On the Sparsity of the Strong Lottery Ticket Hypothesis

Emanuele Natale, Davide Ferre', Giordano Giambartolomei, Frédéric Giroire, Frederik Mallmann-Trenn

On the Sparsity of the Strong Lottery Ticket Hypothesis

Emanuele Natale
Université Côte d'Azur,
CNRS, Inria, I3S, France

Davide Ferré
Université Côte d'Azur,
CNRS, Inria, I3S, France

Giordano Giambartolomei
Department of Informatics,
King's College London

Frédéric Giroire
Université Côte d'Azur, CNRS,
Inria, I3S, France

Frederik Mallmann-Trenn
Department of Informatics,
King's College London

Abstract

Considerable research efforts have recently been made to show that a random neural network N contains subnetworks capable of accurately approximating any given neural network that is sufficiently smaller than N, without any training. This line of research, known as the Strong Lottery Ticket Hypothesis (SLTH), was originally motivated by the weaker Lottery Ticket Hypothesis, which states that a sufficiently large random neural network N contains *sparse* subnetworks that can be trained efficiently to achieve performance comparable to that of training the entire network N. Despite its original motivation, results on the SLTH have so far not provided any guarantee on the size of subnetworks. Such limitation is due to the nature of the main technical tool leveraged by these results, the Random Subset Sum (RSS) Problem. Informally, the RSS Problem asks how large a random i.i.d. sample Ω should be so that we are able to approximate any number in $[-1, 1]$, up to an error of ε, as the sum of a suitable subset of Ω.

We provide the first proof of the SLTH in classical settings, such as dense and equivariant networks, with guarantees on the sparsity of the subnetworks. Central to our results, is the proof of an essentially tight bound on the Random Fixed-Size Subset Sum Problem (RFSS), a variant of the RSS Problem in which we only ask for subsets of a given size, which is of independent interest.

1 Introduction

The Lottery Ticket Hypothesis (LTH) is a research direction that has attracted considerable attention over the years, stemming from the empirical contrast between the fact that, while large neural networks can be successfully trained to achieve good performance on a given task and successively pruned to a great level of sparsity without compromising their performance, researchers have struggled to train sparse neural networks from scratch. The authors of [13] observed that, using a simple pruning strategy (namely Iterative Magnitude Pruning while rewinding the original weights of the remaining edges to their value at initialization), *starting from a sufficiently large random neural networks, it is possible to identify sparse subnetworks that can be trained to achieve the performance achievable by the starting network* (see Figure 2 in the appendix for an illustration). The previous statement, namely the LTH, soon gave rise to an even stronger one, corroborated by empirical works [30, 27] which proposed "training-by-pruning" algorithms (see Section 2 for details), providing evidence that *starting from a sufficiently large random neural networks, it is possible to identify sparse subnetworks that exhibit good performance as they are, without changing the original weights* (see Figure 3 in the appendix for an illustration). By removing the need to analyze the dynamics of

Preprint.

training, the last statement, namely the Strong Lottery Ticket Hypothesis (SLTH), allowed a fruitful series of rigorous proofs for increasingly more general architectures (see Section 2 for an overview). Such rigorous results can informally be stated as follows:

Theorem 1 (Informal statement of previous SLTH results)**.** *With high probability, a random artificial neural network N_Ω with m parameters can be pruned so that the resulting subnetwork N_S ε-approximates (i.e., approximates up to an error ε) any target artificial neural network N_t with $O\left(m/\log_2(1/\varepsilon)\right)$ parameters.*

It is important to note that, to this day, we only have proofs on the existence of such subnetworks, also called winning tickets, but it remains an open question how to find them reliably.

All theoretical results on the SLTH however have so far provided no guarantee on the number of parameters of the winning ticket N_S. This is in contrast to the original motivation of the LTH and to the practical application of the aforementioned training-by-pruning algorithms that motivated the SLTH, such as [16, 15]. In fact, to approximate target networks with $O\left(m/\log_2(1/\varepsilon)\right)$ parameters, essentially all winning tickets N_S have $\Theta(m)$ parameters (see Appendix A), thus being roughly of the same size of the original network N_Ω. We thus ask the following natural question:

> If we want to ε-approximate a family of target artificial neural networks with m_t parameters by pruning a fraction α, called sparsity, of the m parameters of a random artificial neural network N_Ω, how big should m be?

We are particularly interested in the regime in which the density parameter $\gamma = 1 - \alpha$ vanishes as the size of the network increases, so that the size of the winning ticket N_S is $\gamma m = o(m)$.

The above question has so far remained unanswered as a consequence of the limitation inherited from the core technical tool that has been leveraged so far to prove SLTH results, namely the Random Subset Sum (RSS) Problem [18]. Informally, the RSS asks how large a random i.i.d. sample Ω should be so that we are able to approximate any number in $[-1, 1]$ as the sum of a suitable subset of Ω. The applicability of RSS to the SLTH was first recognized by [25] within the proof strategy previously developed in [21].

1.1 Our Contribution

We answer the aforementioned question by introducing and proving a refined variant of the RSS Problem, namely the Random Fixed-Size Subset Sum Problem (RFSS), in which the approximation of the target values should be achieved by only considering subsets of fixed size k from a set of n samples (Theorem 2). We focus on subsets of fixed size k rather than subsets of size up to k for two main reasons. From a theoretical point of view, it is a stronger requirement, and practically speaking, using fixed-size subsets enables us to achieve SLTH results where the layers of the lottery ticket exhibit a uniform structure, potentially offering a computational advantage in their implementation.

In Section 4, we show how the density γ impacts the *overparameterization*, i.e., the ratio (m/m_t) between the number of parameters of the original network N_Ω and that of the class of target networks N_t that can be ε-approximated by pruning N_Ω down to a subnetwork N_S with γm parameters. In our analysis, we also compare and recover as special cases previous SLTH results such as [25, 21, 7, 3, 10]. For instance, when $\gamma m = \Theta(m)$, we recover up to a logarithmic factor the result of [25], which states that the overparameterization needed is $O(\log_2\left(m_t^2/\varepsilon^2\right))$. In the case of Dense Neural Networks, Theorem 3 thus bridges the gap between the two extreme cases of $\gamma m = \Theta(m_t)$ and $\gamma m = \Theta(m)$ considered in [21] and [25], respectively. It is worth noting that [25] is often considered an improvement over [21], as it exponentially reduces the overparameterization, albeit at the cost of a trivial sparsity level. Finally, we prove that our bounds on the overparameterization as a function of the subnetwork sparsity are essentially tight.

Organization of the paper. After reviewing the literature on the SLTH in Section 2, we introduce the Random Fixed-Size Subset Sum Problem in Section 3. In Section 4, we explore some applications of the RFSS Problem to the SLTH, and finally draw our conclusions in Section 5. Some limitations of our work, along with its potential impact, are discussed in Section 6.

2 Related Work

The SLTH is named after the LTH, which was introduced by Frankle and Carbin in [13]. At the time of writing, this paper has received over 3,300 citations, attesting to the significance and impact of the research topic. Surveying the LTH is thus besides the scope of this work, and we defer the reader to dedicated surveys such as [17].

The SLTH was empirically motivated by work investigating training-by-pruning algorithms such as [30, 27], namely algorithms that leverage the gradient of the network parameters to learn a good *mask* of the edges to be retained (i.e., a good subnetwork, called the winning ticket). [30] achieves this by learning a probability associated to each edge, which is then used to sample the edges that should be included in the subnetwork. [27] gets rid of the stochasticity involved in the aforementioned strategies by learning a score associated to each edge; the subnetwork is then determined by including the edges with the highest score. Such strategies are leveraged in [16, 15] in a federated learning setting, in order to improve the communication cost of distributed training by communicating the sampled masks of a fixed shared network, rather than the entire weights. However, these training-by-pruning algorithms are generally not computationally less expensive than classical training, since they also make use of backpropagation to update scores and are applied to a sufficiently large network to find a winning ticket. To reduce the computational cost of finding a good subnetwork, [14] shows, both theoretically and experimentally, that randomly pre-pruning the source network before looking for a winning ticket can be an effective approach. In [24], on top of randomly pruning the source network, some parameters are also frozen. Frozen parameters are forced to be part of the winning ticket and they do not have an associated score, which effectively reduces the search space for the training-by-pruning algorithms.

The first rigorous proof of the SLTH in the case of dense neural networks has been provided by [21], which establishes a framework that was inherited by the subsequent works. [25] crucially shows that the framework in [21] allows the application of the RSS analysis in [18], proving that, with no constraint on the size of the subnetworks, a random network with m-parameters can be pruned to approximate target networks with $m/\log(1/\varepsilon)$ parameters (we defer the reader to Theorem 3 for details on further constraints on the parameters). An alternative proof of the result in [25] was simultaneously shown in [23]. [8] and [3] successively extended [25] and [23] to convolutional neural networks (CNNs). By leveraging multidimensional generalizations of RSS [9, 2], [6] further extended the SLTH to structured pruning of CNNs and, as a special case, dense networks. Finally, [10] provided a general framework that proves the SLTH for equivariant networks.

As for refinements and generalizations of the above results, [4] shows that, at the cost of a quadratic overhead in the overparameterization w.r.t. [25], the number of layers of the random network N_Ω can be reduced to $\ell + 1$, where ℓ is the number of layers of the target networks N_z; furthermore, while previous results only considered networks with ReLU activation, [4] shows how to extend the proof in [25] to a more general class of activations functions. [5] introduces the notion of universal lottery ticket, and show that it is possible to prune a sufficiently overparameterized random network so that the resulting subnetwork (the lottery ticket) can approximate certain class of functions up to an affine transformation of the output of the subnetwork (in this sense being universal). [12] shows how to extend the proof in [25] when neurons have random biases, and adapts the training-by-pruning algorithm of [27] to find a strong lottery ticket with a desired sparsity level. Motivated by theoretical insights on the existence of sparse strong lottery tickets, [11] develops a framework to plant the latter in large random network and investigates training-by-pruning algorithms, providing evidence that sparse strong lottery tickets typically exists for common machine learning tasks, and the difficulty to find them is of algorithmic nature.

Our proof of the RFSS Problem in Section 3 is based on the second moment method approach first explored by [19], and which has recently been refined to prove multidimensional generalizations of RSS by [9] and [2].

3 Fixed-Size Random Subset Sum

In this section we present our technical contributions on the RFSS, which are the foundation of our proofs regarding the sparsity of the SLTH.

Let us start by introducing some notation. We denote by $[n]$ the set $\{1, \ldots, n\}$, for $n \in \mathbb{N}$. Given a set $\Omega = \{X_1, ..., X_n\}$ and a set of indices $S \subseteq [n]$ we define $\Sigma_S^\Omega = \sum_{i \in S} X_i$, and we omit Ω when clear from the context. We now define a class of distributions for which our RFSS result holds.

Definition 1 (sum-bounded)**.** *We say that a probability density function f is* sum-bounded *if there exist positive constants c_l and c_u such that, for all $k \in \mathbb{N}$, given k independent samples $X_1, ..., X_k$ with density f, the density of their sum $f_{\Sigma_{[k]}}$ satisfies*

$$\frac{c_l}{\sqrt{k}} \leq f_{\Sigma_{[k]}}(x) \leq \frac{c_u}{\sqrt{k}},$$

with the lower bound holding for all $x \in \left[-\sqrt{k}, \sqrt{k}\right]$ and the upper bound holding for all $x \in \mathbb{R}$.

At first, our definition of sum-bounded could look as a weaker version of a classical local limit theorem on the sum of random variables (e.g., see [26, Chapter VII, Theorem 7]). However, that is not the case, since we require a lower bound on the sum for any k, which is needed to prove our main result.

Denote, for all $x \in [0, 1]$, the binary entropy as

$$H_2(x) = -x \log_2 x - (1 - x) \log_2 (1 - x).$$

Our main technical result is the following proof of a fixed-size subset variant of the RSS Problem.

Theorem 2. *Let $0 < \varepsilon < 1$, $c_{hyp} \geq 1$, k, n be integers with $1 \leq k \leq \frac{n}{2}$, and let $\Omega = \{X_1, ..., X_n\}$ where the X_i's are i.i.d. random variables with sum-bounded density. There exists a constant c_{thm} such that, if*

$$n \geq c_{hyp} \frac{\log_2 \frac{k}{\varepsilon}}{H_2\left(\frac{k}{n}\right)}, \tag{1}$$

then for every fixed $z \in \left[-\sqrt{k}, \sqrt{k}\right]$ it holds that

$$\Pr\left(\exists S \subset [n], |S| = k : |\Sigma_S - z| < \varepsilon\right) \geq c_{thm}.$$

Remark. The proof of Theorem 2 is given in Section 3.1, and it actually holds for any $1 \leq k \leq \lambda n$, for an arbitrary $\lambda \in [1/n, 1)$. We state the theorem this way for readability and because we are primarily interested in high-sparsity settings (i.e., small size k of the subsets), so considering values of $k \geq \frac{n}{2}$ does not add much to our analysis. The same remark also holds for Corollary 1.

The sum-bounded condition of Definition 1 is easily verified for distributions such as the Gaussian distribution. Previous SLTH results rely on a classical resampling argument by [18, Corollary 3.3], which shows how RSS results for Uniform$[-1, 1]$ independent random variables naturally extend to independent random variables that *contains* a uniform distribution, in the sense that they can be expressed as the mixture of distributions one of which is Uniform$[-1, 1]$ with constant probability.[1] The next lemma thus proves that the Uniform$[-1, 1]$ distribution is sum-bounded[2]. A detailed proof is provided in Appendix C.

Lemma 1. *The Uniform$[-1, 1]$ probability density function is sum-bounded, i.e., given a set $\mathcal{U}_n = \{U_i\}_{i \in [n]}$ of i.i.d. variables U_i with Uniform$[-1, 1]$ probability density function, there exist constants c_l and c_u such that the probability density function $f(x, n)$ of the sum $\Sigma_{[n]}^{\mathcal{U}_n}$ of these variables, for all $n \in \mathbb{N}$,*

$$\frac{c_l}{\sqrt{n}} \leq f(x, n) \leq \frac{c_u}{\sqrt{n}}, \tag{2}$$

with the lower bound holding for all $x \in [-\sqrt{n}, \sqrt{n}]$, and the upper bound holding for all $x \in \mathbb{R}$.

Finally, in our proofs on the Sparse SLTH in Section 4, we make use of the following corollary of Theorem 2, which ensures a uniform high probability of hitting any target $z \in [-\sqrt{k}, \sqrt{k}]$, considering independent random variables that contain a uniform distribution.

[1] The definition in [18, Corollary 3.3] is actually more general, since it concerns a different problem.

[2] We believe that Lemma 1 is known, but we could not find a reference.

Corollary 1. *Let $0 < p \leq 1$ and $\varepsilon \in (0, 1/2)$ be constants, k, n with $1 \leq k \leq \frac{n}{2}$, and let $\Omega = \{X_1, ..., X_n\}$ be i.i.d. random variables whose density is a mixture of a Uniform$([-1, 1])$ with probability p, and some other density otherwise. There exists a positive constant c_{amp} that only depends on p such that, if*

$$n \geq c_{amp} \frac{\log_2^2 \frac{k}{\varepsilon}}{H_2\left(\frac{k}{n}\right)}, \tag{3}$$

then

$$\Pr\left(\forall z \in \left[-\sqrt{k}, \sqrt{k}\right], \exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon\right) \geq 1 - \varepsilon.$$

Proof Idea. The corollary follows from three arguments. First, by a standard sampling argument, we can assume that a constant fraction of the sample follows a Uniform$[-1, 1]$ distribution. Secondly, by Lemma 1, the uniform probability density function is sum-bounded. We can thus apply Theorem 2, which guarantees a success probability of c_{thm} for approximating a given target. Finally, by a standard probability amplification argument and a union bound applied to Theorem 2, by paying an extra factor $\log_2(k/\varepsilon)$ in Eq. 1, the constant c_{thm} can be assumed to be $1 - \varepsilon$, and the existence of a suitable subset S_z holds simultaneously for all $z \in [-\sqrt{k}, \sqrt{k}]$. Details are given in Appendix D. □

For k big enough, we can get rid of the squared logarithmic dependency on k in the right hand side of Equation 3, as shown in the following Corollary, whose proof can be found in Appendix E.

Corollary 2. *Let $0 < p \leq 1$ and $\varepsilon \in (0, 1/2)$ be constants, k, n be integers with $1 \leq k \leq \frac{n}{2}$ and $k \geq 2c_{amp}\left(\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}\right)$. Let $\Omega = \{X_1, ..., X_n\}$ be i.i.d. random variables whose density is a mixture of a Uniform$([-1, 1])$ with probability p, and some other density otherwise. There exists a positive constant c_{amp} that only depends on p such that, if*

$$n \geq 2c_{amp} \frac{\log_2^2 \frac{1}{\varepsilon}}{H_2\left(\frac{k}{n}\right)}, \tag{4}$$

then

$$\Pr\left(\forall z \in \left[-\sqrt{k}, \sqrt{k}\right], \exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon\right) \geq 1 - \varepsilon.$$

As customary in conference versions of papers, our proofs adopt the convention of taking ceilings and floors as suitable for non integer fractional terms. This is done in the interest of the reader (and ours), and does not impact the results in any significant way.

3.1 Proof of Theorem 2

Proof of Theorem 2. For simplicity, throughout the proof we will often use c to denote any positive constant. Let $\mathcal{S}_k = \{S \subset [n] \,|\, |S| = k\}$ and define, for a fixed $z \in [-\sqrt{k}, \sqrt{k}]$,

$$Y = Y(z) = \sum_{S \in \mathcal{S}_k} Z_S$$

where $Z_S = Z_S(z) = \mathbf{1}_{\{|\Sigma_S - z| < \varepsilon\}}$. Following [19], we exploit the second moment method for RFSS, generalising it to arbitrary k.

$$\Pr(Y > 0) \geq \frac{(\mathbb{E}[Y])^2}{\mathbb{E}[Y^2]}, \tag{5}$$

it thus suffices to prove that

$$\mathbb{E}\left[Y^2\right] \leq c\left(\mathbb{E}[Y]\right)^2. \tag{6}$$

We first rewrite Eq. 5 in a more convenient form. Let $\tilde{S}$ and $\tilde{S}'$ be two independently and uniformly at random chosen subsets of $[n]$ of size k, and denote $H_S(z)$ as the event that Σ_S ε-approximates z, namely

$$H_S = H_S(z) = \{|\Sigma_S - z| < \varepsilon\}.$$

We have

$$\mathbb{E}[Y] = \sum_{S \in \mathcal{S}_k} \mathbb{E}[Z_S] = \sum_{S \in \mathcal{S}_k} \Pr(H_S) = \binom{n}{k} \Pr\left(H_{\tilde{S}}\right) \tag{7}$$

and

$$\begin{aligned}\mathbb{E}[Y^2] &= \mathbb{E}\left[\left(\sum_{S \in \mathcal{S}_k} Z_S\right)\left(\sum_{S' \in \mathcal{S}_k} Z_{S'}\right)\right] = \sum_{S,S' \in \mathcal{S}_k} \mathbb{E}\left[Z_S Z_{S'}\right] \\ &= \sum_{S,S' \in \mathcal{S}_k} \Pr\left(H_S \wedge H_{S'}\right) = \binom{n}{k}^2 \Pr\left(H_{\tilde{S}} \wedge H_{\tilde{S}'}\right). \end{aligned} \tag{8}$$

Using Eqs. 7 and 8 we can rewrite the r.h.s. of Eq. 5 as follows

$$\frac{\left(\mathbb{E}\left[Y\right]\right)^2}{\mathbb{E}\left[Y^2\right]} = \frac{\left[\Pr\left(H_{\tilde{S}}\right)\right]^2}{\Pr\left(H_{\tilde{S}} \wedge H_{\tilde{S}'}\right)} = \frac{\Pr\left(H_{\tilde{S}}\right)}{\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}\right)}.$$

Eq. 6 thus becomes

$$\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}\right) \leq c \Pr\left(H_{\tilde{S}}\right). \tag{9}$$

Let I_i denote the event $\{|\tilde{S} \cap \tilde{S}'| = i\}$ and $I_{a,b}$ the event $\bigcup_{a \leq i \leq b} I_i$. Fix $\mu \in (\lambda, 1)$. By the law of total probability and independence of I_i and $H_{\tilde{S}}$, we rewrite the l.h.s. of Eq. 9 as follows:

$$\begin{aligned}&\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}\right) \\ &= \Pr\left(H_{\tilde{S}'} \wedge I_k \,|\, H_{\tilde{S}}\right) + \Pr\left(H_{\tilde{S}'} \wedge I_{\mu k,k-1} \,|\, H_{\tilde{S}}\right) + \Pr\left(H_{\tilde{S}'} \wedge I_{0,\mu k-1} \,|\, H_{\tilde{S}}\right) \\ &= \Pr\left(I_k\right) \cdot \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_k\right) && (10) \\ &\quad + \Pr\left(I_{\mu k,k-1}\right) \cdot \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) && (11) \\ &\quad + \sum_{i=0}^{\mu k-1} \left(\Pr\left(I_i\right) \cdot \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_i\right)\right). && (12)\end{aligned}$$

To conclude the proof, it suffices to show that each addendum in Eqs. 10, 11 and 12 are upper-bounded by some constant multiple of $\varepsilon/\sqrt{k}$, since the lower bound in Definition 1 ensures that

$$\frac{\varepsilon}{\sqrt{k}} \leq c \Pr\left(H_{\tilde{S}}\right). \tag{13}$$

As for the first addendum (Eq. 10), since $\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_k\right) = 1$, then

$$\begin{aligned}&\Pr\left(I_k\right) \cdot \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_k\right) = \Pr\left(I_k\right) = \frac{1}{\binom{n}{k}} \overset{(a)}{\leq} \sqrt{\frac{8k(n-k)}{n}} 2^{-nH_2\left(\frac{k}{n}\right)} \\ &\overset{(b)}{\leq} \sqrt{\frac{8k(n-k)}{n}} 2^{-c_{\text{hyp}} \log_2 \frac{k}{\varepsilon}} \overset{(c)}{\leq} 2\sqrt{2}\frac{\varepsilon}{\sqrt{k}}, \end{aligned} \tag{14}$$

where inequality (a) in Eq. 14 is a standard lower bound on $\binom{n}{k}$ holding for all $k \leq n-1$; in inequality (b) in Eq. 14 we used Eq. 1, namely $nH_2\left(\frac{k}{n}\right) \geq c_{\text{hyp}} \log_2 \frac{k}{\varepsilon}$; in inequality (c) in Eq. 14 we used that $c_{\text{hyp}} \geq 1$.

As for the second addendum (Eq. 11), we next show that

$$\Pr\left(I_{\mu k,k-1}\right) \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \leq c\frac{\varepsilon}{\sqrt{k}} \tag{15}$$

by proving that

$$\Pr\left(I_{\mu k,k-1}\right) \leq \frac{c}{\sqrt{k}} \tag{16}$$

and

$$\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \leq c\varepsilon. \tag{17}$$

First, observe that $I = |\tilde{S} \cap \tilde{S}'|$ follows a Hypergeometric(n, k, k) distribution, thus by Chebyshev's inequality

$$\Pr\left(I_{\mu k,k-1}\right) \leq \Pr\left(I \geq \mu k\right) = \Pr\left(I - \frac{k^2}{n} \geq \mu k - \frac{k^2}{n}\right) \leq \frac{\mathrm{Var}\left[I\right]}{\mu^2 k^2 \left(1 - \frac{k}{\mu n}\right)^2}$$

$$\leq c' \frac{\frac{k^2}{n}\frac{n-k}{n}\frac{n-k}{n-1}}{k^2} \leq \frac{c}{\sqrt{k}}, \tag{18}$$

having set $c' = \mu^2(1 - \lambda/\mu)^2 > 0$, thus proving Eq. 16. Secondly, define $A = \tilde{S}' \backslash \tilde{S}$ and observe that

$$\Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \tag{19}$$

$$= \sum_{i=\mu k}^{k-1} \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_i\right) \Pr\left(I_i \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \tag{20}$$

$$= \sum_{i=\mu k}^{k-1} \int_{-\infty}^{\infty} \Pr\left(|\Sigma_A - (z-y)| < \varepsilon \,|\, \Sigma_I = y, I_i, H_{\tilde{S}}\right) \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy$$
$$\cdot \Pr\left(I_i \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \tag{21}$$

$$= \sum_{i=\mu k}^{k-1} \int_{-\infty}^{\infty} \Pr\left(|\Sigma_A - (z-y)| < \varepsilon \,|\, \Sigma_I = y, I_i\right) \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy$$
$$\cdot \Pr\left(I_i \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \tag{22}$$

$$\leq c\varepsilon \sum_{i=\mu k}^{k-1} \int_{-\infty}^{\infty} \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy \Pr\left(I_i \,|\, H_{\tilde{S}}, I_{\mu k,k-1}\right) \leq c\varepsilon, \tag{23}$$

where from Eq. 19 to Eq. 20 and from Eq. 20 to Eq. 21 we used the law of total probability;[3] from Eq. 21 to Eq. 22 we dropped the redundant event $H_{\tilde{S}}$ in the conditioning, due to conditional independence; finally, from Eq. 22 to Eq. 23 we used Definition 1 which implies that for any $i \in \{\mu k, ..., k-1\}$ it holds

$$\Pr\left(|\Sigma_A - (z-y)| < \varepsilon \,|\, \Sigma_I = y, I_i\right) = \Pr\left(\left|\Sigma_{[k-i]} - (z-y)\right| < \varepsilon\right) \leq c\varepsilon.$$

This concludes the proof of Eq. 17.

As for the third addendum (Eq. 12), analogously to the calculations from Eq. 20 to Eq. 22, by the law of total probability we have

$$\sum_{i=0}^{\mu k-1} \Pr\left(I_i\right) \cdot \Pr\left(H_{\tilde{S}'} \,|\, H_{\tilde{S}}, I_i\right)$$

$$= \sum_{i=0}^{\mu k-1} \Pr\left(I_i\right) \cdot \int_{-\infty}^{\infty} \Pr\left(|\Sigma_A - (z-y)| < \varepsilon \,|\, \Sigma_I = y,\, I_i\right) \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy$$

$$= \sum_{i=0}^{\mu k-1} \Pr\left(I_i\right) \cdot \int_{-\infty}^{\infty} \Pr\left(\left|\Sigma_{[k-i]} - (z-y)\right| < \varepsilon\right) \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy \tag{24}$$

$$\leq c\frac{\varepsilon}{\sqrt{k}} \sum_{i=0}^{\mu k-1} \Pr\left(I_i\right) \cdot \int_{-\infty}^{\infty} \Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right) dy \tag{25}$$

$$\leq c\frac{\varepsilon}{\sqrt{k}}, \tag{26}$$

where from Eq. 24 to Eq. 25 we used Definition 1, which implies that for any $i \in \left\{0, ..., \frac{9}{10}k - 1\right\}$ it holds

$$\Pr\left(\left|\Sigma_{[k-i]} - (z-y)\right| < \varepsilon\right) \leq c'\frac{\varepsilon}{\sqrt{k-i}} \leq c\frac{\varepsilon}{\sqrt{k}}.$$

[3]For simplicity, we denote the density of Σ_I conditional on $H_{\tilde{S}} \cap I_i$ as $\Pr\left(\Sigma_I = y \,|\, H_{\tilde{S}}, I_i\right)$.

The three bounds on the addenda in Eqs. 10, 11 and 12 (respectively Eqs. 14, 15 and 26), combined with Eq. 13, conclude the proof. □

4 Sparse Strong Lottery Ticket Hypothesis (SSLTH)

We now apply our results on the RFSS problem to the SLTH and obtain guarantees on the sparsity of winning tickets for Dense Neural Networks (DNNs, Theorem 3) and Equivariant NNs (Theorem 4).

The next theorem essentially interpolates between the two extremes of [21][Theorem 2.1] (where $\gamma m = \Theta(m_t)$) and [25][Theorem 1] (where $\gamma m = \Theta(m)$), where we recall that m and m_t represent the number of parameters of the overparameterized and the target networks, respectively, and γ is the density of the winning ticket.

We use $\sigma(\cdot)$ to denote the ReLU activation function, i.e., $\sigma(x) = x \cdot \mathbf{1}_{x \geq 0}$, and $\|\mathbf{W}\|$ to denote the spectral norm of the matrix $\mathbf{W}$. Let $\mathcal{F}$ to be a set of target ReLU neural networks $f : \mathbf{R}^{d_0} \to \mathbf{R}^{d_l}$ of depth l such that

$$\mathcal{F} = \{f : f(\mathbf{x}) = \mathbf{W}_l \sigma(\mathbf{W}_{l-1} \dots \sigma(\mathbf{W}_1 \mathbf{x})),\ \forall i\ \ \mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}} \text{ and } \|\mathbf{W}_i\| \leq 1\} \tag{27}$$

For a given $f \in \mathcal{F}$, for all $i \in [\ell]$, let $\rho_i = \max\{{}^{d_{i-1}}/_{d_i}, {}^{d_i}/_{d_{i-1}}\}$, and $\rho = \max_i \rho_i$. Then, recalling that c_{amp} is the constant defined in Corollary 1, we have the following result.

Theorem 3 (SSLTH for DNNs)**.** *Let g be a randomly initialized feed-forward 2ℓ-layer neural network, in which each weight is drawn from a Uniform$[-1, 1]$ distribution, of the following form:*

$$g(\mathbf{x}) = \mathbf{M}_{2l} \sigma(\mathbf{M}_{2l-1} \dots \sigma(\mathbf{M}_1 \mathbf{x})).$$

Let $\gamma' = \gamma'(\varepsilon) \in (0, 1)$, $\mathbf{M}_{2i} \in \mathbb{R}^{d_i \times 2d_{i-1} n_i^*}$ and $\mathbf{M}_{2i-1} \in \mathbb{R}^{2d_{i-1} n_i^* \times d_{i-1}}$, with n_i^* satisfying

$$n_i^* = c_{amp} \frac{\log_2^2\left(\frac{2\ell d_{i-1} d_i \gamma' n_i^*}{\varepsilon}\right)}{H_2(\gamma')}. \tag{28}$$

With probability at least $1 - \varepsilon$, for every $f \in \mathcal{F}$, where $\mathcal{F}$ is defined as in Eq. 27, g can be pruned to obtain a subnetwork of sparsity at least $\alpha = 1 - \gamma$ that approximates f up to an error ε, having defined $\gamma = \rho\gamma'$.

Proof Idea. The theorem follows from a slight variation of the same approach detailed in [25], in which we use our Corollary 1 instead of [18][Corollary 2.5] when pruning g, allowing us to have control over the size of the pruned network. A detailed proof is provided in Appendix F. □

To illustrate a simple example of how Theorem 3 addresses the main question asked in the introduction, consider the case where we want to approximate a target network with m_t parameters and ℓ layers, each of width d (so $\rho = 1$ and $\gamma' = \gamma$), by pruning an overparameterized network to achieve a desired sparsity level of $\alpha = 1 - \gamma$. The condition expressed by Equation 28 in Theorem 3 comes from the use of Corollary 1 when pruning network g, as shown in the proof. If, instead of Corollary 1, we use its simplified variant Corollary 2, it is easy to observe that Equation 28 would become

$$n_i^* = c_{\text{amp}} \frac{\log_2^2\left(\frac{2\ell d_{i-1} d_i}{\varepsilon}\right)}{H_2(\gamma')}. \tag{29}$$

Using this condition, Theorem 3 then tells us that we need to prune a randomly initialized network with twice as many layers and a number of parameters of the order of $d^2 \frac{log^2 \frac{\ell d^2}{\varepsilon}}{H(\gamma)}$.

We will now clarify the connection between Theorem 3 and the earlier results from [21] and [25]. Figure 1 provides a quick visual comparison.

Malach et al.[21]. When all layers have the same width d, [21] showed that any target network with l layers and a total of $m_t = d^2 l$ parameters can be ε-approximated by pruning a randomly initialized network with $2l$ layers. The overparameterization of this network, relative to the target network, is $O\left(\frac{m_t^2}{\varepsilon^2} \log_2 \frac{m_t}{\varepsilon}\right) = \tilde{O}\left(\frac{m_t^2}{\varepsilon^2}\right)$. More specifically, the winning ticket found after pruning

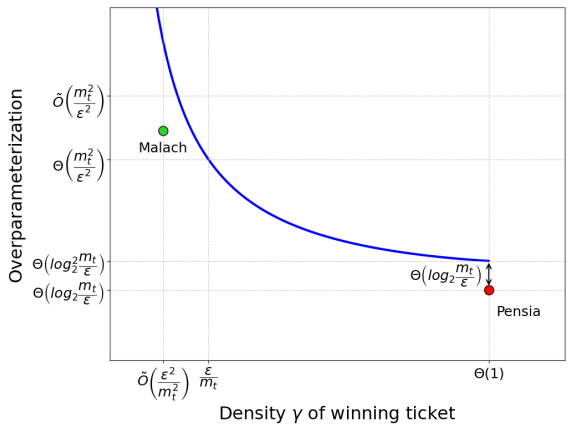

Figure 1: A qualitative plot showing the relationship between the density γ of a winning ticket and the overparameterization required by Theorem 3 for a target network with m_t parameters. Earlier results from Pensia et al.[25] and Malach et al.[21] are shown for comparison.

has a parameter count of the same order as the target network, resulting in a density of $\gamma = \tilde{O}\left(\frac{\varepsilon^2}{m_t^2}\right)$. Notably, this density γ is the inverse of the overparameterization, as the size of the winning ticket matches that of the target network.

Next, we show that Theorem 3 also yields a density that is polynomial in $\frac{\varepsilon}{m_t}$, when using an overparametrization of $\Theta\left(\frac{m_t^2}{\varepsilon^2}\right)$. Let $z = \left(\frac{m_t}{\varepsilon}\right)$, and note that $\gamma' = \gamma$ in Theorem 3, since all layers have the same width. As n_i^* in Theorem 3 represents the overparametrization with respect to the target network, let us set $n_i^* = cz^2$, for some constant c. Equation 28 then becomes

$$cz^2 \geq c_{\text{amp}} \frac{\log_2^2(cz^3\gamma)}{H(\gamma)} \tag{30}$$

We show that the inequality $cz^2 \geq c_{\text{amp}} \frac{\log_2^2(cz^3\gamma)}{\gamma \log_2(1/\gamma)}$ holds for some big enough constant c when setting $\gamma = \frac{\varepsilon}{m_t} = \frac{1}{z}$, which implies that Equation 30 is also satisfied. We get $cz \geq c_{\text{amp}} \frac{\log_2^2(cz^2)}{\log_2(z)}$, which is satisfied for a big enough constant c (see Appendix H). Overall, when using an overparametrization $\Theta\left(\frac{m_t^2}{\varepsilon^2}\right)$, we find a winning ticket with density $\frac{\varepsilon}{m_t}$, as shown in Figure 1.

Pensia et al.[25]. For simplicity, let us still consider target networks where all layers have the same width d, and we apply Theorem 3 using the simplified condition from Equation 29. When $\gamma m = \Theta(m)$, i.e. the density γ is a constant as in [25] (see Appendix A), the entropy term $H_2(\gamma')$ in the right-side of Equation 29 also becomes a constant. In this setting, we indeed recover the result shown in [25][Theorem 1], up to a logarithmic factor, as shown in Figure 1.

Quite similarly to Theorem 3, the next result essentially generalizes [10] up to a factor $\log_2 \frac{1}{\varepsilon}$. The theorem is stated with the understanding that for G-equivariant networks, in order to preserve G-equivariance, pruning is best done not with respect to the parameters expressing the network in the canonical basis (i.e. directly on the weights of the network), but with respect to the *equivariant parameters*, that is those coefficients expressing the linear layers of the network as a linear combination of the elements of the corresponding equivariant basis [10]. For simplicity, due to the technical set-up, we assume all feature spaces being $\mathbb{F} = (\mathbb{R}^d, \sigma)$, with σ the linear representation of the group G, and the same number n of such feature spaces being stacked in each layer. A G-equivariant linear map from the ith feature space to the $i+1$st can be decomposed in a corresponding equivariant basis denoted $\mathcal{B}_{i\to i+1} = \mathcal{B}$. Since all feature spaces are the same, we omit the layers' indices. When stacking n feature spaces in the input and output of the ith layer, the full equivariant basis is denoted $k_{n\to n}$, and finally the basis of the G-equivariant maps from $\mathbb{F}^n$ to $\mathbb{F}^n$ can be written as the Kronecker product $k_{n\to n} \otimes \mathcal{B}$. For any basis $\mathcal{B} = \{b_1, \ldots, b_p\}$, we denote its cardinality $p = |\mathcal{B}|$

and define $\|\mathcal{B}\| = \max_{\|\beta\|_\infty} \| \sum_{k=1}^{p} \beta_k b_k \|$, with $\|\cdot\|$ in the r.h.s. being the operator norm inherited from the ℓ_p norm.

Theorem 4 (SSLTH for Equivariant Networks)**.** *Let h be a random 2ℓ-layer G-equivariant network where all equivariant parameters are drawn from a Uniform$[-1,1]$ distribution, every odd layer expressed in the associated equivariant basis $k_{\tilde{n}\to n} \otimes \mathcal{B}$ and every even layer expressed in the associated equivariant basis $k_{n\to\tilde{n}} \otimes \mathcal{B}$. Let $\gamma = \gamma(\varepsilon) \in (0,1)$, with $\tilde{n}$ satisfying*

$$\tilde{n} = c_{amp} \frac{\log_2^2 \left(\frac{2\ell n^2 \max\{|\mathcal{B}|, \|\mathcal{B}\|\} \gamma \tilde{n}}{\varepsilon} \right)}{H_2(\gamma)}.$$

With probability at least $1-\varepsilon$, for every ℓ-layer G-equivariant neural network f, with all layers expressed in the associated equivariant basis $k_{n\to n} \otimes \mathcal{B}$, h can be pruned to obtain a G-equivariant subnetwork of sparsity at least $\alpha = 1-\gamma$ that approximates f up to an error ε.

The proof, which we omit, is analogous to that of Theorem 3, since [10][Theorem 1] exploits the exact same pruning strategy of [25], except for the fact that it is applied not to the original parameters of the equivariant network, but to the network expressed in terms of its equivariant basis (the sparsity α is here also intended with respect to the equivariant parameters count). This allows the construction to apply without losing the property of equivariance in the pruned approximating subnetwork obtained. The crucial step is when Corollary 1 is applied in [10][Lemma 1], instead of [18][Corollary 2.5]. This is done in parallel, multiple times, across non-overlapping coefficients of the equivariant basis. Thanks to the careful preprocessing devised by the authors, this preserves equivariance and at the same time ensures that each application of Corollary 1 is independent of the others.

To conclude the section, we mention that Theorem 4 applies in particular to vanilla CNNs, which are a special case of equivariant neural networks where the group is $G = (\mathbb{Z}^2, +)$, recovering previous SLTH results on CNN [7, 3]. Furthermore, we remark that Theorem 3 can be revisited through the improvement upon the 2ℓ-depth overparameterization devised in [4], i.e., it is possible to provide sparsity guarantees also for overparameterizations requiring depth $\ell+1$ only. The analysis is more technical and we omit it, but the ideas are analogous to what shown in [4]. An analogous improvement is suggested as future work in [10].

$$\mathcal{F} := \{h_W : W \in \mathbb{R}^{d\times d}, \|W\| \leq \sqrt{k}\}, \quad \text{where} \quad h_W(x) = Wx. \tag{31}$$

The formal claim states that, if a network with n parameters can approximate every $h_W \in \mathcal{F}$ with probability at least $1/2$ (after it is pruned down to k parameters), then the hypothesis of Theorem 2 in Eq. 1 must hold.[4]

Theorem 5. *Let $n, k \in \mathbb{N}$, with $1 \leq k \leq \lambda n$, having set $\lambda = 1 - 1/2\pi \approx 0.84$. Consider a neural network g with n parameters, and let $\mathcal{G}_k$ be the set of neural networks that can be formed by pruning g down to k parameters. Let $\mathcal{F}$ be as defined in Eq. 31. If it holds that, for some $\varepsilon < 1/16$,*

$$\forall h_W \in \mathcal{F}, \mathbb{P}\left(\exists g' \in \mathcal{G}_k : \max_{\mathbf{x}:\|x\|\leq 1} \|h_W(x) - g'(x)\| < \varepsilon \right) \geq \frac{1}{2}, \tag{32}$$

then it holds that

$$n \geq \frac{d^2}{2} \frac{\log_2 \frac{k}{\varepsilon}}{H_2\left(\frac{k}{n}\right)}.$$

The theorem follows by adapting the packing argument of [25]. A detailed proof is provided in Appendix G.

5 Conclusions

In this work, we have extended previous results on the Strong Lottery Ticket Hypothesis by quantifying the required overparameterization as a function of the sparsity of the subnetworks. Central to our results is a proof of the Random Fixed-size Subset Sum (RFSS) Problem, a refinement of the seminal Random Subset Sum (RSS) Problem in which the subsets have a required fixed size.

[4] Equivalently, the hypothesis of Corollary 1 must hold up to a factor $\Theta(\log_2 \frac{k}{\varepsilon})$.

A challenging open problem is to extend our analysis of RFSS to the multidimensional case, in which the random samples and targets are vectors in $\mathbb{R}^d$. Previous extension of RSS to the Multidimensional RSS have indeed allowed to prove structured-pruning version of the SLTH [8]. A Multidimensional RFSS result would then allow to quantify, in the structured pruning case, the dependency of the overparameterization w.r.t. the sparsity of the (structured) subnetworks.

Another future direction is to refine our analysis of the RFSS in Theorem 2 in order to improve the probability of success to $1 - \varepsilon$ rather than constant, thus allowing to avoid shaving off the extra factor $\log_2(1/\varepsilon)$ in our corollaries w.r.t. our lower bound, which is due to the amplification done in Corollary 1 to get to probability $1 - \varepsilon$.

Finally, an important future direction is to improve training-by-pruning methods such as [30, 27, 12, 11, 24] or to develop new ones, in order to allow to efficiently find strong lottery tickets of a desired sparsity, thus empirically validating our theoretical predictions.

6 Limitations and Impact

Limitations Similar to all the research conducted on the LTH and the SLTH, this work only proves the existence of lottery tickets. To this date, it is not clear if these subnetworks can be found reliably (no formal proof exists) in an efficient manner - however, empirical evidence suggests that efficient algorithms exist (e.g., [30, 27]).

Impact The contribution of this work is primarily theoretical and not confined to a specific domain. Its potential societal impact would, therefore, be closely tied to the particular scenarios to which it is applied. It could be interesting to compare the environmental impact of finding lottery tickets inside overparameterized networks. We also believe that our work has the potential to have a strong environmental impact as sparse NNs have massively reduced inference costs.

Acknowledgments and Disclosure of Funding

This research is supported by the EPSRC grant EP/W005573/1, and by the France 2030 program, managed by the French National Research Agency under grant agreements No. ANR-23-PECL-0003 and and ANR-22-PEFT-0002. It was also funded in part by the European Network of Excellence dAIEDGE under Grant Agreement Nr. 101120726, by SmartNet and LearnNet, and by the French government National Research Agency (ANR) through the UCA JEDI (ANR-15-IDEX-01), EUR DS4H (ANR-17-EURE-004), and the 3IA Côte d'Azur Investments in the Future project with the reference number ANR-19-P3IA-0002.

References

[1] Giampietro Allasia. "Approximation of the normal distribution functions by means of a spline function". In: *Statistica* 41.2 (1981), pp. 325–332.

[2] Sander Borst et al. "On the Integrality Gap of Binary Integer Programs with Gaussian Data". In: *Mathematical Programming* 197.2 (Feb. 2023), pp. 1221–1263. ISSN: 1436-4646. DOI: 10.1007/s10107-022-01828-1. (Visited on 06/05/2023).

[3] Rebekka Burkholz. "Convolutional and Residual Networks Provably Contain Lottery Tickets". In: *Proceedings of the 39th International Conference on Machine Learning*. Baltimore: PMLR, July 2022, pp. 2414–2433. (Visited on 02/27/2023).

[4] Rebekka Burkholz. "Most Activation Functions Can Win the Lottery Without Excessive Depth". In: *Thirty-Sixth Conference on Neural Information Processing Systems*. Dec. 2022. (Visited on 02/27/2023).

[5] Rebekka Burkholz et al. "On the Existence of Universal Lottery Tickets". In: *International Conference on Learning Representations*. virtual, Apr. 2022. (Visited on 02/27/2023).

[6] Arthur da Cunha, Francesco D'Amore, and Emanuele Natale. "Polynomially Over-Parameterized Convolutional Neural Networks Contain Structured Strong Winning Lottery Tickets". In: *Thirty-Seventh Conference on Neural Information Processing Systems*. Nov. 2023. (Visited on 04/23/2024).

[7] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. "Proving the Strong Lottery Ticket Hypothesis for Convolutional Neural Networks". In: *ICLR 2022 - 10th International Conference on Learning Representations*. Virtual, France, Apr. 2022. (Visited on 08/04/2022).

[8] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. "Proving the Strong Lottery Ticket Hypothesis for Convolutional Neural Networks". In: *ICLR 2022 - 10th International Conference on Learning Representations*. Virtual, France, Apr. 2022. (Visited on 08/04/2022).

[9] Arthur Carvalho Walraven Da Cunha et al. "Revisiting the Random Subset Sum Problem". In: *DROPS-IDN/v2/Document/10.4230/LIPIcs.ESA.2023.37*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. DOI: 10.4230/LIPIcs.ESA.2023.37. (Visited on 04/23/2024).

[10] Damien Ferbach et al. "A General Framework For Proving The Equivariant Strong Lottery Ticket Hypothesis". In: *The Eleventh International Conference on Learning Representations*. Sept. 2022. (Visited on 03/04/2024).

[11] Jonas Fischer and Rebekka Burkholz. "Plant 'n' Seek: Can You Find the Winning Ticket?" In: *International Conference on Learning Representations*. Apr. 2022. (Visited on 02/27/2023).

[12] Jonas Fischer, Advait Gadhikar, and Rebekka Burkholz. *Lottery Tickets with Nonzero Biases*. June 2022. arXiv: 2110.11150 [cs]. (Visited on 05/13/2024).

[13] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. Sept. 2018. (Visited on 10/20/2023).

[14] Advait Harshal Gadhikar, Sohom Mukherjee, and Rebekka Burkholz. "Why Random Pruning Is All We Need to Start Sparse". In: (June 2023). (Visited on 08/16/2023).

[15] Berivan Isik et al. "Adaptive Compression in Federated Learning via Side Information". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2024, pp. 487–495. (Visited on 05/16/2024).

[16] Berivan Isik et al. "Sparse Random Networks for Communication-Efficient Federated Learning". In: *The Eleventh International Conference on Learning Representations*. Sept. 2022. (Visited on 01/18/2024).

[17] Bohan Liu et al. *A Survey of Lottery Ticket Hypothesis*. Mar. 2024. DOI: 10.48550/arXiv.2403.04861. arXiv: 2403.04861 [cs]. (Visited on 05/16/2024).

[18] George S. Lueker. "Exponantially small bounds on the expected optimum of the partition and subset sum problem". In: *Ramdom Structures and Algorithms* 12 (1998), pp. 51–62.

[19] George S. Lueker. "On the average difference between the solutions to linear and integer knapsack problems". In: *Applied Probability - Computer Science, The Interface*. Vol. 1. Birkhäuser, 1982.

[20] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The Theory of Error-Correcting Codes*. Vol. 16. North-Holland Mathematical Library. North-Holland Publishing Company, 1977.

[21] Eran Malach et al. "Proving the Lottery Ticket Hypothesis: Pruning Is All You Need". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, July 2020, pp. 6682–6691. (Visited on 03/26/2023).

[22] James E Marengo, David L Farnsworth, Lucas Stefanic, et al. "A geometric derivation of the Irwin-Hall distribution". In: *International Journal of Mathematics and Mathematical Sciences* 2017 (2017).

[23] Laurent Orseau, Marcus Hutter, and Omar Rivasplata. "Logarithmic Pruning Is All You Need". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 2925–2934. ISBN: 978-1-71382-954-6. (Visited on 03/26/2023).

[24] Hikari Otsuka et al. *Partial Search in a Frozen Network Is Enough to Find a Strong Lottery Ticket*. Feb. 2024. DOI: 10.48550/arXiv.2402.14029. arXiv: 2402.14029 [cs, stat]. (Visited on 05/15/2024).

[25] Ankit Pensia et al. "Optimal lottery tickets via SUBSETSUM: logarithmic over-parameterization is sufficient". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. , Vancouver, BC, Canada, Curran Associates Inc., 2020. ISBN: 9781713829546.

[26] Valentin V. Petrov. *Sums of Independent Random Variables*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 2. Folge. Springer Berlin, Heidelberg, 1975, 348 pages. DOI: https://doi.org/10.1007/978-3-642-65809-9.

[27] Vivek Ramanujan et al. "What's Hidden in a Randomly Weighted Neural Network?" In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020, pp. 11890–11899. DOI: 10.1109/CVPR42600.2020.01191.

[28] N. Shakhaidarova. "Uniform local and global theorems for densities". In: *Izv. Akad. Nauk UzSSR Ser. Fiz-Mat. Nauk* 5 (1966), pp. 90–91.

[29] Irina Shevtsova. *On the absolute constants in the Berry Esseen type inequalities for identically distributed summands*. 2011. arXiv: 1111.6554 [math.PR].

[30] Hattie Zhou et al. "Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NIPS 2019)*. 2019, pp. 3592–3602.

A Lower Bound on the Ticket Size in [25]

The claim is a direct consequence of the proof of [25, Theorem 2] (Appendix B). There, in Step 3, it is shown that

$$|\mathcal{G}| \geq \frac{1}{2}\left(\frac{1}{2\varepsilon}\right)^{d^2},$$

where $\mathcal{G}$ is the set of subnetworks that can be formed. Let m be the number of parameters of the original network. If we consider subnetworks of size at most γm ($0 \leq \gamma \leq 1$), we have[5]

$$|\mathcal{G}| \leq \sum_{i=1}^{\gamma m}\binom{m}{\gamma m} \leq 2^{\gamma m \log_2(\frac{m}{\gamma m}e)},$$

which combined with the previous inequality implies

$$\gamma m \log_2\left(\frac{e}{\gamma}\right) \geq d^2 \log_2\left(\frac{1}{2\varepsilon}\right) - 1$$

If we have an overparameterized network of size $m = \mathcal{O}(d^2 \log_2\left(\frac{1}{2\varepsilon}\right))$, as in [25], we need $\gamma m = \Theta(m)$ for the last inequality to be satisfied (note that $\log_2\left(\frac{e}{\gamma}\right) \leq 1$, as $0 \leq \gamma \leq 1$).

B Visualizations

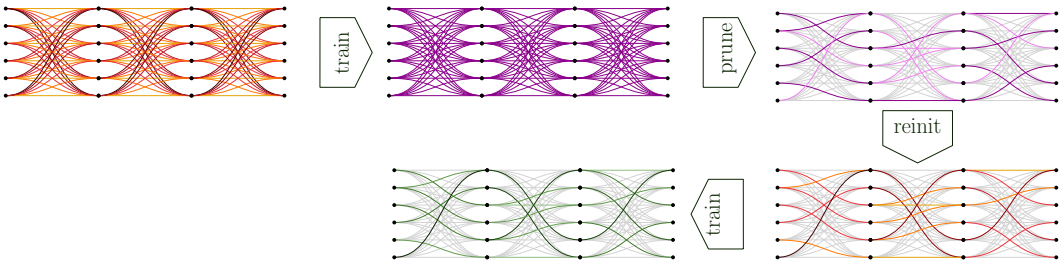

Figure 2: **Simplified representation of the procedure for finding Lottery Tickets (LTH)**. A large random neural network (step 1) is trained by iterative pruning with rewind: when the loss reaches a local minimum (step 2), some weights with smallest absolute value are pruned (step 3) and the value of the remaining edges is then reset to that of the initialization (step 4); finally, training is resumed and the final network is obtained (step 5). Remarkably, the sparser subnetwork is consistently able to reach a loss not larger than that right after pruning.

Figure 3: **Simplified representation of the procedure for finding Strongly Lottery Tickets (SLTH) / Training by pruning**. Previous work has shown that it is possible to sparsify large random neural network in order to obtain subnetworks that achieve good performance for a task under consideration, motivating the *Strong Lottery Ticket Hypothesis*. No training is required.

C Proof of Uniform$[-1, 1]$ being Sum-Bounded

In this section we provide a detailed proof of Lemma 1, which states that the uniform distribution in $[-1, 1]$ is sum-bounded, as stated in Definition 1. We remark that, while the proof is written for uniform random variables, it should be possible to extend it to a family of densities which are unimodal, with bounded variance, and bounded third moment.

[5] follows from the upper bound $\sum_{i=1}^{k}\binom{n}{i} \leq \left(\frac{en}{k}\right)^k$ on the partial sum of binomial coefficients.

Proof of Lemma 1. Note first that the distribution of the sum of n i.i.d. variables in $[0,1]$ is known as the Irwin–Hall distribution I_n.[6] We will use that $\mathrm{Var}[I_n] = \frac{n}{12}$, where $\mathrm{Var}[X]$ denotes the variance of the random variable X.

For $n \geq 2$, $f(x,n)$ can be defined as the convolution of $f(x) = f(x,1)$ and $f(x,n-1)$, i.e.,

$$f(x,n) = \int_{-\infty}^{+\infty} f(x-\tau, n-1) f(\tau) d\tau.$$

It is straightforward to show, by induction and an elementary substitution in the integral above, which is relied upon in the inductive step, that $f(x,n)$ is symmetric about 0, that is $f(x,n) = f(-x,n)$.

Let us now prove by induction that $f(x,n)$ is nondecreasing on the interval $[-n,0]$ and nonincreasing over $[0,n]$ (for simplicity, since it vanishes outside $[-n,n]$, we can consider directly the negative half and positive half of the real line, respectively, in the argument that follows).

The claims hold trivially for $f(x)$; also note that

$$f(\tau) = \begin{cases} \frac{1}{2} & \text{if} \quad -1 \leq \tau \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad \Longrightarrow \qquad f(x,n) = \frac{1}{2} \int_{-1}^{+1} f(x-\tau, n-1) d\tau.$$

If $x \leq x' \leq -1$. Since $x - \tau \leq x' - \tau \leq 0$, by inductive hypothesis we have that $f(x-\tau, n-1) \leq f(x'-\tau, n-1)$ over the whole interval $\tau \in [-1,1]$. Taking integrals yields $f(x,n) \leq f(x',n)$.

Now, consider the case when $x \leq -1 \leq x' \leq 0$. If $x+1 \leq -x'-1$, $x - \tau \leq x+1 \leq -x'-1 \leq -x'+\tau \leq -x+\tau$. By the symmetry about the origin, the inductive hypothesis is $f(x-\tau, n-1) = f(-x+\tau, n-1) \leq f(-x'+\tau, n-1) = f(x'-\tau, n-1)$ over the whole interval $\tau \in [-1,1]$, since $-1 \leq -x'+\tau \leq -x+\tau$. Taking integrals yields $f(x,n) \leq f(x',n)$. Otherwise, there exists τ_0 such that $x - \tau_0 = -x' - 1$, $x - \tau > -x' - 1$ for all $\tau \in [-1, \tau_0)$ and $x - \tau < -x' - 1$ for all $\tau \in (\tau_0, 1]$. By symmetry, using $-x = x' + 1 - \tau_0$, $f(x-\tau, n-1) = f(-x+\tau, n-1) = f(x'+1+\tau-\tau_0, n-1)$. Thus, for all $\tau \in [-1, \tau_0]$, via the change of variable $\sigma = -(1+\tau-\tau_0)$ in the middle integral below, we obtain that

$$\int_{-1}^{\tau_0} f(x-\tau, n-1) d\tau = \int_{-1}^{\tau_0} f(x'+1+\tau-\tau_0, n-1) d\tau = \int_{-1}^{\tau_0} f(x'-\sigma, n-1) d\sigma. \quad (33)$$

For all $\tau \in (\tau_0, 1]$, $x - \tau < -x' - 1 \leq -x' + \tau \leq -x + \tau$, by symmetry about the origin we have that $f(x-\tau, n-1) \leq f(x'-\tau, n-1)$ by the inductive hypothesis with the same reasoning of the case $x + 1 \leq -x' - 1$. Taking integrals over the range $[\tau_0, 1]$ for each term of the inductive hypothesis yields

$$\int_{\tau_0}^{1} f(x-\tau, n-1) d\tau \leq \int_{\tau_0}^{1} f(x'-\tau, n-1) d\tau \quad (34)$$

Eqs. 33 and 34 imply that $f(x,n) \leq f(x'n)$.

Trivially, if $-1 \leq x \leq x' \leq 0$, analogous ideas are put in place as for the previous case, therefore we omit the details. We have thus shown the nondecreasing monotonicity of $f(x,n)$ on the negative half of the real line. By the symmetry of $f(x,n)$ about the origin, on the positive half of the real line the nondecreasing monotonicity turns into nonincreasing monotonicity, and the proof is complete.

Lower bound (first inequality in Eq. 2). The variance of $\Sigma_{[n]}^{\mathcal{U}_n}$ is $n/3$ since $\Sigma_{[n]}^{\mathcal{U}_n} = 2(I_n(n) - n/2)$ and $\mathrm{Var}[I_n(n)] = n/12$. We define $Z_n^u = \frac{\Sigma_{[n]}^{\mathcal{U}_n}}{\sqrt{n/3}}$ and we note with F_n its cumulative distribution function. Z_n^u has expectation 0 and standard deviation 1. Consider the probability

$$P_L(n) = \Pr(\sqrt{n} \leq \Sigma_{[n]}^{\mathcal{U}_n} \leq 2\sqrt{n}) = \Pr(\sqrt{3} \leq Z_n^u \leq 2\sqrt{3}).$$

[6] It should be known that I_n is unimodal with a mode in $n/2$, but we were not able to find a reference. It is instructive to note, assuming that I_n is unimodal with a mode in $n/2$, it directly follows that its probability density function is increasing on the interval $[0, n/2]$, and then decreasing over $[n/2, n]$. This implies that $f(x,n)$ (the density of $\Sigma_{[n]}^{\mathcal{U}_n}$), is non decreasing in the interval $[-n, 0]$, has maximum at 0, and non increasing $[0,n]$ for all $n \geq 2$.

Now, we use the following form of Berry–Esseen inequality, discussed in [22][p.2]).[7]

Theorem 6 (Allasia [1]). *For all $n \geq 1$,*

$$|F_n(z) - \Phi(z)| \leq \frac{\sqrt{3}}{20\sqrt{n}},$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution.

Theorem 6 implies

$$P_L(n) \geq \Phi(2\sqrt{3}) - \Phi(\sqrt{3}) - 2 \cdot \frac{\sqrt{3}}{20\sqrt{n}}.$$

When $n \geq 18$,

$$\Phi(2\sqrt{3}) - \Phi(\sqrt{3}) - 2 \cdot \frac{\sqrt{3}}{20\sqrt{n}} \geq \Phi(2\sqrt{3}) - \Phi(\sqrt{3}) - 2 \cdot \frac{\sqrt{3}}{20\sqrt{18}} = C_{18} > 0.$$

That is $P_L(n) \geq C_{18} > 0$. When $2 \leq n < 18$, $P_L(n) = F_n(2\sqrt{3}) - F_n(\sqrt{3}) = c_n > 0$. We thus have

$$P_L(n) \geq \min\{C_i, \text{ for } 2 \leq i \leq 18\} = c_l' > 0.$$

Recall that $P_L(n) = \Pr(\sqrt{n} \leq \Sigma_{[n]}^{\mathcal{U}_n} \leq 2\sqrt{n})$. As the density $f(x, n)$ is decreasing on $\mathbb{R}^+$, we have

$$P_L(n) \leq f(\sqrt{n}, n)\sqrt{n}.$$

Thus,

$$f(\sqrt{n}, n) \geq \frac{P_L(n)}{\sqrt{n}}.$$

Since $P_L(n) \geq c_l'$ then for all $n \geq 2$

$$f(\sqrt{n}, n) \geq \frac{c_l'}{\sqrt{n}}.$$

When $n = 1$, the density $f(1, 1) = \frac{1}{2}$. So, by setting $c_l = \min(c_l', \frac{1}{2})$, we get that, for all $n \geq 1$, for all $0 \leq x \leq \sqrt{n}$:

$$f(x, n) \geq f(\sqrt{n}, n) \geq \frac{c_l}{\sqrt{n}}.$$

By a symmetric argument, we also have for all $n \geq 1$, for all $-\sqrt{n} \leq x \leq 0$:

$$f(x, n) \geq f(-\sqrt{n}, n) \geq \frac{c_l}{\sqrt{n}}.$$

Upper bound (second inequality in Eq. 2). Here, we bound the probability distribution function $f(x, n)$ of $\Sigma_{[n]}^{\mathcal{U}_n} = \sqrt{n/3}Z_n$, where we recall that $Z_n^u = \frac{\Sigma_{[n]}^{\mathcal{U}_n}}{\sqrt{n/3}}$. Denoting f_z the probability distribution function of Z_n^u, we have

$$f_z(x, n) = f\left(\sqrt{\frac{n}{3}}x, n\right)\sqrt{\frac{n}{3}}.$$

We use the following local limit theorem, discussed in [26][p.214].

Theorem 7 (Sahaidarova [28]). *Let $\{X_n\}$ be a sequence of independent random variables with a common density $p(x)$, such that $E[|X_1|^3] < \infty$, $E[X_1] = 0$, $E[X_1^2] = 1$ and $\sup p(x) \leq C$. Let $p_n(x)$ be the density of the random variable $\frac{1}{\sqrt{n}}\sum_{j=1}^n X_j$. Then*

$$\sup_x |p_n(x) - \phi(x)| \leq \frac{A\beta_3}{\sqrt{n}} \max(1, C^3),$$

where ϕ is the probability distribution function of a standard gaussian, A is an absolute constant, and $\beta_3 = E[|X_1|^3]$.

[7] It is also possible to obtain our result via classical Berry-Esseen inequality, due to the improved upper bound of 0.4748 on the absolute constant, provided in [29]. This would require replacing with 900 the cut-off value for n, which is 18 in the current version of the argument.

The theorem can be applied to a uniform continuous distribution with density $p^u(x) = \frac{1}{2\sqrt{3}}$ in the interval $[-\sqrt{3}, \sqrt{3}]$, which has mean 0 and variance 1. We thus get, for every $x \in \mathbb{R}$,

$$f_z(x,n) = p_n^u(x) \leq \phi(0) + \frac{A\beta_3}{\sqrt{n}} = \frac{1}{2\pi} + \frac{A\beta_3}{\sqrt{n}} \leq \frac{1}{2\pi} + A\frac{3\sqrt{3}}{4} = c'_u.$$

In conclusion, setting $c_u = \sqrt{3}c'_u$, for every $x \in \mathbb{R}$ it holds that

$$f(x,n) = \sqrt{\frac{3}{n}} f_z\left(\sqrt{\frac{3}{n}}x, n\right) \leq \frac{\sqrt{3}c'_u}{\sqrt{n}} = \frac{c_u}{\sqrt{n}}.$$

□

D Proof of Corollary 1

Proof of Corollary 1. As anticipated, we proceed in three steps.

Step 1: Hoeffding bound. We start by showing, following the idea at the base of [18, Corollary 3.3], that if n' is large enough, a standard Hoeffding bound ensures that with high probability a constant fraction of the sample follows a Uniform$[-1,1]$ distribution. Since we assumed that every X_i is a mixture of a Uniform$[-1,1]$ distribution with probability p, and another distribution with density g (given by the factors G_i), we can rewrite $X_i = B_i \cdot U_i + (1 - B_i) \cdot G_i$, with U_i being the uniform random variable, G_i being the random variable with density g, B_i being independent Bernoulli random variables with probability p.

Fix $\alpha = \alpha(p) \neq p$, and assume, for now, that n' satisfies Eq. 1, and therefore, since $\varepsilon < 1/2$, choosing $c_{\text{hyp}} = c_{\text{hyp}}(p) \geq (\alpha - p)^{-2}$, ensures that, defining $\varepsilon' = \varepsilon/2$,

$$n' \geq c_{\text{hyp}} \log_2 \frac{1}{\varepsilon} \geq \frac{1}{2(\alpha - p)^2} \ln \frac{1}{\varepsilon'}$$

and therefore

$$\Pr\left(\sum_i^{n'} B_i \leq \alpha n'\right) \leq e^{-2(\alpha-p)^2 n'} \leq e^{-\ln \frac{1}{\varepsilon'}} = \varepsilon'.$$

Thus

$$\Pr\left(\sum_i^{n'} B_i > \alpha n'\right) \geq 1 - \varepsilon',$$

that is, with high probability, there is a set of indices $I \subseteq [n']$ of size $|I| \geq \alpha n'$, such that for each $i \in I$ it holds $B_i = 1$, i.e. X_i is uniformly distributed.

Step 2: Application of Theorem 2 via rejection-sampling. Lemma 1 ensures that the uniform distribution of the $|I|$ random variables selected in *Step 1* is sum-bounded. Conditionally on the event $\{\sum_i^{n'} B_i > \alpha n'\}$, we can discard all random variables indexed outside I and apply directly Theorem 2 to $\alpha n'$ of the remaining ones, for any fixed k and $z \in [-\sqrt{k}, \sqrt{k}]$, since $\alpha c_{\text{hyp}} \geq 1$ by construction. This guarantees a success probability of c'_{thm} for approximating the given target z; thus,

$$\begin{aligned}
&\Pr\left(\exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon'\right) \geq \\
&\Pr\left(\exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon' \middle| \sum_i^{n'} B_i > \alpha n'\right) \Pr\left(\sum_i^{n'} B_i > \alpha n'\right) \geq \\
&\Pr\left(\exists S_z \subset I, |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon' \middle| |I| > \alpha n'\right)(1 - \varepsilon') \geq c'_{\text{thm}}(1 - \varepsilon') \geq \frac{3}{4} c'_{\text{thm}} = c_{\text{thm}}.
\end{aligned}$$

Step 3: Amplification. Finally, by a standard probability amplification argument and a union bound applied to Theorem 2, by paying an extra factor $\log_2(k/\varepsilon)$ in Eq. 1, the constant c_{thm} can be amplified to $1-\varepsilon$, and the existence of a suitable subset S_z holds simultaneously for all $z \in \left[-\sqrt{k}, \sqrt{k}\right]$. We now give more details on this amplification.

Recall that $\varepsilon' = \frac{\varepsilon}{2}$, and let $c_{\text{amp}} = c_{\text{amp}}(p) = 8\frac{c_{\text{hyp}}}{c_{\text{thm}}}$ and $r = \frac{4}{c_{\text{thm}}} \ln \frac{k}{\varepsilon}$. By assumption,

$$n \geq c_{\text{amp}} \frac{\log_2^2 \frac{k}{\varepsilon}}{H_2\left(\frac{k}{n}\right)} \geq 2r c_{\text{hyp}} \frac{\log_2 \frac{k}{\varepsilon}}{H_2\left(\frac{k}{n}\right)} \geq r c_{\text{hyp}} \frac{\log_2 \frac{k}{\varepsilon'}}{H_2\left(\frac{k}{n}\right)},$$

where the last inequality is ensured by $\varepsilon < 1/2$. By *Step 2*, we can apply Theorem 2, with ε' and $n' \geq c_{\text{hyp}} \frac{\log_2 \frac{k}{\varepsilon'}}{H_2\left(\frac{k}{n}\right)} = n^*$, allowing us to prove that we can ε'-approximate any target z with probability at least c_{thm}. The probability of failing to approximate some given z is then at most $1-c_{\text{thm}}$. From the sample Ω of sum-bounded random variables take r subsamples (without replacement) of cardinality n^* each, $\Omega_1, \ldots, \Omega_r$. The probability of failing to approximate some given z with subsetsums from Ω is less than that of failing to approximate it with subsetsums from within every Ω_i's, and the latter probability is at most $(1-c_{\text{thm}})^r$; thus, for every $z \in [-\sqrt{k}, \sqrt{k}]$,

$$\Pr\left(\nexists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon'\right) \leq (1-c_{\text{thm}})^r.$$

By an union bound, we also have that

$$\begin{aligned}
&\Pr\left(\forall z \in \left[-\sqrt{k}, \sqrt{k}\right], \exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon\right) \\
&\geq \Pr\left(\forall z \in \left\{-\sqrt{k} + i\varepsilon' : i \in \left[\frac{2}{\varepsilon'}\sqrt{k}\right]\right\}, \exists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon', \right) \\
&= 1 - \Pr\left(\exists z \in \left\{-\sqrt{k} + i\varepsilon' : i \in \left[\frac{2}{\varepsilon'}\sqrt{k}\right]\right\}, \nexists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon'\right) \\
&\geq 1 - \sum_{z \in \left\{-\sqrt{k} + i\varepsilon' : i \in \left[\frac{2}{\varepsilon'}\sqrt{k}\right]\right\}} \Pr\left(\nexists S_z \subset [n], |S_z| = k : |\Sigma_{S_z} - z| < \varepsilon'\right) \\
&\geq 1 - \frac{2}{\varepsilon'}\sqrt{k}\left(1-c_{\text{thm}}\right)^r = 1 - \frac{2}{\varepsilon'}\sqrt{k}\exp\left(\frac{4}{c_{\text{thm}}}\ln\left(\frac{k}{\varepsilon}\right) \cdot \ln(1-c_{\text{thm}})\right) \\
&\geq 1 - \frac{2}{\varepsilon'}\sqrt{k}\exp\left(-4\ln\frac{k}{\varepsilon}\right) = 1 - \frac{2}{\varepsilon'}\sqrt{k}\frac{\varepsilon^4}{k^4} \geq 1 - 4\varepsilon^3 \geq 1 - \varepsilon,
\end{aligned}$$

where the last inequality is ensured by $\varepsilon < 1/2$. This completes the proof.

□

E Proof of Corollary 2

Proof of Corollary 2. By definition of binary entropy, we have

$$H_2\left(\frac{k}{n}\right) = \frac{k}{n}\log_2\left(\frac{n}{k}\right) + \left(1-\frac{k}{n}\right)\log_2\frac{n}{n-k} \tag{35}$$

In particular, since both terms in the previous equation are positive, we get

$$H_2\left(\frac{k}{n}\right) \geq \frac{k}{n}\log_2\left(\frac{n}{k}\right) \tag{36}$$

We now use eq. (36) to derive an upper bound for the quantity $\frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)}\frac{\log_2^2 k + 2log_2 k \cdot log_2 1/\varepsilon}{n}$, which will be used later:

$$\frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)}\frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n} \leq \frac{c_{\text{amp}}}{\frac{k}{n}\log_2\left(\frac{n}{k}\right)}\frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n}$$

$$= c_{\text{amp}} \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{k} \frac{1}{\log_2\left(\frac{n}{k}\right)} \tag{37}$$

$$\leq c_{\text{amp}} \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{k} \tag{38}$$

$$\leq \frac{1}{2}, \tag{39}$$

where from eq. (37) to eq. (38) we used that $log_2 n/k \geq 1$ for $k \leq n/2$, and then the hypothesis $k \geq 2c_{\text{amp}}\left(\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}\right)$ directly gives eq. (39). Let us now rewrite eq. (3) in a more convenient form:

$$n\frac{H_2\left(\frac{k}{n}\right)}{c_{\text{amp}}} \geq \log_2^2 \frac{k}{\varepsilon}$$

$$n\frac{H_2\left(\frac{k}{n}\right)}{c_{\text{amp}}} \geq \log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon} + \log_2^2 \frac{1}{\varepsilon}$$

$$n\left(\frac{H_2\left(\frac{k}{n}\right)}{c_{\text{amp}}} - \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n}\right) \geq \log_2^2 \frac{1}{\varepsilon}$$

$$n\left(1 - \frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)} \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n}\right) \geq \frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)} \log_2^2 \frac{1}{\varepsilon} \tag{40}$$

$$n \geq \frac{c_{\text{amp}}}{\left(1 - \frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)} \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n}\right)} \frac{\log_2^2 \frac{1}{\varepsilon}}{H_2\left(\frac{k}{n}\right)} \tag{41}$$

Using eq. (39) we get

$$\frac{c_{\text{amp}}}{\left(1 - \frac{c_{\text{amp}}}{H_2\left(\frac{k}{n}\right)} \frac{\log_2^2 k + 2log_2 k \cdot log_2 \frac{1}{\varepsilon}}{n}\right)} \leq 2c_{\text{amp}} \tag{42}$$

To satisfy eq. (41), we can then choose n such that

$$n \geq 2c_{\text{amp}} \frac{\log_2^2 \frac{1}{\varepsilon}}{H_2\left(\frac{k}{n}\right)}, \tag{43}$$

and then apply Corollary 1 to end the proof. □

F Proof of Theorem 3

In the proof we will refer to the following results, upon which [25][Theorem 1] relies (the statement below slightly differ as we fix two small typos in their notation and mixing coefficients). With the understanding that by a mixture D of a distribution D_1 and D_2 with probability p it is meant that the pdf (we adopt the convention that this term includes generalised functions, such as Dirac deltas for point masses) of D can be written as a convex combination of the pdf of D_1 and that of D_2, that is $f_D = pf_{D_1} + (1-p)f_{D_2}$. For the unfamiliar reader, we note that in the literature this is often stated in short as $D = pD_1 + (1-p)D_2$.

Lemma 2 ([25][Corollary 1]). *Let $X \sim$Uniform$[0,1]$ (or $X \sim$Uniform$[-1,0]$) and $Y \sim$Uniform$[-1,1]$ be independent random variables. Let P be the distribution of the XY and δ_0 the Dirac delta at 0. Let D be the distribution obtained as mixture of δ_0 and P with probability $1/2$. Then D is the mixture of a Uniform$[-1/2, 1/2]$ and some distribution Q with probability $\ln(2)/4$.*

Corollary 3 ([25][Corollary 2]). *Let $X_1, \ldots, X_n$ be iid with distribution D as defined in Lemma 2, where $n \geq C\ln(2/\varepsilon)$ for some universal constant C. Then*

$$\Pr\left(\forall\, z \in [-1,1],\ \exists\, S \subset [n]\ : |z - \sum_{i \in S} X_i| \leq \varepsilon\right) \geq 1 - \varepsilon.$$

Proof of Theorem 3. The key idea is exploiting Corollary 1 at each step of the pruning strategy established in [25][Theorem 1], where Corollary 3 is used instead. Without loss of generality, we replace their $\min\{\varepsilon, \delta\}$ with ε. For the sake of easily following the approach adopted in [25], let us define $n^*(x)$ as the function

$$n^*(x) = c_{\text{amp}} \frac{\log_2^2(kx)}{H_2\left(\frac{k}{n^*(x)}\right)} \tag{44}$$

where $k = \gamma' n^*(x)$. In the following, we use n^* as short for $n^*(1/\varepsilon)$, and we will only explicitely provide an argument for n^* when it is different than $1/\varepsilon$. For instance, in the last step of the proof, we will use $n^*(2\ell d_i d_{i-1}/\varepsilon)$, which matches the definition of n_i^* given in Eq. 28.

Consider [25][Lemma 1]. When approximating a single link (that is, a weight), after the overparameterization (which creates an additional layer of width $2n^*$ in between the input and the output node) via $4n^*$ links, instead of pruning via Corollary 3, we prune via Corollary 1 twice in the second layer, that is we ensure that only $k = \gamma' n^*$ edges yield the desired approximation, both in the edges corresponding to the positive part of the input weights and in those corresponding to the negative part. Thus we obtain at most $4k$ surviving edges, after the preprocessing step and the pruning mask is applied. This yields a sparsity of at least $\alpha' = 1 - \gamma'$. Note that it is because of the preprocessing step that we go from distributions Uniform$[-1, 1]$ to distributions D, as defined in Lemma 2, which are shown to be a mixture with Uniform$[-1, 1]$ and therefore can be also handled via Corollary 1.

Consider [25][Lemma 2]. When approximating a real-valued multivariate linear function, after the overparameterization (which creates an additional layer of width $2dn^*(d/\varepsilon)$ in between the d input nodes and the output node) one simply iterates the ideas of the previous case d times. For each input node, the overparameterization surviving the preprocessing step on the weights of the input layer is $4n^*(d/\varepsilon)$. Pruning the second layer of the overparameterized link for each input via Corollary 1 with $k = \gamma' n^*(d/\varepsilon)$ (again, performing this both on the edges corresponding to the positive part of the input weights and in those corresponding to the negative part), instead of exploiting Corollary 3, yields that at most $4dk$ edges survive after the pruning mask is applied. This yields a sparsity of at least $\alpha' = 1 - \gamma'$.

Finally, consider [25][Lemma 3]. When approximating a layer with input dimension d_1 and output dimension d_2, after the overparameterization (which creates an additional layer of width $2d_1 n^*(d_1 d_2/\varepsilon)$ in between the input nodes and the output nodes) one iterates the ideas of the previous case d_1 times in the input layer through the same preprocessing step, and d_2 times in the output layer, one for each of the d_1 blocks created by the preprocessing (essentially the weights in the input layer are *re-used* d_2 times). For each input node, the overparameterization surviving the preprocessing step is at most $2(d_2 + 1)n^*(d_1 d_2/\varepsilon)$. Overall, after the preprocessing step, we have at most $2d_1(d_2 + 1)n^*(d_1 d_2/\varepsilon)$ parameters. We then use Corollary 1 (with $k = \gamma' n^*(d_1 d_2/\varepsilon)$) to prune the number of parameters between the introduced additional layer and the d_2 outputs down to $2d_1 d_2 \gamma' n^*(d_1 d_2/\varepsilon)$. As for the edges between the d_1 inputs and the additional layer, only those that reach a neuron in the additional layer, from which there is at least one outgoing edge towards the d_2 outputs, are used; since for each of the d_1 blocks of $2n^*(d_1 d_2/\varepsilon)$ neurons in the additional layer we only kept $2\gamma' n^*(d_1 d_2/\varepsilon)$ outgoing edges to each of the d_2 output neurons, in the worst case (all the nodes involved in the subsetsums are disjoint) we keep $2d_2\gamma' n^*(d_1 d_2/\varepsilon)$ of them for each of the d_1 neurons. Globally, we are left with a total of at most $2d_1 d_2 \gamma' n^*(d_1 d_2/\varepsilon)$ edges both in the input layer and in the output layer, thus a total of $4d_1 d_2 \gamma' n^*(d_1 d_2/\varepsilon)$ edges survive the pruning. The density of the surviving edges is then less than

$$\frac{4d_1 d_2 \gamma' n^*(d_1 d_2/\varepsilon)}{2d_1^2 n^*(d_1 d_2/\varepsilon) + 2d_1 d_2 n^*(d_1 d_2/\varepsilon)} = \frac{2d_2\gamma'}{d_1 + d_2} = \frac{(d_1 \frac{d_2}{d_1} + d_2)\gamma'}{d_1 + d_2} \leq \rho_1 \gamma',$$

where $\rho_1 = \max\{d_1/d_2, d_2/d_1\}$ and in the last inequality we used that $d_1 d_2/d_1 + d_2 \leq \rho_1(d_1 + d_2)$ since $\rho_1 \geq 1$. This ensures a sparsity $\alpha' \geq 1 - \rho_1\gamma'$.

[25][Theorem 1] consists of performing, for every $i \in [\ell]$, the previous step on layer i with input dimension d_{i-1} and output dimension d_i. The overparameterization creates an additional layer of nodes of width $2d_{i-1} n^*(2\ell d_{i-1} d_i/\varepsilon)$ in between the d_{i-1} input nodes and the d_i output nodes. Since the construction is stacked ℓ times, this generates 2ℓ layers for the overparameterized network, which

will therefore have a starting number of parameters

$$m = \sum_{i=1}^{\ell} 2d_{i-1}^2 n^*(2\ell d_{i-1}d_i/\varepsilon) + 2d_{i-1}d_i n^*(2\ell d_{i-1}d_i/\varepsilon).$$

Corollary 1 applied to each stacked overparameterized layer instead of Corollary 3 as in the previous step yields that the total number of parameters left after the pruning is

$$m_t \leq \sum_{i=1}^{\ell} 4d_{i-1}d_i k_i,$$

where $k_i = \gamma' n^*(2\ell d_{i-1}d_i/\varepsilon)$. Recall that $\rho = \max_i \rho_i$, where $\rho_i = \max\{d_i/d_{i-1}, d_{i-1}/d_i\} \geq 1$. Recall that $\gamma = \rho\gamma'$. We obtain that

$$\begin{aligned}
m_t \leq & \sum_{i=1}^{\ell} 2d_{i-1}\frac{d_i}{d_{i-1}}d_{i-1}k_i + 2d_{i-1}d_i k_i \\
\leq & \sum_{i=1}^{\ell} 2d_{i-1}\rho_i d_{i-1}k_i + 2d_{i-1}d_i\rho_i k_i \\
\leq & \rho\sum_{i=1}^{\ell} 2d_{i-1}^2 k_i + 2d_{i-1}d_i k_i \\
= & \rho\gamma'\sum_{i=1}^{\ell} 2d_{i-1}^2 n^*(2\ell d_{i-1}d_i/\varepsilon) + 2d_{i-1}d_i n^*(2\ell d_{i-1}d_i/\varepsilon) = \gamma m
\end{aligned}$$

We then get that the density of the edges surviving the pruning is $m_t/m \leq \gamma$, which implies a sparsity of at least $\alpha = 1 - \gamma$.

□

G Proof of Theorem 5

Proof of Theorem 5. Consider the space $\mathcal{W}_k = \{W \in \mathbb{R}^{d\times d} : \|W\| \leq \sqrt{k}\}$, and let $\mathcal{P}_k$ be a 2ε-separated set of $\mathcal{W}_k$, i.e. a subset $\mathcal{P}_k \subset \mathcal{W}_k$ such that for all distinct $W, W' \in \mathcal{P}_k$ it holds $\|W - W'\| > 2\varepsilon$. We denote $\mathcal{W} = \mathcal{W}_1$, $\mathcal{P} = \mathcal{P}_1$, and the set of all possible subnetworks of g as $\mathcal{G}$ (note that this does not denote $\mathcal{G}_1$, the set of all subnetworks of size 1).

Step 1: Packing argument. In [25][Theorem 2, Step 1], it is shown that any function g' can only approximate at most one member of $\mathcal{P}$ for bounded input x (say, $\|x\| \leq 1$). In particular, this also applies to functions g' representing the elements of $\mathcal{G}_k$.

Step 2: Relation between $|\mathcal{G}_k|$ and $|\mathcal{P}_k|$. By *Step 1*, in [25][Theorem 2, Step 2] it is shown that $|\mathcal{P}| \leq 2|\mathcal{G}|$, under the assumption of Eq. 32, with $\mathcal{G}_k$ replaced by $\mathcal{G}$). Therefore, also by *Step 1*, replacing $\mathcal{P}$ with $\mathcal{P}_k$ and $\mathcal{G}$ with $\mathcal{G}_k$ in [25][Theorem 2, Step 2], it holds that $|\mathcal{P}_k| \leq 2|\mathcal{G}_k|$. Note that $|\mathcal{G}_k| = \binom{n}{k}$, the number of different ways in which we can select k parameters out of n, so we actually get

$$\binom{n}{k} > \frac{|\mathcal{P}_k|}{2}. \tag{45}$$

Step 3: Lower bound on $|\mathcal{P}_k|$. Let us now consider a 2ε-separated set $\mathcal{P}_k^{\max}$ of maximal cardinality. In [25][Theorem 2, Step 3] it is shown that

$$|\mathcal{P}^{\max}| \geq \frac{\text{Vol}(\mathcal{W})}{\text{Vol}(\{W \in \mathcal{W} : \|W\| \leq 2\varepsilon\})} = \left(\frac{1}{2\varepsilon}\right)^{d^2}.$$

Here Vol is the Lebesgue measure in $\mathbb{R}^{d\times d}$ identified with $\mathbb{R}^{d^2}$.By the exact same argument, replacing $\mathcal{W}$ with $\mathcal{W}_k$ and thus $\mathcal{P}^{\max}$ with $\mathcal{P}_k^{\max}$, it holds that

$$|\mathcal{P}_k^{\max}| \geq \frac{\text{Vol}(\mathcal{W}_k)}{\text{Vol}(\{W \in \mathcal{W}_k : \|W\| \leq 2\varepsilon\})} = \left(\frac{\sqrt{k}}{2\varepsilon}\right)^{d^2}.$$

Combining this fact with Eq. 45 applied to $\mathcal{P}_k^{\max}$ implies that

$$\binom{n}{k} > \frac{1}{2}\left(\frac{\sqrt{k}}{2\varepsilon}\right)^{d^2}. \tag{46}$$

Step 4: Lower bound on n. Consider the standard bound found in [20]

$$\binom{n}{k} \leq \sqrt{\frac{n}{2\pi k(n-k)}}2^{nH_2(k/n)}.$$

and combine it with with Eq. 46. It follows that

$$2^{nH_2\left(\frac{k}{n}\right)} \geq \frac{1}{2}\sqrt{\frac{2\pi k(n-k)}{n}}\left(\frac{\sqrt{k}}{2\varepsilon}\right)^{d^2}$$

and taking the logarithm of both sides yields the sought lower bound on n:

$$nH_2\left(\frac{k}{n}\right) \geq \frac{1}{2}\log_2\left(\frac{2\pi k(n-k)}{n}\right) + d^2\log_2\frac{\sqrt{k}}{2\varepsilon} - 1 \tag{47}$$

$$\geq d^2\left(\frac{1}{2}\log_2 k + \log_2\frac{1}{\varepsilon} - 1\right) - 1 \tag{48}$$

$$\geq \frac{d^2}{2}\log_2\frac{k}{\varepsilon}, \tag{49}$$

where from Eq. 47 to Eq. 48 we exploited the definition of λ, which ensures that the first term in the r.h.s. of Eq. 47 is nonnegative;[8] from Eq. 48 to Eq. 49 we used that for all $\varepsilon < 1/16$ it holds that

$$d^2\left(\log_2\frac{1}{\varepsilon} - 1\right) \geq 1.$$

□

H Details of comparison with Malach et al. [21]

We show that $cz \geq c_{\text{amp}}\frac{\log_2^2(cz^2)}{\log_2(z)}$ holds for a big enough constant c. Recall that $z = \frac{m_t}{\varepsilon}$, so we can always assume $\log_2(z) \geq 1$. We have

$$\log_2^2(cz^2) = (\log_2(c) + 2\log_2(z))^2 \tag{50}$$

$$= \log_2^2(c) + 4\log_2(c)\log_2(z) + 4\log_2^2(z) \tag{51}$$

$$\overset{(a)}{\leq} 6(\log_2^2(c) + \log_2^2(z)) \tag{52}$$

$$\overset{(b)}{\leq} 12\log_2^2(c)\log_2^2(z), \tag{53}$$

where in (a) we used that $2ab \leq a^2 + b^2$, and in (b) that $a + b \leq 2ab$ for a and b greater than 1.

We can then focus on showing that there is a big enough constant c such that $cz \geq 12c_{\text{amp}}\log_2^2(c)\log_2^2(z)$. We get $c \geq 12c_{\text{amp}}\log_2^2(c)\frac{\log_2^2(z)}{z}$, and we have

$$\log_2^2(c)\frac{\log_2^2(z)}{z} \leq \log_2^2(c) \tag{54}$$

$$\leq \sqrt{c}. \tag{55}$$

We can then focus on $c \geq 12c_{\text{amp}}\sqrt{c}$, which is satisfied for $c \geq 144c_{\text{amp}}^2$.

[8] This term being nonnegative is equivalent to $k(1 - k/n) \geq 1/2\pi$, and since $1 \leq k \leq \lambda n$, any $\lambda \leq 1 - 1/2\pi$ ensures it.