrm{d}x - \int_{z-\epsilon}^{z+\epsilon} \varphi(x)\,\mathrm{d}x.$$

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

$$\begin{aligned}\int_{-\epsilon}^{\epsilon} \varphi(x)\,\mathrm{d}x - \int_{z-\epsilon}^{z+\epsilon} \varphi(x)\,\mathrm{d}x &= \int_{-\epsilon}^{z-\epsilon} \varphi(x)\,\mathrm{d}x - \int_{\epsilon}^{z+\epsilon} \varphi(x)\,\mathrm{d}x \\ &= \int_{-\epsilon}^{z-\epsilon} \varphi(x) - \varphi(x+2\epsilon)\,\mathrm{d}x\end{aligned}$$

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

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

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

Lemma 16 (Chernoff-Hoeffding bounds Dubhashi and Panconesi [2009])**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $\Pr\left(0 \leq X_i \leq 1\right) = 1$ 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 $\Pr\left(X \geq (1+\delta)\mu_+\right) \leq \exp\left(-\frac{\delta^2\mu_+}{3}\right)$;*
2. *if $0 \leq \mu_- \leq \mu$, then $\Pr\left(X \leq (1-\delta)\mu_-\right) \leq \exp\left(-\frac{\delta^2\mu_+}{2}\right)$.*

Lemma 17 (Corollary of [Laurent and Massart, 2000, Lemma 1])**.** *Let $X \sim \chi^2_d$ be a chi-squared random variable with d degrees of freedom. For any $t > 0$, it holds that*

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

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 $\tilde{Z}_1, \ldots, \tilde{Z}_n$ are identically distributed random variables following a standard normal distribution, $\tilde{Z}$ is a half-normal distribution,[8] and $\tilde{Z}, \tilde{Z}_1, \ldots, \tilde{Z}_n$ are independent, then Y follows an NSN distribution.*

Proof. By Definition 4, Y is NSN if, for each $i \in [d]$, $Y_i = Z \cdot Z_i$ where $Z, Z_1, \ldots, Z_n$ are i.i.d. random variables following a standard normal distribution. If $\tilde{Z} = |Z|$, we can rewrite $\tilde{Z}_i = \operatorname{sign}\left(Z\right)\operatorname{sign}\left(Z_i\right)|Z_i|$ for each $i = 1, \ldots, n$, where $Z, Z_1, \ldots, Z_n$ are i.i.d. standard normal

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

random variables, as $\operatorname{sign}(Z)\operatorname{sign}(Z_i)$ is independent of $\operatorname{sign}(Z)$ and of $\operatorname{sign}(Z)\operatorname{sign}(Z_j)$ for $i \neq j$. Then,

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

implying the thesis.

□

Corollary 19 (of Theorem 5)**.** *Let d, k, and n be positive integers with $n \geq C_1 k^2 \log\left(\frac{1}{\epsilon}\right)$ and $k \geq C_2 d^3 \log\frac{d}{\epsilon}$ for some universal constants $C_1, C_2 \in \mathbb{R}_{>0}$. Let $X_1, \ldots, X_n$ be d-dimensional i.i.d. NSN random vectors. For any $0 < \epsilon \leq \frac{1}{4}$ and $\vec{z} \in \mathbb{R}^d$ with $\|\vec{z}\|_1 \leq \sqrt{k}$ it holds*

$$\Pr\left(\exists S : |S| = k,\ \left\|\left(\sum_{i\in S} X_i\right) - \vec{z}\right\|_{\max} \leq \epsilon\right) \geq 1 - \epsilon.$$

Proof. Let $s = \left\lceil C_1 \log\left(\frac{1}{\epsilon}\right)\right\rceil$ and let us partition the n vectors $X_1, \ldots, X_n$ in s disjoint sets $G_1, \ldots, G_s$ 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 ($i \in [s]$)

$$\Pr\left(\exists S \subset G_i : |S| = k,\ \left\|\left(\sum_{i\in S} X_i\right) - \vec{z}\right\|_{\max} \leq \epsilon\right) \geq c. \tag{37}$$

It follows that

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

where the latter inequality comes from Eq. 37 and the independence of the variables across different G_i. By choosing C_1 large enough,

$$1 - (1-c)^{\left\lceil C_1 \log\left(\frac{1}{\epsilon}\right)\right\rceil} \geq 1 - \epsilon.$$

□

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

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

$$\leq \max_{i,j\in[D],\ell\in[c_1]} \|K\|_1 \cdot \|X\|_{\max}$$
$$= \|K\|_1 \cdot \|X\|_{\max}.$$

□