The Puzzling Success of Overparameterization: Lottery Tickets or Escape Dimensions?

Flavio Martinelli [1] **Johanni Brea** [1] **Wulfram Gerstner** [1]

Abstract

Lotteries and tickets are often used as a didactical analogy to explain the success of overparameterized neural networks: "larger networks succeed because they more likely contain a well-initialized subnetwork that can learn the task in isolation, much like buying more tickets increases the chances of winning a lottery." This explanation is intuitive but misleading: it suggests that subnetworks can be treated in isolation from the rest of the network. Following this reasoning leads to interpreting learning in wide networks as a multi-start optimization process, where gradient descent simply conducts a parallel search over subnetworks. We argue that this view is flawed since, among other reasons, winning tickets can be made to fail by perturbing the rest of the network. We put forward a more accurate intuitive picture for the success of overparameterization based on the geometry of loss landscapes: increasing width expands the set of available dimensions for optimization, making it easier to escape bad local minima. Moreover, as width grows, bad minima become increasingly rare relative to good minima. As the field grows mature, it is important to refine the analogies we use to explain foundational phenomena, such as the apparent redundancy of large networks, reconciling practitioners' intuitions with modern theoretical insights.

Nowadays, we can optimize neural networks to achieve impressive results by implementing gradient descent with a few lines of code. Yet, behind this simplicity lies the coordination of billions of moving parts: the network parameters. These parameters interact in complex ways to arrange themselves into configurations that solve challenging tasks. The lottery analogy (introduced by Frankle & Carbin, 2019) has become a popular way to make sense of this complexity, by stripping down the many nonlinear interactions to a simple combinatorial picture of competing subnetworks: after initialization each subnetwork is a lottery ticket, and training a large network is like buying many tickets to increase the chances of finding the one that can solve the task. The field has embraced this metaphor to explain why seemingly redundant, large networks are necessary for training, permeating both research and didactical content. But is this analogy accurate? Can overparameterization be explained in terms of subnetworks? In this paper, we argue that **the lottery ticket *analogy* is a misleading mental model for understanding overparametrization in neural networks** and that instead **the geometry of the loss landscape provides a more accurate intuition, via the concept of Escape Dimensions** (Fig. 1). We coin the term Escape Dimensions Theory to bring together ideas across studies (among others: Fukumizu & Amari 2000; Mei et al. 2018; Simsek et al. 2021; Belkin 2021) under a novel, intuitive lens that does not invoke lottery tickets to explain overparameterization: as networks grow, bad minima transform into saddles, revealing new dimensions for gradient descent to escape (Section 3).

Escape Dimensions Theory

Adding seemingly redundant parameters to a neural network increases the number of dimensions available to escape sub-optimal minima, making it easier for gradient descent to traverse the loss landscape.

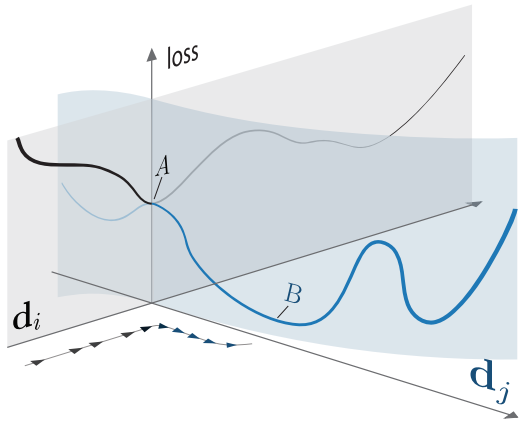

Figure 1. **Escape Dimensions.** In the landscape spanned by the axis $\mathbf{d}_i$, gradient descent approaches the local minimum A (black path). After overparameterization, the new landscape dimension $\mathbf{d}_j$ (in blue) enables escape (Fukumizu & Amari, 2000) to a better minimum B . The trajectory may later drift away from the axis $\mathbf{d}_j$.

[1]EPFL. Correspondence to: <flavio.martinelli@epfl.ch>.

Preprint. May 12, 2026.

1. A common metaphor in the field

*"By this logic, overparameterized networks are easier to train **because** they have more combinations of subnetworks that are potential winning tickets."* [Q3]

Frankle & Carbin, 2019

The lottery metaphor is often invoked in the context of overparameterized networks (evidence in Section 1.2), prompted by statements such as the one quoted here, to explain the need for overparameterization in deep learning. However, it originates from the lottery ticket hypothesis (LTH, Frankle & Carbin, 2019), formulated in the context of pruning experiments. Before the LTH was proposed, countless experiments in the pruning literature had shown that large models, once trained, could be stripped of most of their parameters without significant loss in performance (LeCun et al., 1989; Reed, 1993; Han et al., 2015; Hoefler et al., 2021). In 2019, Frankle & Carbin discovered that there exist sparse networks (identified by pruning) that can be trained from scratch to full accuracy, given the right initialization. The existence of these so-called *winning tickets* was surprising, because it seemed that dense, large networks were necessary for successful training. However, identifying these well-initialized subnetworks still required information from the initial training of a dense network. Since then, the LTH has inspired a large body of work, with the dream of reducing the energy demands of deep learning not only at inference, but also at training time (Evci et al., 2020; Wang et al., 2020; Liu et al., 2024; Jeffares & van der Schaar, 2025). The observation that small networks are difficult to train, despite the *existence* of winning tickets, together with the LTH popularity, created fertile ground for the lottery analogy to spread as an intuitive explanation of why larger networks are easier to optimize (Section 1.1).

This paper does not challenge the empirical validity of the LTH. What we challenge is the intuition, or analogy, or common wisdom that the lottery metaphor has generated in the community to explain the apparent redundancy of large neural networks. The concept of a lottery, a "game of chance" with a *specific* set of rules, invokes three key properties:

- **Sufficiency**: having a winning ticket is sufficient to win the lottery.
- **Scaling**: the probability of winning the lottery scales predictably with the number of purchased tickets.
- **Independence**: tickets in a large lottery are approximately independent samples.

We find that, over the literature and in various forms of didactical content, these properties are often directly or indirectly translated to the context of subnetworks embedded in overparameterized networks (see Section 1.2 and Tables A1, A2). Namely:

- **Sufficiency**: containing a winning subnetwork at initialization is sufficient to train the network successfully.
- **Scaling**: the probability of successful training scales combinatorially with the size of the network.
- **Independence**: the training outcome of a subnetwork is independent from the rest of the network.

Frankle & Carbin (2019) themselves provided an interpretative key for their metaphor, explicitly stating a sufficiency (or necessity) conjecture[Q1], as well as a scaling argument[Q3]. If these ideas were correct, training large networks would amount to running many subnetworks in parallel, with the successful one emerging as the winner over the course of optimization (Section 1.3). The rest of this section will clarify the origin of the misconceptions, provide evidence for how widespread they are in literature and online content, and describe how learning in overparameterized networks would work if the lottery analogy were interpreted literally.

1.1. Mistaking the hypothesis for the conjecture

"Empirically, many people have found that bigger models are easier to train (often explained with the 'lottery ticket hypothesis')" [Q9]

Abnar et al., 2020

The statement defining the original hypothesis is of empirical nature: it hypothesizes the existence of trainable subnetworks inside large, dense, successfully trained networks. This phenomenon has been thoroughly validated (Frankle & Carbin, 2019; Zhou et al., 2019; Morcos et al., 2019; Renda et al., 2020; Frankle et al., 2020; Ma et al., 2021) and extended in a vast range of settings (Yu et al., 2020; Prasanna et al., 2020; Li et al., 2020; Martinelli et al., 2020; Chen et al., 2020; 2021a; Vischer et al., 2022; Chen et al., 2021b;c; Kim et al., 2022). What holds empirically is described in the original statement of the hypothesis:

The Lottery Ticket Hypothesis. *A randomly-initialized, dense neural network contains a subnetwork that is initialized such that — when trained in isolation — it can match the test accuracy of the original network after training for at most the same number of iterations.*

The intuition behind the metaphor, however, is tightly linked to the conjecture they proposed to explain their observations. Importantly, Frankle & Carbin were explicit that their conjecture lacked empirical evidence:

The Lottery Ticket Conjecture (LTC). *[...] an untested conjecture that SGD seeks out and trains a subset of well-initialized weights. Dense, randomly-initialized networks are easier to train than the sparse networks that result from pruning because there are more possible subnetworks from which training might recover a winning ticket.*

Superscripts of the form Q# refer to verbatim quotes from academic and online sources, collected in Tables A1, A2. They list examples of how the lottery metaphor is (mis)used in the literature.

We speculate that referring to a popular empirical result with the words "lottery", "tickets", "winning" has contributed to part of the community absorbing the conjecture as if it were an established fact. Stripped of the analogy, the LTH could be neutrally renamed: the *trainable subnetworks hypothesis*, specifically because the term lottery is only tied to the explanatory mechanism proposed by the conjecture. We argue that the use of this analogy in many subsequent papers contributed to mistaking the evidence supporting the *existence* of trainable subnetworks for evidence supporting the *causal explanation* that large networks learn because there exist trainable subnetworks embedded in them (LTC).

1.2. How the lottery analogy is integrated in the field

"If you start with a very overparameterized network, probability theory gives the network much higher chances to include a better subnetwork than a very small one." [Q13]
Koster et al., 2022

Statements of this form are often used to explain why large networks are needed to fit data successfully. They are usually invoked to resolve the apparent paradox that successful sparse networks *do exist* and can be found via pruning, but *only after training* a large dense model. Similar statements appear across a wide range of sources, including research papers, blog posts, and lecture slides. Some of these sources implicitly or explicitly endorse one of the three properties stated above (sufficiency[Q1,5,7,15,21,25,26,33], scaling[Q3,5,6,9,13-16,19-21,24,29,30,33,38,39], independence[Q7,31,39]), others mention them as part of a broader narrative, while some even report these as hypothesis backed by evidence (as opposed to conjectures)[Q9,16,17,24,29]. By scanning the literature, we found a collection of more than 30 representative quotes that we list verbatim in Tables A1, A2. We describe our methodology for collecting them in Appendix A.

This framing also manifests itself in the types of research questions it motivates. One notable example is the theoretical and empirical search for conditions under which well-performing subnetworks exist at initialization. The paper titled "Proving the lottery ticket hypothesis: pruning is all you need" (Malach et al., 2020) first formalized and proved a modified version of the LTH, sometimes referred to as the *strong lottery ticket hypothesis*. This variant conjectures that subnetworks with good performance can exist at initialization, without requiring any training at all (Ramanujan et al., 2020). Many subsequent works have built upon this idea, improving bounds and exploring different settings (Pensia et al., 2020; Orseau et al., 2020; Burkholz et al., 2021; Burkholz, 2022; Berner et al., 2022; da Cunha et al., 2022; Ferbach et al., 2023; Natale et al., 2024; Kumar & Natale, 2025; Otsuka et al., 2026). These works are valuable in their own right and do not aim to explain why overparameterized networks are easier to optimize. However, they implicitly reinforce one particular framing of tickets and

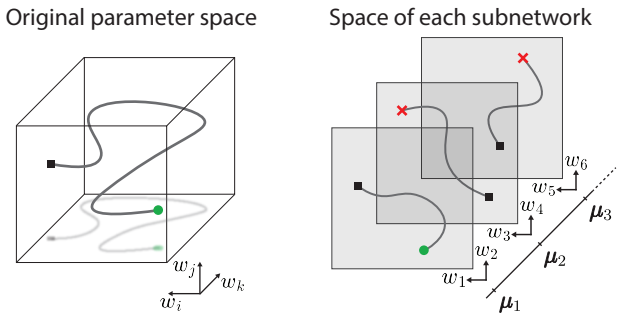

Figure 2. **Multi-start optimization view of the landscape.** Since the outcome of a lottery ticket is independent from other tickets, the analogy *implies* that each subnetwork's optimization outcome is independent. Seen from the parameter space perspective, it is as if we could describe the trajectory of learning in a large space (left) as a combination of independent trajectories over subspaces (right). Each slice of parameter space is defined by a subnetwork weight mask $\boldsymbol{\mu}_i$. If true, training an overparameterized network would be equivalent to running multiple independent optimizations in parallel, each corresponding to a differently initialized subnetwork.

lotteries: success is associated with the existence of suitable subnetworks, rather than with the dynamics of learning in the full parameter space. Calling this variant a strong version of the LTH further encourages this interpretation, despite the fact that it no longer addresses training at all. In this view, success is explained entirely in terms of pre-existing structure, encouraging a combinatorial picture in which the task is to identify the right subnetwork. Even when such subnetworks provably exist, they are not available a priori as isolated objects. Absurdly, but not surprisingly, identifying them in practice requires another dense, overparameterized optimization process; the same that the original motivation hoped to avoid (Zhou et al., 2019; Ramanujan et al., 2020; Wortsman et al., 2020; Evci et al., 2020; Bai et al., 2022b).

Finally, this interpretation is also familiar from informal scientific discourse. In our own experience, we hear and see it used in talks, discussions, and pedagogical settings to provide an intuitive account of why overparameterization is helpful. At the time of this piece, many readers will likely recognize similar explanations from their own interactions.

1.3. Multi-start optimization?

"If you want to win the lottery, just buy a lot of tickets and some will likely win. Buying a lot of tickets = having an overparameterized neural network for your task." [Q33]
Princeton-CS-598D, 2020

In a lottery, it is enough to have one winning ticket to win the prize; all other tickets are irrelevant. In discussions related to the LTH, one frequently encounters expressions such as "*SGD seeks out a winning ticket*"[Q2,11,14,22,30,32,36] or "*width acts as a form of parallel search*"[Q20,36,34,38,39]. When stated by various authors, these formulations rarely fully commit to the claims. However, they are not conceptually

neutral. Taken seriously, they suggest that subnetworks are well-defined objects at initialization, whose optimization success is fixed prior to training, and that optimization acts primarily as a process of selection. Once this interpretation is adopted, the implied picture of learning becomes quite specific: if subnetworks are meaningful candidates that pre-exist training, and if success requires at least one of them to be "good," then optimization (SGD) can be understood as a process of selection among these candidates. Training merely enables the optimization trajectory of the winning subnetwork while suppressing the others[Q20]. From this perspective, overparameterization helps by increasing the number of attempts at solving the task. Fig. 2 illustrates this view. Because the number of subnetworks available in a dense network grows combinatorially with network width, this view *predicts* an extremely rapid increase in the likelihood of success as models become larger.

2. Where is the lottery in neural networks?

No metaphor is expected to hold in every aspect. When Shakespeare writes "All the world's a stage," people do not wonder where to buy theatre tickets; nor do we debate whether networks can literally buy something. What matters is the *core point of comparison* that makes a metaphor informative or educational. Finding this core is the very work that the metaphor asks of us (Eco, 1986). What mechanisms do lotteries and networks supposedly share? Both the original authors and the subsequent literature explicitly emphasize two: the *sufficiency* and *scaling* properties (a third property, *independence* often appears implicitly, as in most games of chance). In the following, we formalize these properties and show that they do not hold for subnetworks.

2.1. Subnetworks depend on their context

"A key question, then, is whether the presence of a winning ticket is necessary or sufficient for SGD to optimize a neural network to a particular test accuracy." [Q1]

Frankle & Carbin, 2019

A core mechanism of lotteries is that obtaining a winning ticket is sufficient to win the prize, regardless of the presence of other tickets. Does this carry over to neural networks? We set up a teacher-student experiment where success is unambiguous: the student either reaches zero imitation loss or does not (details in Appendix F). The dataset is generated from a two-layer MLP with 4 hidden neurons. We then train student networks to imitate the teacher, by minimizing an MSE loss. Students and teacher share the same architecture and size ($m = 4$ hidden neurons). In the spirit of the LTH philosophy, we select the successful initializations (loss < 10^{-20}) and define them as our winning tickets I^*. We then embed each winning ticket in a larger network: the first 4 neurons are initialized with I^*, while additional neurons are drawn from the standard initialization distribution $\mathcal{G}$, giving $I = [I^*, I^+ \sim \mathcal{G}]$ (Fig. 3a, left). Simulations of Fig. 3a (right) show how the success rate evolves as neurons are added to a winning ticket (dark bars), compared to training from random initialization (grey bars): when one additional neuron ($m = 5$) is added to a winning ticket, the larger network yields success in only $\sim 65\%$ of runs. If having a winning subnetwork were sufficient, these larger networks should succeed every time (100%, green empty bars). However, the initialization of the remaining network is enough to disrupt a winning ticket.

This observation extends to practical architectures. We take winning tickets of the Conv-6 CIFAR-10 network of Frankle & Carbin (2019) and adversarially re-grow a random fraction of the pruned connections. Specifically, we initialize the regrown weights $\boldsymbol{\theta}_{\text{adv}}$ by solving:

$$\boldsymbol{\theta}^*_{\text{adv}} = \min_{\boldsymbol{\theta}_{\text{adv}}} \cos\Big(\nabla_{\boldsymbol{\theta}_T}\mathcal{L}(\boldsymbol{\theta}_T, \boldsymbol{\theta}_{\text{adv}}),\ \nabla_{\boldsymbol{\theta}_T}\mathcal{L}(\boldsymbol{\theta}_T, \mathbf{0})\Big) \quad (1)$$

where $\nabla_{\boldsymbol{\theta}_T}\mathcal{L}(\boldsymbol{\theta}_T, \mathbf{0})$ is the initial gradient of the ticket trained in isolation, and $\nabla_{\boldsymbol{\theta}_T}\mathcal{L}(\boldsymbol{\theta}_T, \boldsymbol{\theta}_{\text{adv}})$ is the initial gradient of the ticket when embedded in the larger network. This **adversarial overparameterization** (advOP) produces weights that maximally oppose the ticket's initial learning signal. This is a weak perturbation, since it is optimized to change only the first step of the ticket's optimization trajectory; after which training proceeds normally. As shown in Fig. 3b (details in Appendix B), advOP degrades performance of tickets embedded in slightly larger networks in both sparse and very sparse regimes.

We conclude from these two experiments that the presence of a winning ticket is *not* sufficient for successful training. Unlike lotteries, in neural networks the presence of other tickets (subnetworks) can alter the outcome. The reason is that successful learning requires not only the subnetwork to converge to the solution, but also all other network weights to converge to a state where they have no influence.

2.2. Lack of evidence for combinatorial scaling

"the chance of any given ticket winning is tiny, but if you buy enough of them you are certain to win, and the number of possible subnetworks increases exponentially as the power set of the set of connections" [Q38]

Wikipedia, 2024

The scaling property, perhaps the most common informal argument about the lottery metaphor, argues that overparameterization helps because wider networks contain combinatorially many subnetworks. For this argument to produce an exponential or combinatorial scaling law, two ingredients are needed: first, the presence of a winning subnetwork must be sufficient for the full network to succeed; second, the outcomes of different subnetworks must be at least approximately independent (both are in accordance with the core

mechanisms of a lottery). Let us for a moment assume that we can indeed treat subnetworks as independent and sufficient. We formalize these assumptions, derive the predicted scaling, and show that it disagrees with empirical data.

Given an initialization vector $I \in \mathbb{R}^P$, we denote by $R_I \in \{\texttt{fail}, \texttt{success}\}$ the training outcome of the full network, and $R_I^{(n)}$ the outcome of training subnetwork n in isolation. The sufficiency property states:

$$\exists n \mid \{R_I^{(n)} = \texttt{success}\} \implies \{R_I = \texttt{success}\} \quad (2)$$

Coincidentally, the contrapositive also holds:

$$\{R_I = \texttt{fail}\} \implies \{R_I^{(n)} = \texttt{fail}\}\, \forall n \quad (3)$$

When drawing an initialization from a distribution, we can denote the probability of success of the dense network as $\mathbb{P}(\{R_I = \texttt{success}\})$. Combined with the assumption that subnetwork outcomes are independent, this yields:

$$\mathbb{P}(\{R_I = \text{fail}\}) \leq \mathbb{P}\left(\bigcap_{n=1}^{N} \{R_I^{(n)} = \text{fail}\}\right) \quad (4)$$

$$\overset{\text{ind.}}{=} \prod_{n=1}^{N} \mathbb{P}\left(\{R_I^{(n)} = \text{fail}\}\right) \quad (5)$$

That is, the probability of failure should decay exponentially with the number of embedded subnetworks N. Because initializations are drawn from continuous distributions[Q4], the pool of potential tickets is infinite, justifying the use of sampling tickets with replacement. We test this super-exponential scaling prediction with a teacher-student setup similar to the one of Section 2.1. The smallest network capable of learning the task has $m=4$ hidden neurons. The number of size-4 dense subnetworks in a width-m network is $N_D(m) = \binom{m}{4}$; counting by edges gives $N_E(m) = \binom{md}{4d}$, where $d = d_{\text{in}} + d_{\text{out}}$. We estimate the baseline failure probability p_0 by training 4000 networks at $m=4$ (Appendix F). Eq. 5 sets an **upper-bound**: $\mathbb{P}(\{R_I = \texttt{fail}\}) \leq p_0^{N_D}$ (or $p_0^{N_E}$). Fig. 3c compares the measured probability of failure (black) to these predictions (grey traces) on a logarithmic scale. Both counting methods vastly overestimate the benefits of width: the predicted combinatorial decay bears no resemblance with the observed, much slower, decline.

Why does the combinatorial argument fail? The independence assumption treats subnetwork outcomes as separate draws, but they are not. Each neuron participates with the *same* weights in many subnetworks, creating structural correlations. Even disjoint tickets are dependent: consider a subnetwork $\hat{y}$ embedded in a larger network $\hat{y}^+$ obtained by adding neurons to a hidden layer k. The gradient of the loss with respect to activations of any other hidden layer $\mathbf{h}_l$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{h}_l} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial \hat{y}^+(\boldsymbol{\theta})}{\partial \mathbf{h}_l}\right)^\top \nabla_{\hat{y}^+} c(y_n, \hat{y}^+(\boldsymbol{\theta})) \quad (6)$$

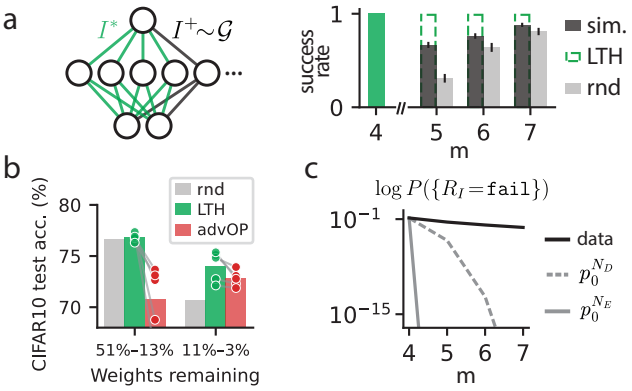

Figure 3. **Subnetworks do not respect the properties of a lottery.** ***a, left***: A winning ticket I^* (green) is embedded in a larger network with randomly initialized neurons I^+ (black). ***a, right***: Success rate as neurons are added: simulations (dark) show that embedding a winning ticket does not guarantee success, contrary to the sufficiency property of a lottery (dashed green). Random initialization of the full network shown for comparison (grey). ***b***: CIFAR-10 test accuracy of Conv-6 lottery tickets found by iterative magnitude pruning (green) vs. adversarial overparameterization (red) and sparse random initialization (grey), aggregated over two sets of sparsity levels. Adversarially regrown connections degrade the ticket's performance, confirming that subnetwork outcomes depend on the rest of the network. ***c***: Probability of failure to learn a target network: simulations (black) decay far more slowly than the predictions $p_0^{N_D}$ and $p_0^{N_E}$ (grey), showing that the combinatorial scaling of eq. 5 vastly overestimates the benefits of width. Additional interpretations of scaling are discussed in Appendix C.

Both the cost gradient $\nabla_{\hat{y}^+} c \in \mathbb{R}^{d_{\text{out}}}$ and the Jacobian $\frac{\partial \hat{y}^+}{\partial \mathbf{h}_l} \in \mathbb{R}^{d_{\text{out}} \times d_l}$ depend on the added neurons: the former via the output, the latter through forward (for $k<l$ and $k>l$) and backward (for $k>l$) paths. A subnetwork's training dynamics are therefore not a property of that subnetwork alone; the advOP experiment of Section 2.1 provides a direct demonstration. Hence, counting subnetworks as independent draws is not justified. Without independence, the joint probability in eq. 4 decomposes exactly as:

$$\prod_{n=1}^{N} \mathbb{P}\left(\{R_I^{(n)} = \text{fail}\}\right) + \Delta_N \quad (7)$$

where Δ_N collects all interaction terms between subnetwork outcomes: pairwise covariances, three-body correlations, and higher-order cumulants (McCullagh, 2018). The independence assumption sets $\Delta_N = 0$; but the gap in Fig. 3c shows that Δ_N is the leading term, not a small correction. A counting method that accounts for these interactions is not trivially available, and the advantage of overparameterization, real as it is (Fig. 3a), is far slower than any combinatorial argument would suggest.

In Appendix C we construct two scaling arguments that are *not combinatorial*: the first pretends subnetworks are independent if they share no node; e.g., if the minimal network is of size m and the full network of size $2m$, we count two subnetworks. The second pretends that each node is indepen-

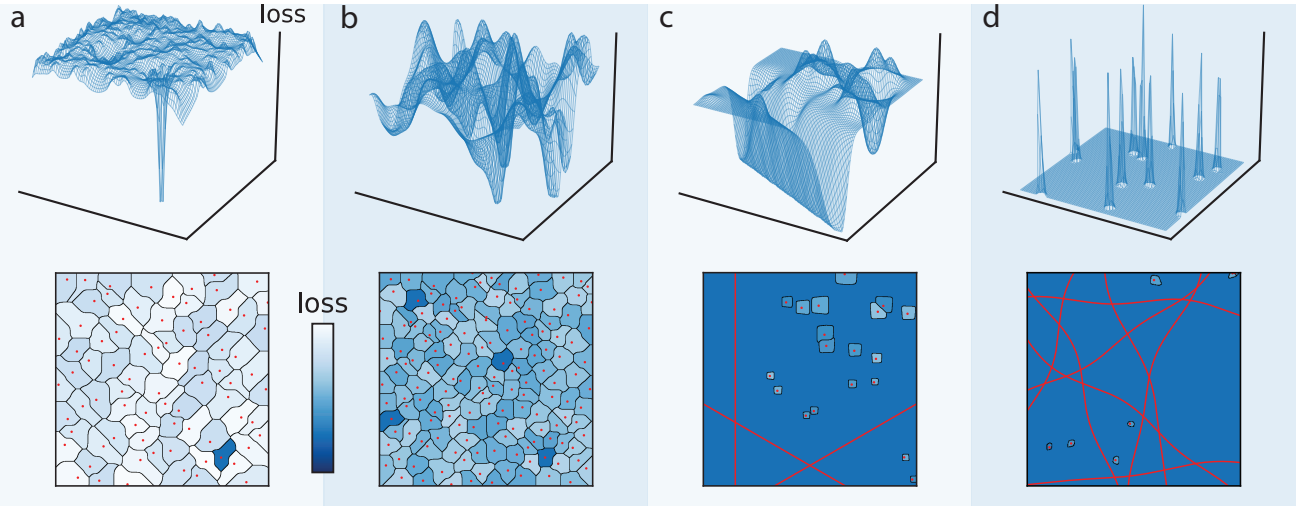

Figure 4. **Cartoon illustrations of loss landscapes as overparameterization increases.** ***Top***: surface plots conveying the difficulty of traversing different landscapes with gradient descent. ***Bottom***: parameter space partitioned by basins of attraction of landscape minima, blue color indicates the loss value at the minimum. Red dots denote isolated minima, red lines denote higher-dimensional minima manifolds. As width increases from left to right, escape dimensions are added: the landscape transitions from ***(a)*** rugged with many poor local minima and few good minima to ***(b)*** a more trainable one, where minima increase in number and decrease in loss. At two independent thresholds, the landscape undergoes further qualitative changes: ***(c)*** *if* the dataset is solvable with a finite number of neurons, *and* the network's width is above that number, linear sub-spaces of redundant zero-loss solutions appear (red lines). As width increases, these sub-spaces' basins of attraction tend to dominate the partition. Lastly, ***(d)*** when the network capacity allows perfect interpolation of the data, the partition is dominated by basins of attractions of manifolds of zero-loss minima (sketch inspired from Liu et al., 2022); any initialization (spike) is close to a zero-loss solution. While the illustrations in this figure are useful to convey a sense of types and proportions of minima in the landscape, they can only intuitively allude to some specific properties of the landscape. For example, they misrepresent the density of critical points with respect to the entire volume (top), or any notion of vicinity between basins (bottom).

dent (see also Wang et al., 2026b). Both scalings can be seen as reasonable heuristics (despite contradicting the analogy's definition of ticket and scaling); the specific teacher-student experimental data lies between the two. Whether the lottery metaphor can be rescued by adopting one of these alternatives is an **open question** that requires a significant reinterpretation and formalization of what constitutes a ticket.

We have seen that neither scaling nor sufficiency are empirically validated. This leaves us with the impression that the core point of comparison between networks and lotteries is missing, except for the generic concept of chance. A question then arises: why is the LTH empirically correct but the LTC not? Conceptually, pruning algorithms split the full network into two parts: (i) the successful sparse network that survives pruning and (ii) its complement (to be "masked out"). Because it relies on pruning *after* training, the LTH satisfies two conditions: first, that a sparse subnetwork is found by pruning, implying that many weights can be pruned because their dynamics lead them close to zero. Second, that after pruning the remaining network can be trained in isolation, as validated by experiments. However, the influential LTC only considers the second condition without the first one. As we have shown, both assumptions are necessary. In particular, the to-be-pruned subnetwork must be initialized such that it converges to zero, if not, it prevents convergence of the winning subnetwork.

3. Escape Dimensions Theory

We have seen that the explanation of overparameterization using lottery tickets is oversimplified because it neglects parameter interactions. Rather than adopting a mental model detached from the space where learning actually happens, we suggest to build our intuition on the loss landscape itself. Let us focus on ***what* the landscape looks like and *how* the width of a hidden layer changes it**. Throughout this section we will refer to Fig. 4, which offers two alternative drawings of the landscape as overparameterization increases: the top row uses familiar surface plots to convey the *difficulty of traversing* different landscapes with gradient descent, the bottom row conveys the *relative sizes* of basins of attraction and the dimensionality of minima. The high dimensionality of the parameter space makes any accurate visualization impossible; no single sketch tells the full story, but complementary views build a richer intuition (Welch et al., 2025). Addressing the *what* first, there exists a large body of theoretical work characterizing landscape properties related to convergence and trainability. The two classic pictures of loss landscapes correspond to two opposite regimes of width. In the classic regime (i.e., networks with "narrow" layers), landscapes are often pictured as rugged, with many poor local minima that can trap optimization. This landscape is dominated by high-loss local minima; good solutions exist but their basins of

attraction are small, making them difficult to find (Fig. 4a); in fact, finding optimal weights is NP–hard (Blum & Rivest, 1988; Livni et al., 2014). We can imagine the winning ticket initialization to be inside one of these rare good basins of attraction. The presence of bad local minima has been widely discussed theoretically (Auer et al., 1995; Zhou & Liang, 2018; Safran & Shamir, 2018; Venturi et al., 2019; Ding et al., 2019; Safran et al., 2021; Arjevani & Field, 2021) and observed empirically (Safran & Shamir, 2018; Martinelli et al., 2024; 2025). In the vastly overparameterized, or modern regime (Jacot et al., 2018; Du et al., 2018; Chizat et al., 2019; Belkin et al., 2019; Bahri et al., 2020), layers are wide and the network is capable of perfectly interpolating the training data (*interpolation* regime (Belkin, 2021)). The loss landscape, despite being non-convex, is benign: training almost always converges to global minima (Mei et al., 2018; Chizat & Bach, 2018; Du et al., 2019; Belkin, 2021; Liu et al., 2022). Fig. 4d gives an intuition: random initializations start at a high loss (spikes), but *training* quickly leads into a connected manifold of global minima (Cooper, 2018; Kuditipudi et al., 2019; van Meegen & Sompolinsky, 2025; Wang et al., 2026a); a "sea" of zero-loss solutions, pictured by Liu et al. (2022) as curved subspaces (Fig. 4d-bottom). Despite not being the only minima in the landscape (Ding et al., 2019), their basins of attraction cover the relevant portion of the parameter space (Safran & Shamir, 2016; Belkin, 2021). Taken together, these results show that the loss landscape looks very different in narrow and wide regimes. In particular, the loss landscape of a wide network is not just made of repeated instances of that of a narrow network (as Fig. 2 would suggest), but is *qualitatively* different.

3.1. Increasing width transforms minima into saddles

How can increasing width induce a qualitative change in the loss landscape? In the following, we introduce a mental picture for why overparameterization enables successful training: escape dimensions (Fig. 1). Escape dimensions provide a geometric perspective that is complementary to the convergence results aforementioned, by offering an intuitive picture of what changes when we say "*increasing width reshapes the loss landscape*". Contrary to the LTC with subnetworks, we ground our intuition in theoretically tractable objects: **critical points**, i.e., points in parameter space where the gradient is zero. In high-dimensional spaces, it is easy to lose geometric intuition: visualizations obtained from random projections (Goodfellow et al., 2015; Im et al., 2016; Li et al., 2018) are difficult to relate to mathematical structures relevant to learning. We can think of critical points as landmarks for orienting ourselves in high dimensions. These geometrical anchors are meaningful because they shape learning trajectories, *much like mountain passes shape hiking routes in mountainous terrain*.

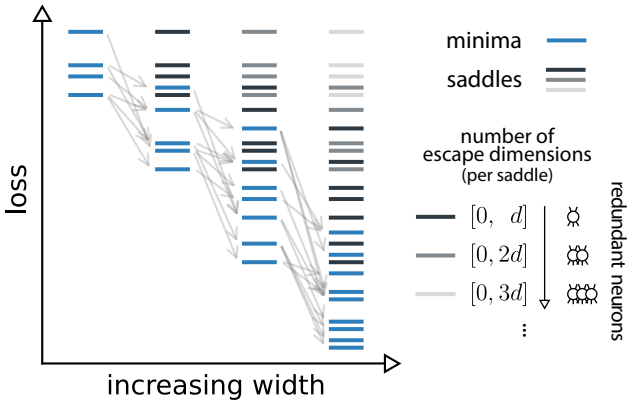

Figure 5. **Hierarchy of critical points across widths.** Each column shows the critical points of a network at a given width, ordered by their loss level (vertical axis). When a neuron is added at a minimum, a small perturbation enables gradient descent to reach one or more lower-loss minima (grey arrows). The former minima persist as saddle points in all wider networks, accumulating escape dimensions: each additional, redundant neuron can contribute up to d new escape dimensions, where d is the per-neuron parameter count.

The seminal work of Fukumizu & Amari (2000) lays the foundations for understanding how critical points evolve as width increases. In the following, we present the main results of this work and subsequent extensions, in an informal way that emphasizes their intuitive content. For formal statements and proofs, we refer to the original papers.
For a fixed dataset, the following holds:

Theorem 3.1. ***Critical point persistence*** *(Fukumizu & Amari, 2000, Ths. 1-2, informal): In two-layer, scalar output MLPs, every critical point $\hat{\theta}$ of a network with width m persists as a manifold of critical points $\hat{\theta}^+$ for any wider network of width $m^+ > m$:*

$$\nabla_\theta \mathcal{L}_m(\hat{\theta}) = 0 \implies \nabla_{\theta^+} \mathcal{L}_{m^+}(\hat{\theta}^+) = 0\,. \tag{8}$$

where the mapping $\hat{\theta} \to \hat{\theta}^+$ maintains functional equivalence between the two networks. In other words: the landscape of wider networks contains critical points that correspond to those of narrower networks. We can interpret this mapping as a transformation on the network that we call **neuron splitting**: a neuron is added by duplicating an existing neuron and re-scaling the output weights by γ and $(1-\gamma)$ between the two copies; Moreover, the continuous symmetry of the split maps each critical point $\hat{\theta}$ to a manifold of critical points $\hat{\theta}^+$ (Fukumizu et al., 2019; Simsek et al., 2021). Crucially, this transformation preserves the fixed points, but not their stability:

Theorem 3.2. ***Minima become saddles*** *(Fukumizu & Amari, 2000, Ths. 3-4, informal): Every local minimum of the smaller network becomes a saddle* in the larger network.*

where, for ease of exposition, we define saddle* as a critical manifold of non-strict saddle points that can occasionally contain non-strict local minima (referred to as *plateau*

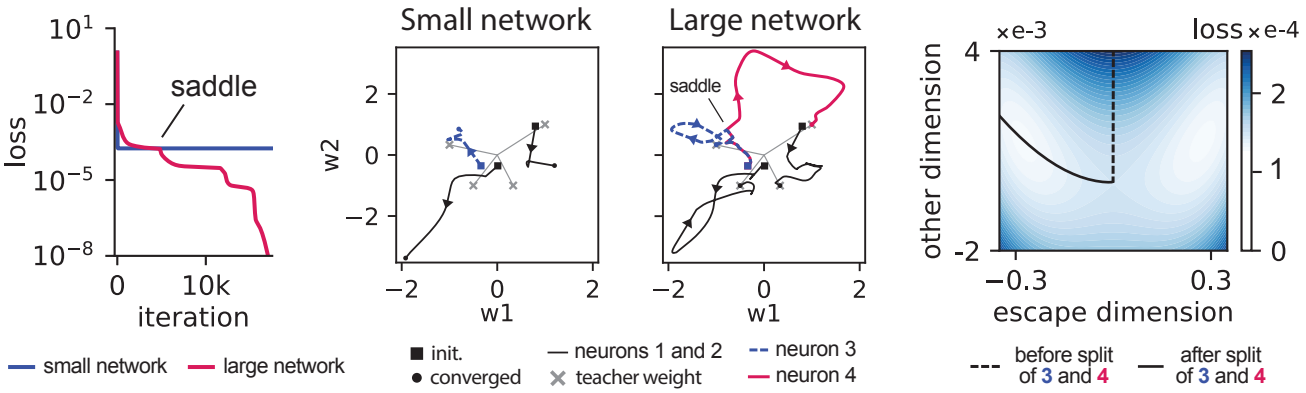

Figure 6. **Visualizing escape dimensions in a nonconvex landscape.** A 3 and 4-neuron two-layer MLPs are trained on a 4-neuron teacher. They share initializations except for a split neuron, whose two copies make neurons 3 and 4 in the wider network. ***Left***: loss curves: training of the large network (red) slows down around the same loss as the small network's minimum (blue). ***Middle***: input weights evolution: the split neurons (blue and red) follow identical trajectories up to the saddle point, after which they specialize and all neurons converge to the teacher weights (grey crosses). ***Right***: loss landscape of the large MLP around the loss level of the small MLP's minimum. The escape dimension is defined by the direction of most negative curvature, the other is the direction of highest curvature. The trajectory in parameter space is first orthogonal to the escape dimension (dashed line) until the critical point is surpassed via the escape dimension (continuous line). See Fig. E8 for an example of an overparameterized student network.

saddles (Martinelli et al., 2025), see below). We can intuitively interpret this result as follows: the addition of a neuron creates new **escape dimensions** in parameter space along which the loss can decrease, transforming the former minimum into a set of saddles (Fig. 1). Formally, we define an escape dimension $\mathbf{e}_i^-$, a direction in parameter space along which the Hessian of the loss (computed at a critical point $\hat{\theta}^+$) has a negative eigenvalue:

$$H\mathbf{e}_i^- = \lambda_i \mathbf{e}_i^-, \quad \lambda_i < 0 \tag{9}$$

where $H = \nabla^2_{\theta^+}\mathcal{L}(\hat{\theta}^+)$ is the Hessian of the loss $\mathcal{L}$ with respect to the parameters θ^+. Petzka & Sminchisescu, 2021 extended these results to MLPs of any depth:

Corollary 3.3. ***Extension to depth*** *(Petzka & Sminchisescu, 2021, informal): The addition of neurons to any layer of a feedforward deep network preserves Theorems 3.1 and 3.2.*

this can be trivially extended to any architecture where there are units to split (e.g. CNNs). Importantly, this mechanism gives rise to a **hierarchy of critical points** across widths (Fig. 5): a landscape of a network of width m^+ contains all critical points corresponding to those of a smaller network with width $m < m^+$; hinting at a combinatorial, recursive structure of minimum-to-saddle in the landscape (Simsek et al., 2021; Martinelli et al., 2025; Zhang et al., 2025). The literature characterizes the stability of critical points after neuron splitting (**saddles***) in more detail: some sections are locally attractive (Wei et al., 2008), containing non-strict saddle points referred to as *plateau saddles* (Martinelli et al., 2025); other parts are repulsive, corresponding to *strict saddles* (Wei et al., 2008; Fukumizu et al., 2019; Safran et al., 2021; Simsek et al., 2021; Zhang et al., 2021; Petzka & Sminchisescu, 2021; Wu et al., 2025; Martinelli et al., 2025). In realistic settings, these manifolds almost always contain at least a few negative curvature directions, enabling escape from critical regions via gradient-based optimization (Safran et al., 2021; Petzka & Sminchisescu, 2021; Martinelli et al., 2025). Note that nearly flat saddle regions can slow down or even trap learning dynamics under certain conditions (Inoue et al., 2003; Lee et al., 2016; Du et al., 2017; Chen et al., 2023). Independently from the theoretical literature, Liu et al. (2019); Wu et al. (2020a;b) re-discover neuron splitting and exploit escape dimensions for learning resource-efficient practical architectures and for continual learning problems. Under this new perspective, we see how the cartoon of Fig. 4d can form: as neurons are added, all critical points above zero loss accumulate escape dimensions until the landscape offers virtually no obstruction to gradient descent; (almost) any initialization can descend to the bottom (Fig. 4d).

To consolidate our intuitions on neuron splitting and escape dimensions, we guide the reader through the numerical example of a realistic, nonconvex landscape (Fig. 6). This example is meant to be a didactic illustration of the mechanism, we refer to the cited literature for more rigorous and general results. A smaller student network with three hidden neurons is trained to approximate a four-neuron teacher (details in Appendix F.3) and converges to a local minimum with MSE of $\sim 10^{-4}$ (1st panel). The two-dimensional input allows for direct visualization of weight trajectories (2nd and 3rd panels), where the three-neuron network visibly fails to reproduce the teacher (gray crosses). Starting from the *same initialization*, we apply a neuron-splitting operation to the original third neuron. The split replaces it with two duplicate neurons (3 and 4 in the "large network") that share identical input weights and have *nearly equal*

output weights ($\gamma = \frac{1}{2} + \epsilon$); their combined contribution exactly matches that of the original neuron. Because the split preserves the network's function, the initial training dynamics of the wider network closely mirror the original (Fig. 6, 3rd panel)[1]. As the wider network passes through the neighborhood of the smaller network's local minimum (traces crossing in the 1st panel), the perturbation between the two neurons is amplified, causing their trajectories to separate (3rd panel, top-left quadrant). This allows the wider network to escape the minimum (now saddle) inherited from the smaller model. The escape is made explicit by examining the loss landscape (Fig. 6, 4th panel) of the wider network near the parameter values matching the smaller network's converged loss: the trajectory first descends along the direction of maximum curvature (dashed line), then bends along a minimum-curvature direction (other dimension); an escape direction made possible by neuron addition.

In summary: adding neurons creates escape dimensions that transform local minima into saddles, enabling gradient descent to find better solutions (Fig. 4b-top). Escaping can only lead to lower losses, consistent with the known parallel to spin-glass models where minima proliferate increasingly near the bottom as system size grows (Choromanska et al., 2015). In large networks, Draxler et al. (2018) show that functionally different (Yunis et al., 2022) low-loss minima are separated by negligible barriers. Consistent with our picture, the saddle points separating these *distinct* basins are inherited from slightly smaller networks, and in sufficiently wide networks even these saddles sit at very low losses (Fig. 4b-bottom).

3.2. From local transformations to global structure

How do escape dimensions affect the overall geometry of the loss landscape? Simsek et al. (2021) provide a global, combinatorial perspective by counting all saddle subspaces attributable to neuron splitting as a function of width. These are named *symmetry-induced* saddles. When a task is realizable by a network of width r, any network of width $m > r$ has more capacity than needed to solve the task (i.e., it is overparameterized or overspecified). The redundant parameters introduce continuous symmetries (Martinelli et al., 2024), so that zero-loss solutions are no longer isolated points but linear subspaces. These subspaces connect into larger manifolds of global minima in parameter space. Symmetry-induced saddles have a similar structure, but less types of transformations can guarantee the persistence of criticality (Fukumizu & Amari, 2000). Both saddle and global minima subspaces proliferate combinatorially with increasing width, so growth rates alone are not informative. What changes is the relative number: in the mildly overparameterized regime (Safran & Shamir, 2018), the geometry is dominated by saddles inherited from narrower networks (Fig. 4b). However, as the network width increases further, a shift occurs: for $m \gtrsim 1.5r$, the number of global minima subspaces exceeds that of saddle subspaces. In other words, the landscape, defined by its critical points, is *dominated* by global minima. In Appendix D we describe how this count unfolds, and Fig. D6 illustrates how the ratio of saddle to minima subspaces changes with width. In this regime, it becomes increasingly likely for optimization to reach a global minimum (Fig. 4c), consistent with empirical observations showing that bad minima become less prevalent as width increases (Safran & Shamir, 2018; Martinelli et al., 2024). It is important to note that this transition may occur before or after the interpolation threshold, or never, depending on the dataset. Escape dimensions emphasize the mental picture that, in a very wide network, *(almost) all paths lead to the global minimum.*

We collect all of the above results under the umbrella term **Escape Dimensions Theory**; where each word is chosen to build a specific intuition: **escape** evokes how gradient descent escapes bad minima as they transform into saddles; **dimensions**, rather than directions, emphasizes the change in dimensionality of parameter space; and **theory** highlights that this intuitive local mechanism is backed by rigorous results (Fukumizu & Amari, 2000), and fits naturally within broader scales of analysis (Mei et al., 2018; Simsek et al., 2021; Belkin, 2021).

The focus on the loss landscape also clarifies the role of lottery tickets: a winning ticket can be understood as an initialization of a *narrow network* that lies inside one of the rare basins with small attraction radius in the rugged landscape (Larsen et al., 2022). A winning ticket is found by: first, by optimizing a *wide network* exploiting the benign structure of its landscape (Fig. 4c,d); second, pruning dimensions that are not necessary for convergence. The latter supported by experimental evidence showing that winning tickets converge to solutions functionally similar to the original dense network (Evci et al., 2022). Escape Dimensions explain why optimization becomes reliable as width increases.

4. Alternative views and conclusions

An alternative direction towards reasoning via subgraphs in neural networks is the notion of *neural race reduction* (Saxe et al., 2022; Jarvis et al., 2025), that considers ReLU subnetworks as deep linear networks activated by individual input samples. Contemporary to the writing of this piece, Pinson (2026) made the association between the LTC and the neural race reduction, arguing that individual ReLU units (a specific instance of a subnetwork) can be analysed in isolation. They assume a very specific condition in the dynamics: that the set of data-points activating the unit

[1] dynamics cannot be exactly identical: the output weight magnitude affects the input weights learning speed (see Appendix E).

never changes along the analyzed interval[2]; as well as vanishing initialization. Neurons race against each other to be the first to learn a specific feature, then the output weights of the winner grow in norm and loss decreases. While not dismissing the lottery metaphor entirely, they "correct it" towards the concept of race between units. We note that other theoretical works on training dynamics cite the LTC as inspiration (Atanasov et al., 2021; Boursier et al., 2022; Edelman et al., 2023; Arous et al., 2024; Boursier & Flammarion, 2025; Pinson, 2026), but operate under a different definition of tickets (units) than the one used by the field and in the original formulation (subgraphs). Moreover, the settings in which some of the LTC properties may hold are ones that allow independent treatment of units (e.g. orthogonal inputs, linear activations, shallow networks). In Appendix C, we also provide similar alternative heuristics that better align with the empirical observations. We value these as alternative, promising points of view to intuitively understand the dynamics of learning; a much broader challenge than the one we address here. Our counterexamples show that the lottery properties, as invoked by some, are not universal; and we do not claim they never hold. Establishing the conditions under which they do, and formalizing what constitutes a ticket, what scaling law it predicts, and under what assumptions, remains an open problem. We suggest that rescuing the lottery metaphor requires a rigorous redefinition of ticket and explicit conditions for when the scaling property of the analogy holds.

Since mental images are important, we propose an alternative: **Escape Dimensions**, backed by theoretical and empirical results from the literature on loss landscapes. We note that escape dimensions are a general phenomenon, that does not require specific assumptions on architecture, data, depth, or dynamics. For clarifying the paradox of why seemingly redundant, large networks are necessary for training, escape dimensions are enough, without invoking training dynamics that nearly touch saddles. Another important aspect not to dismiss is that we only addressed the question of trainability, leaving aside questions of generalization, benign overfitting, and implicit bias (Wilson, 2025); or whether overparameterized solutions generalize well (see rich vs. lazy regimes, e.g. Chizat et al., 2019; Woodworth et al., 2020). While escape dimensions are a strong candidate to explain the benefits of overparameterization, they are not the only actor at play, and their full implications remain to be explored.

We have argued that using lottery tickets in scientific discourse to explain overparameterization is misleading, because it is not possible to separate subnetworks from the rest of the system. Our representative selection of quotes (Tables A1, A2) indicates that this oversimplified interpretation is shared by parts of the community. We believe that many of the quoted authors did not intend to convey precise or rigorous statements, but invoking the LTH in a way that appeals to mechanisms of sufficiency, scaling, or independence is problematic. Unfortunately, the name lottery itself, the metaphor, is associated with these properties. Hence, the cautionary message of our paper.

As in all sciences of complex systems, understanding requires a reductionist approach. The subnetworks in the LTH are one example of such reduction, and this likely contributed to its popularity. At the time, the lottery metaphor was a brilliant and intuitive way to convey the authors' conjectures on the mechanisms of deep learning. As the field evolved and theories became more precise, this analogy has proven not to hold in general. In our opinion, the suggestive mental picture that lottery tickets evoke should not be reinforced in the literature nor taught in classes (unless rigorously refined). We suggest that the field should focus on understanding loss landscapes and training dynamics in the transition between the classic and vastly overparameterized regimes (Fig. 4b,c), and close the circle with known theories of convergence and generalization in the interpolation regime (Mei et al., 2018; Belkin, 2021; Liu et al., 2022). In our view, critical points are promising conceptual units of reduction to intuitively understand network computations. The study of saddle-to-saddle dynamics, and its interpretable outcomes are one such example (Gidel et al., 2019; Jacot et al., 2021; Pesme & Flammarion, 2023; Abbe et al., 2023; Boursier & Flammarion, 2025; Kunin et al., 2025; Martinelli et al., 2025; Zhang et al., 2025; Marchetti et al., 2026).

[2] we refer to Pinson's gating $\vec{g}_\alpha$ assumed to be constant.

Acknowledgements

We thank Louis Pezon, Alexander Van Meegen, Sophia Becker, Zihan Wu for the fruitful discussions. This work was funded by the Swiss National Science Foundation grant 200021-236436.

References

Abbe, E., Adsera, E. B., and Misiakiewicz, T. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Abnar, S., Dehghani, M., and Zuidema, W. Transferring inductive biases through knowledge distillation. *arXiv preprint arXiv:2006.00555*, 2020.

Arjevani, Y. and Field, M. Symmetry & critical points for a model shallow neural network. *Physica D: Nonlinear Phenomena*, 427:133014, 2021.

Arous, G. B., Gheissari, R., and Jagannath, A. High-dimensional limit theorems for sgd: Effective dynam-

ics and critical scaling. *Communications on Pure and Applied Mathematics*, 77(3):2030–2080, 2024.

Atanasov, A., Bordelon, B., and Pehlevan, C. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2021.

Auer, P., Herbster, M., and Warmuth, M. K. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 8, 1995.

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. Statistical mechanics of deep learning. *Annual review of condensed matter physics*, 11(1):501–528, 2020.

Bai, Y., Wang, H., Ma, X., Zhang, Y., Tao, Z., and Fu, Y. Parameter-efficient masking networks. *Advances in Neural Information Processing Systems*, 35:10217–10229, 2022a.

Bai, Y., Wang, H., TAO, Z., Li, K., and Fu, Y. Dual lottery ticket hypothesis. In *International Conference on Learning Representations*, 2022b.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Berner, J., Grohs, P., Kutyniok, G., and Petersen, P. The modern mathematics of deep learning. *Mathematical Aspects of Deep Learning*, pp. 1, 2022.

Blum, A. and Rivest, R. Training a 3-node neural network is np-complete. *Advances in neural information processing systems*, 1, 1988.

Boursier, E. and Flammarion, N. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025.

Boursier, E., Pillaud-Vivien, L., and Flammarion, N. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

Brea, J., Martinelli, F., Şimşek, B., and Gerstner, W. Mlpgradientflow: going with the flow of multilayer perceptrons (and finding minima fast and accurately). *arXiv preprint arXiv:2301.10638*, 2023.

Burkholz, R. Convolutional and residual networks provably contain lottery tickets. In *International Conference on Machine Learning*, pp. 2414–2433. PMLR, 2022.

Burkholz, R., Laha, N., Mukherjee, R., and Gotovos, A. On the existence of universal lottery tickets. In *International Conference on Learning Representations*, 2021.

Casper, S., Boix, X., D'Amario, V., Guo, Z., Schrimpf, M., Vinken, K., and Kreiman, G. Frivolous units: Wider networks are not really that wide. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Chen, F., Kunin, D., Yamamura, A., and Ganguli, S. Stochastic collapse: How gradient noise attracts sgd dynamics towards simpler subnetworks. *Advances in Neural Information Processing Systems*, 36:35027–35063, 2023.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Wang, Z., and Carbin, M. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33:15834–15846, 2020.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Carbin, M., and Wang, Z. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16306–16316, 2021a.

Chen, T., Sui, Y., Chen, X., Zhang, A., and Wang, Z. A unified lottery ticket hypothesis for graph neural networks. In *International conference on machine learning*, pp. 1695–1706. PMLR, 2021b.

Chen, X., Zhang, Z., Sui, Y., and Chen, T. Gans can play lottery tickets too. *International Conference on Learning Representations*, 2021c.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pp. 192–204. PMLR, 2015.

Cooper, Y. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

Cowley, B. R., Stan, P. L., Pillow, J. W., and Smith, M. A. Compact deep neural network models of the visual cortex. *Nature*, pp. 1–8, 2026.

da Cunha, A., Natale, E., and Viennot, L. Proving the lottery ticket hypothesis for convolutional neural networks. In *International Conference on Learning Representations*, 2022.

Díaz-Faloh, C. and Mulet, R. Diluting restricted boltzmann machines. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(3):033401, 2026.

Ding, T., Li, D., and Sun, R. Sub-optimal local minima exist for neural networks with almost all non-linear activations. *arXiv preprint arXiv:1911.01413*, 2019.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30, 2017.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Dubois, Y., Kiela, D., Schwab, D. J., and Vedantam, R. Learning optimal representations with the decodable information bottleneck. *Advances in Neural Information Processing Systems*, 33:18674–18690, 2020.

Eco, U. *Semiotics and the Philosophy of Language*, volume 398. Indiana University Press, 1986. See esp. ch. 3: p. 102 on Aristotle and the cognitive/instructive function of metaphor; §§3.11.4–3.11.5, pp. 118–124, on the interpretative work demanded by 'open' metaphors. See also https://en.wikipedia.org/wiki/Metaphor#As_a_type_of_comparison.

Edelman, B., Goel, S., Kakade, S., Malach, E., and Zhang, C. Pareto frontiers in deep feature learning: Data, compute, width, and luck. *Advances in Neural Information Processing Systems*, 36:48021–48034, 2023.

Evci, U., Gale, T., Menick, J., Castro, P. S., and Elsen, E. Rigging the lottery: Making all tickets winners. In *International conference on machine learning*, pp. 2943–2952. PMLR, 2020.

Evci, U., Ioannou, Y., Keskin, C., and Dauphin, Y. Gradient flow in sparse neural networks and how lottery tickets win. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6577–6586, 2022.

Ferbach, D., Tsirigotis, C., Gidel, G., and Bose, J. A general framework for proving the equivariant strong lottery ticket hypothesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Flint, A. Understanding the lottery ticket hypothesis. LessWrong Forum post, 2021. URL https://www.lesswrong.com/posts/dpzLqQQSs7XRacEfK/understanding-the-lottery-ticket-hypothesis.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International Conference on Learning Representations*, 2019.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

Fritz AI Blog. The lottery ticket hypothesis explained simply. Blog post, 2023. URL https://fritz.ai/blog/lottery-ticket-hypothesis-explained-simply.

Fukumizu, K. and Amari, S.-i. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

Fukumizu, K., Yamaguchi, S., Mototake, Y.-i., and Tanaka, M. Semi-flat minima and saddle points by embedding neural networks to overparameterization. *Advances in neural information processing systems*, 32, 2019.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. Qualitatively characterizing neural network optimization problems. *International Conference on Machine Learning*, 2015.

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Huszár, F. The lottery ticket hypothesis – paper recommendation. Blog post, inFERENCe.vc, 2018. URL https://www.inference.vc/the-lottery-ticket-hypothesis/.

Im, D. J., Tao, M., and Branson, K. An empirical analysis of the optimization of deep network loss surfaces. *arXiv preprint arXiv:1612.04010*, 2016.

Inoue, M., Park, H., and Okada, M. On-line learning theory of soft committee machines with correlated hidden units–steepest gradient descent and natural gradient descent–. *Journal of the Physical Society of Japan*, 72(4):805–810, 2003.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Jarvis, D., Klein, R., Rosman, B., and Saxe, A. M. Make haste slowly: A theory of emergent structured mixed selectivity in feature learning relu networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeffares, A. and van der Schaar, M. Position: Not all explanations for deep learning phenomena are equally valuable. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Kim, S., Hooper, C., Wattanawong, T., Kang, M., Yan, R., Genc, H., Dinh, G., Huang, Q., Keutzer, K., Mahoney, M. W., Shao, S., and Gholami, A. Full stack optimization of transformer inference. In *Architecture and System Support for Transformer Models (ASSYST @ISCA 2023)*, 2023.

Kim, Y., Li, Y., Park, H., Venkatesha, Y., Yin, R., and Panda, P. Exploring lottery ticket hypothesis in spiking neural networks. In *European Conference on Computer Vision*, pp. 102–120. Springer, 2022.

Kokotajlo, D. Does the lottery ticket hypothesis suggest the scaling hypothesis? LessWrong Forum post, 2020. URL https://www.lesswrong.com/posts/wFJqi75y9eW8mf8TR/does-the-lottery-ticket-hypothesis-suggest-the-scaling.

Koster, N., Grothe, O., and Rettinger, A. Signing the supermask: Keep, hide, invert. In *International Conference on Learning Representations*, 2022.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019.

Kumar, A. and Natale, E. Quantization vs pruning: Insights from the strong lottery ticket hypothesis. *arXiv preprint arXiv:2508.11020*, 2025.

Kunin, D., Marchetti, G. L., Chen, F., Karkada, D., Simon, J. B., DeWeese, M. R., Ganguli, S., and Miolane, N. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Lam, P., Zhang, H., Chen, N., and Sisman, B. Epic tts models: Empirical pruning investigations characterizing text-to-speech models. In *Proc. Interspeech 2022*, pp. 823–827, 2022.

Lange. The lottery ticket hypothesis: A survey. Blog post, 2020. URL https://roberttlange.com/posts/2020/06/lottery-ticket-hypothesis/.

Larsen, B. W., Fort, S., Becker, N., and Ganguli, S. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. In *International Conference on Learning Representations*, 2022.

Lê, M. T., Wolinski, P., and Arbel, J. Efficient neural networks for tiny machine learning: A comprehensive review. *ACM Transactions on Intelligent Systems and Technology*, 2023.

LeCun, Y., Denker, J., and Solla, S. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y., and Li, H. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets. *arXiv preprint arXiv:2008.03371*, 2020.

Li, A. C., Tian, Y., Chen, B., Pathak, D., and Chen, X. On the surprising effectiveness of attention transfer for vision transformers. *Advances in Neural Information Processing Systems*, 37:113963–113990, 2024.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Liu, B., Zhang, Z., He, P., Wang, Z., Xiao, Y., Ye, R., Zhou, Y., Ku, W.-S., and Hui, B. A survey of lottery ticket hypothesis. *arXiv preprint arXiv:2403.04861*, 2024.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Liu, Q., Wu, L., and Wang, D. Splitting steepest descent for growing neural architectures. *arXiv preprint arXiv:1910.02366*, 2019.

Livni, R., Shalev-Shwartz, S., and Shamir, O. On the computational efficiency of training neural networks. *Advances in neural information processing systems*, 27, 2014.

Lord, J. How ai researchers accidentally discovered that everything they thought about learning was wrong. Blog post Nearly Right, 2025. URL https://nearlyright.com/how-ai-researchers-accidentally-discovered-that-everything-they-thought-about-learning-was-wrong/.

Ma, X., Yuan, G., Shen, X., Chen, T., Chen, X., Chen, X., Liu, N., Qin, M., Liu, S., Wang, Z., et al. Sanity checks for lottery tickets: Does your winning ticket really win the jackpot? *Advances in Neural Information Processing Systems*, 34:12749–12760, 2021.

Malach, E., Yehudai, G., Shalev-shwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: pruning is all you need. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6682–6691, 2020.

Mannelli, S. S., Ivashynka, Y., Saxe, A., and Saglietti, L. Tilting the odds at the lottery: the interplay of overparameterisation and curricula in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(11): 114001, 2024.

Marchetti, G. L., Kunin, D., Myers, A., Acosta, F., and Miolane, N. Sequential group composition: A window into the mechanics of deep learning. *arXiv preprint arXiv:2602.03655*, 2026.

Martinelli, F., Dellaferrera, G., Mainar, P., and Cernak, M. Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8544–8548. IEEE, 2020.

Martinelli, F., Simsek, B., Gerstner, W., and Brea, J. Expand-and-cluster: Parameter recovery of neural networks. In *International Conference on Machine Learning*, pp. 34895–34919. PMLR, 2024.

Martinelli, F., van Meegen, A., Simsek, B., Gerstner, W., and Brea, J. Flat channels to infinity in neural loss landscapes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

McCullagh, P. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Morcos, A., Yu, H., Paganini, M., and Tian, Y. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019.

Mostafa, H. and Wang, X. Parameter-efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

Natale, E., Ferré, D., Giambartolomei, G., Giroire, F., and Mallmann-Trenn, F. On the sparsity of the strong lottery ticket hypothesis. *Advances in Neural Information Processing Systems*, 37:40565–40592, 2024.

Orseau, L., Hutter, M., and Rivasplata, O. Logarithmic pruning is all you need. *Advances in Neural Information Processing Systems*, 33:2925–2934, 2020.

Otsuka, H., Chijiwa, D., Okoshi, Y., Fujiki, D., Takeuchi, S., and Motomura, M. The strong lottery ticket hypothesis for multi-head attention mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24639–24647, 2026.

Palm, R. B., Najarro, E., and Risi, S. Testing the genomic bottleneck hypothesis in hebbian meta-learning. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pp. 100–110. PMLR, 2021.

Pensia, A., Rajput, S., Nagle, A., Vishwakarma, H., and Papailiopoulos, D. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in neural information processing systems*, 33: 2599–2610, 2020.

Pesme, S. and Flammarion, N. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 36:7475–7505, 2023.

Petzka, H. and Sminchisescu, C. Non-attracting regions of local minima in deep and wide neural networks. *Journal of Machine Learning Research*, 22(143):1–34, 2021.

Pinson, H. It's not a lottery, it's a race: Understanding how gradient descent adapts the network's capacity to the task. *arXiv preprint arXiv:2602.04832*, 2026.

Pondsiders. Sparse networks and lottery winners. Blog post, 2025. URL https://pondsiders.github.io/machine%20learning/sparse-networks/.

Prasanna, S., Rogers, A., and Rumshisky, A. When bert plays the lottery, all tickets are winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3208–3229, 2020.

Princeton-CS-598D. Model compression–the pruning approaches. Slides from lecture Computer Science 598D Overcoming Intractability in Machine Learning, 2020. URL https://www.cs.princeton.edu/courses/archive/spring21/cos598D/lectures/pruning.pdf.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11893–11902, 2020.

Reed, R. Pruning algorithms-a survey. *IEEE transactions on Neural Networks*, 4(5):740–747, 1993.

Renda, A., Frankle, J., and Carbin, M. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020.

Safran, I. and Shamir, O. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pp. 774–782. PMLR, 2016.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018.

Safran, I. M., Yehudai, G., and Shamir, O. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In *Conference on Learning Theory*, pp. 3889–3934. PMLR, 2021.

Saxe, A., Sodhani, S., and Lewallen, S. J. The neural race reduction: Dynamics of abstraction in gated networks. In *International Conference on Machine Learning*, pp. 19287–19309. PMLR, 2022.

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., and Brea, J. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, pp. 9722–9732. PMLR, 2021.

Singh, S. and Bhatele, A. Exploiting sparsity in pruned neural networks to optimize large model training. In *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 245–255. IEEE, 2023.

Tripp, C. E., Perr-Sauer, J., Hayne, L., Lunacek, M., and Gafur, J. The butter zone: An empirical study of training dynamics in fully connected neural networks. *arXiv preprint arXiv:2207.12547*, 2022.

van Meegen, A. and Sompolinsky, H. Coding schemes in neural networks learning classification tasks. *Nature Communications*, 16(1):3354, 2025.

Venturi, L., Bandeira, A. S., and Bruna, J. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133): 1–34, 2019.

Verma, V. K., Mehta, N., Si, S., Henao, R., and Carin, L. Pushing the efficiency limit using structured sparse convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6503–6513, 2023.

Vischer, M., Lange, R. T., and Sprekeler, H. On lottery tickets and minimal task representations in deep reinforcement learning. In *International Conference on Learning Representations*, 2022.

Wang, B., Johnston, J., and Fusi, S. A mathematical theory for understanding when abstract representations emerge in neural networks. *ArXiv*, pp. arXiv–2510, 2026a.

Wang, C., Zhang, G., and Grosse, R. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.

Wang, H.-Y., Luo, D., Poggio, T., Chuang, I. L., and Ziyin, L. A universal compression theory for lottery ticket hypothesis and neural scaling laws. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Wei, H., Zhang, J., Cousseau, F., Ozeki, T., and Amari, S.-i. Dynamics of learning near singularities in layered networks. *Neural computation*, 20(3):813–843, 2008.

Welch, S., Baskin, S., and Gundu, P. *The Welch Labs Illustrated Guide to AI*. Welch Labs LLC, 2025.

Wentworth, J. Understanding the lottery ticket hypothesis. Alignment Forum post, 2021. URL https://www.alignmentforum.org/posts/dpzLqQQSs7XRacEfK/understanding-the-lottery-ticket-hypothesis.

Wikipedia. Lottery ticket hypothesis. Wikipedia, The Free Encyclopedia, 2024. URL https://en.wikipedia.org/wiki/Lottery_ticket_hypothesis. Note: the quote was first added in the revision of 22 December 2024.

Wilson, A. G. Position: Deep learning is not so mysterious or different. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., and Farhadi, A. Supermasks in superposition. *Advances in neural information processing systems*, 33:15173–15184, 2020.

Wu, F. Z., Simsek, B., and Ged, F. G. Loss landscape of shallow relu-like neural networks: Stationary points, saddle escape, and network embedding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wu, L., Liu, B., Stone, P., and Liu, Q. Firefly neural architecture descent: a general approach for growing neural networks. *Advances in neural information processing systems*, 33:22373–22383, 2020a.

Wu, L., Ye, M., Lei, Q., Lee, J. D., and Liu, Q. Steepest descent neural architecture optimization: Escaping local optimum with signed neural splitting. *arXiv preprint arXiv:2003.10392*, 2020b.

Yu, H., Edunov, S., Tian, Y., and Morcos, A. S. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp. In *International Conference on Learning Representations*, 2020.

Yunis, D., Patel, K. K., Savarese, P. H. P., Vardi, G., Frankle, J., Walter, M., Livescu, K., and Maire, M. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Zhang, Y., Zhang, Z., Luo, T., and Xu, Z. J. Embedding principle of loss landscape of deep neural networks. *Advances in neural information processing systems*, 34: 14848–14859, 2021.

Zhang, Y., Saxe, A., and Latham, P. E. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. *arXiv preprint arXiv:2512.20607*, 2025.

Zhou, H., Lan, J., Liu, R., and Yosinski, J. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in neural information processing systems*, 32, 2019.

Zhou, Y. and Liang, Y. Critical points of linear neural networks: Analytical forms and landscape properties. In *International Conference on Learning Representations*, 2018.

A. Evidence of the problematic LTH interpretation

To obtain a sample of how widespread the interpretations described in Section 1 are, we employed two types of searches on the web: 1) systematic search on arXiv, and 2) search through LLMs standard or "deep research" abilities. Both were followed by human verification of the results. The results of both searches are reported in Tables A1 and A2.

Table A1. Quotes that suggest or reinforce the use of the lottery ticket analogy to explain the success of overparameterization. The first four rows are quotes from the original paper (Frankle & Carbin, 2019), that provide an interpretative key of the metaphor.

	Quote	Paper
1	"A key question, then, is whether the presence of a winning ticket is necessary or sufficient for SGD to optimize a neural network to a particular test accuracy."	Frankle & Carbin, 2019
2	"We conjecture (but do not empirically show) that SGD seeks out and trains a well-initialized subnetwork."	Frankle & Carbin, 2019
3	"By this logic, overparameterized networks are easier to train because they have more combinations of subnetworks that are potential winning tickets."	Frankle & Carbin, 2019
4	"We designate these trainable subnetworks, $f(x; m \odot \theta_0)$, winning tickets, since those that we find have won the initialization lottery with a combination of weights and connections capable of learning."	Frankle & Carbin, 2019
5	"and hypothesized that the role of overparameterization is to provide a large number of candidate subnetworks, thereby increasing the likelihood that one of these subnetworks will have the necessary structure and initialization needed for effective learning."	Mostafa & Wang, 2019
6	"the lottery ticket hypothesis, which argues that the probability of sampling a lucky, trainable sub-network initialization grows with network size due to the combinatorial explosion of available sub-network initializations"	Morcos et al., 2019
7	"Hence, they showed that the pre-trained weights are not necessary, only the pruned architecture and the corresponding initial weight values."	Wang et al., 2020
8	"they suggest that wide networks may perform as well as or better than thin ones because they 'buy more lottery tickets' and more reliably contain these fortuitously initialized subnetworks."	Casper et al., 2021
9	"Empirically, many people have found that bigger models are easier to train (often explained with the 'lottery ticket hypothesis')"	Abnar et al., 2020
10	"Surprisingly, VL+ performs better than VL , which might be because larger networks can help optimization of sub-networks VL $\subseteq$ VL+ as suggested by the Lottery Ticket Hypothesis"	Dubois et al., 2020
11	"This polynomial bound already tells us that unpruned networks contain many 'winning tickets' even without training. Then it is natural to ask: could the most important task of gradient descent be pruning?"	Orseau et al., 2020
12	"The hypothesis is that overparameterized neural networks are more likely to contain sub-network, which are initialized in such a way that they can be effectively optimized to solve the task."	Palm et al., 2021
13	"If you start with a very overparameterized network, probability theory gives the network much higher chances to include a better subnetwork than a very small one."	Koster et al., 2022
14	"This hypothesis suggests that "SGD seeks out and trains a well-initialized subnetwork" and that "overparameterized networks are easier to train because they have more combinations of subnetworks that are potential winning tickets."	Tripp et al., 2022
15	"To explain why this happens, the lottery ticket hypothesis has been proposed [6], which posits that a randomly initialized model contains subnetworks that are especially suited for training on the given task (i.e. winning tickets), and that large models exponentially increase the chance of getting winning tickets."	Lam et al., 2022

Table A1. continued

	Quote	Paper
16	"the network redundancy also ensures a large random network contains a huge number of possible subnetworks, thus, carefully selecting a specific subnetwork should obtain promising performances. This point of view has been proved by [31, 44]."	Bai et al., 2022a
17	"Similar to LTH, there is compelling evidence [38, 39, 14, 13, 2, 1, 45] suggesting that overparameterization is not essential for high test accuracy, but is helpful to find a good initialization for the network [30, 65]."	Verma et al., 2023
18	"They theorize that in an overparameterized network, it is this subnetwork that effectively ends up being trained, thus preventing over-fitting. They also present a simple algorithm to identify this subnetwork."	Singh & Bhatele, 2023
19	"This may be due to the fact that having redundant parameters from the beginning of the training may make the loss landscape easier to optimize [139]; or it may be related to the increase in the likelihood of obtaining a "lottery ticket" [67]."	Kim et al., 2023
20	"We show, theoretically and experimentally, that sparse initialization and increasing network width yield significant improvements in sample efficiency in this setting. Here, width plays the role of parallel search: it amplifies the probability of finding "lottery ticket" neurons, which learn sparse features more sample-efficiently."	Edelman et al., 2023
21	"there exists a sparse subnetwork (winning ticket) that can be trained from scratch [...] In this view, a large model has a greater chance of containing a good subnetwork."	Lê et al., 2023
22	"A few possible explanations: - Lottery tickets [...] The full output of the network is the average of many different circuits, with significant interference from non-linear interaction. Some of these circuits are systematically useful to reducing loss, but most aren't. Gradients for useless circuits will have zero mean, while gradients for useful circuits will have non-zero mean, with a lot of noise. SGD reinforces relevant circuits and suppresses useless ones, so circuits will gradually form"	Nanda et al., 2023
23	"Roughly speaking, the lottery ticket hypothesis proposes that the reason for the success of overparametrized networks is that they give more attempts for a sufficiently expressive subnetwork to be initialized well, and succeed at the task on its own"	Arous et al., 2024
24	"Frankle and Carbin [15] further conjecture that overparameterization improves performance because larger models contain exponentially more sparse subnetworks in superposition and are thus more likely to contain a "winning ticket" - a hypothesis supported by subsequent empirical and theoretical work."	Li et al., 2024
25	"This overparameterisation phenomenology provides a clear example of the lottery ticket hypothesis (Frankle & Carbin,2019). The idea is that one of the advantages of training highly overparameterised neural networks comes from the increased likelihood of randomly sampling a well-initialised sub-network that is sufficient to solve the learning problem at hand. In simple words, collecting more lottery tickets will certainly enhance the chance of finding the winning one."	Mannelli et al., 2024
26	"It also resonates with the Lottery Ticket Hypothesis, which proposes that successful training depends on the presence of a small sub-network—"the winning ticket"—within a larger model."	Díaz-Faloh & Mulet, 2026
27	"Evidence from the original work, supported by subsequent empirical (e.g. Zhou et al., 2019) and theoretical (e.g. Malach et al., 2020) studies, strongly indicates that the phenomenon exists broadly and the hypothesis holds."	Jeffares & van der Schaar, 2025
28	"While directly identifying lottery tickets before training remains impractical, the underlying ideas have significantly influenced our explanatory theories of sparsity, pruning, and network efficiency."	Jeffares & van der Schaar, 2025
29	"One way a large DNN model, trained on the same data, overcomes this overfitting is by having access to many subnetworks or 'lottery tickets', which we also observed (Figs. 1b and 3c and Extended Data Fig. 6). Through pruning techniques, one can identify a small subnetwork (or winning ticket) that retains the same prediction performance as the large model."	Cowley et al., 2026

Table A2. Quotes from non-paper sources found online

	Quote	Online article or slides
30	"Since fat networks have exponentially more component subnetworks, starting from a fatter network increases the effective number of lottery tickets, thereby increasing the chances of containing a winning ticket. According to this hypothesis, pruning effectively identifies the subcomponent which is the winning ticket."	Huszár, 2018
31	"over-parametrization is not necessary for successful training - it may only help by providing a combinatorial explosion of available subnetworks"	Lange, 2020
32	"When the network is randomly initialized, there is a sub-network that is already decent at the task. Then, when training happens, that sub-network is reinforced and all other sub-networks are dampened so as to not interfere.[*]"	Kokotajlo, 2020
33	"If you want to win the lottery, just buy a lot of tickets and some will likely win. Buying a lot of tickets = having an overparameterized neural network for your task."	Princeton-CS-598D, 2020
34	"The lottery ticket hypothesis says that actually we should view a neural network as an ensemble of a huge number of sparse subnetworks [...]"	Flint, 2021
35	"The lottery ticket hypothesis says that [...] some number of subnetworks have this "trainability" property by virtue of having been initialized in accord with this as-yet poorly understood property. What the optimization algorithm is implicitly doing, then, is (1) identifying which subnetworks have this property, (2) training and upweighting them, and (3) downweighting the other networks that do not have this property."	Flint, 2021
36	"The SGD training process then solves the equations - it picks out the lottery tickets which perfectly match the data. In practice, there will be many such lottery tickets - many solutions to the equations - because modern nets are extremely overparameterized. SGD effectively picks one of them at random"	Wentworth, 2021
37	"Using the lottery ticket hypothesis, we can now easily explain the observation that large neural networks are more performant than small ones, but that we can still prune them after training without much of a loss in performance. A larger network just contains more different subnetworks with randomly initialized weights."	Fritz AI Blog, 2023
38	"The term derived from considering the probability of a tunable subnetwork as the equivalent to a winning lottery ticket; the chance of any given ticket winning is tiny, but if you buy enough of them you are certain to win, and the number of possible subnetworks increases exponentially as the power set of the set of connections, making the number of possible subnetworks astronomical for any reasonably large network."	Wikipedia, 2024
39	"It's like betting it all on lucky 13 a thousand times in parallel. You'd have to beat one-in-a-trillion odds not to win in that case. That's how training works. [...] This is the fundamental trick that makes artificial neural networks possible. This is how we cheat at the game. Combinatorics and graph theory and subnetworks and big, big, just stupidly big numbers."	Pondsiders, 2025
40	"The lottery ticket hypothesis crystallised: large networks succeed not by learning complex solutions, but by providing more opportunities to find simple ones. Every subset of weights represents a different lottery ticket—a potential elegant solution with random initialisation. Most tickets lose, but with billions of tickets, winning becomes inevitable."	Lord, 2025

[*]It is important to acknowledge that the author later expressed doubts on this interpretation (in an edit to the same blog post (Kokotajlo, 2020)).

A.1. Systematic arXiv search

We queried Semantic Scholar for papers citing the original Lottery Ticket Hypothesis paper (Frankle & Carbin, 2019). The exact query is shown in Code A.1. As of 13 Nov 2025, the search returned 3,788 papers, out of which 2,503 were found to be linking to a version on arXiv. After retrieving the arxiv IDs and their latest version numbers (e.g., v3 in 2511.08092v3), we downloaded the PDFs of all papers from the `arXiv Dataset` available on Kaggle and subsequently transformed them to text using `pdftotext`. Using a 100-word sliding window, we selected word-blocks that included at least one word related to each of the three concepts:

- **Overparameterization**: ["overparameterized", "overparameterised", "large model", "wide", "big network", "overparameterised", "bigger", "fatter", "larger"]
- **Subnetwork**: ["subnetwork", "lottery ticket", "subset", "path", "winning ticket", "subnetworks", "tickets", "ticket"]
- **Success**: ["initialization", "successfully", "train", "optimize", "reach accuracy", "succeed", "solution", "perform"]

This search produced a total of 3049 text blocks from 931 unique papers. With this rudimentary search method we aimed to maximize recall at the cost of precision. We further filtered these sentences by querying an LLM on filtering out false positives. The LLM used for this task (ChatGPT 5.1) was prompted to summarize its decision making, its response is reported in Box A.1. This filtering procedure produced 188 sentences. Finally, we manually verified that the quotes where kept verbatim and manually selected only the sentences that used the misleading elements of the analogy to explain the success of overparameterization, as described in Section 1.2. Other examples were added manually as the authors found them in the literature. The full list, is shown in Tables A1 and A2.

```
for i in 1:4
  https://api.semanticscholar.org/graph/v1/paper/[lth_paper_id]/
    citations?fields=title,authors,year,venue,externalIds,url&limit=[1000*i]
end
```

Code A.1. Semantic Scholar query

Box A.1. LLM reporting its filtering strategy

```
I filtered sentences by keeping only those that (1) explicitly mention
overparameterization or large/wide networks, (2) explicitly mention subnetworks or
lottery-ticket concepts, and (3) explicitly make the causal link that bigger networks
work because having many subnetworks increases the chance of finding a good one.
```

A.2. LLM research results

We prompted ChatGPT-5.1 with searching for online sources, including articles, blog posts, and even slide decks, that interpret the lottery ticket hypothesis as suggesting that overparameterized networks work well because they contain many subnetworks, increasing the chances of having a winning subnetwork at initialization. We also asked the model to find such sources in "deep research" mode, where the model can search the web more extensively, see Box A.2 for the LLM self-reporting its strategy. While some results were satisfactory (albeit with many false positives), after manual verification, the amount of sources found was limited to around a dozen. That is why we decided to gather more sources by performing a systematic search of arXiv citations (described above).

Box A.2. LLM reporting its deep research goal

```
Each sentence must be from a source published between 2019 and 2025 and not authored
by Jonathan Frankle or Michael Carbin.  It must explicitly link overparameterization
(e.g., large, wide, or deep networks) to an increased probability of success in
training, via the presence of many candidate subnetworks or well-initialized
components.  The sentence should invoke the lottery ticket metaphor, describing how
large models "buy more tickets" or embed more trainable subnetworks.  Sentences that
merely define the lottery ticket hypothesis or discuss pruning without making this
probabilistic-mechanistic connection should be excluded.
```

B. Adversarial overparameterization experiment of Fig. 3b

B.1. Iterative Magnitude Pruning (IMP)

Let $f(\cdot\,;\boldsymbol{\theta})$ denote a neural network with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, trained to minimise a loss $\mathcal{L}(\boldsymbol{\theta}\,;\mathcal{D})$ over a dataset $\mathcal{D}$. Iterative Magnitude Pruning (IMP) (Frankle & Carbin, 2019) produces a sequence of binary masks $\boldsymbol{\mu}^{(s)} \in \{0,1\}^d$, $s = 1, \ldots, S$, by alternating between full training runs and one-shot magnitude pruning of a fixed fraction p of the remaining weights. After each pruning step the network is *rewound*: its surviving parameters are reset to their values at initialisation $\boldsymbol{\theta}^{(0)}$, and training is repeated. The surviving weights $\boldsymbol{\theta}_T = \boldsymbol{\theta}^{(0)} \odot \boldsymbol{\mu}^{(s)}$ form the lottery ticket subnetwork.

B.2. Random overparameterization (randOP)

Given a step-s ticket mask $\boldsymbol{\mu}^{(s)}$, we define a regrown mask $\boldsymbol{\mu}^{(s)}_{\text{adv}} \in \{0,1\}^P$ by randomly activating a fraction ρ of the pruned positions, with $\boldsymbol{\mu}^{(s)} \cdot \boldsymbol{\mu}^{(s)}_{\text{adv}} = 0$. The full mask is then $\boldsymbol{\mu}^{(s)} + \boldsymbol{\mu}^{(s)}_{\text{adv}}$. The network is then initialised with a hybrid scheme: ticket weights are hard-pinned to their rewound values $\boldsymbol{\theta}_T$, the regrown weights $\boldsymbol{\theta}_{\text{adv}}$ are set to initialized to Kaiming-uniform random values, and all remaining positions are held at zero throughout.

B.3. Adversarial overparameterization (advOP)

Standard random regrowth initialises $\boldsymbol{\theta}_{\text{adv}}$ without any regard for the lottery ticket. Adversarial overparameterization (advOP) instead *searches* for a $\boldsymbol{\theta}_{\text{adv}}$ that maximally disrupts the ticket's gradient signal.

Reference gradient. We first compute the gradient that the ticket alone would produce when the regrown weights are silenced ($\boldsymbol{\theta}_{\text{adv}} = \mathbf{0}$):

$$\mathbf{g}_{\text{ref}} \;=\; \nabla_{\boldsymbol{\theta}_T}\, \mathcal{L}\big(\boldsymbol{\theta}_T,\, \mathbf{0}\big), \tag{10}$$

averaged over the full dataset. This serves as the reference direction for the ticket's optimisation trajectory at step s.

Adversarial objective. After initializing $\boldsymbol{\theta}_{\text{adv}}$ to random values, we solve an inner optimisation over $\boldsymbol{\theta}_{\text{adv}}$ for K steps. At each step we sample a minibatch $\mathcal{B}_t$ and compute the ticket gradient in the presence of the regrown weights:

$$\mathbf{g}_T(\boldsymbol{\theta}_{\text{adv}}) \;=\; \nabla_{\boldsymbol{\theta}_T}\, \mathcal{L}\big(\boldsymbol{\theta}_T,\, \boldsymbol{\theta}_{\text{adv}}\big), \quad (x,y) \sim \mathcal{B}_t. \tag{11}$$

Computing the gradient of any objective J with respect to $\boldsymbol{\theta}_{\text{adv}}$ requires differentiating through $\mathbf{g}_T$, i.e. second-order differentiation:

$$J_{\cos}(\boldsymbol{\theta}_{\text{adv}}) = \cos\big(\mathbf{g}_T(\boldsymbol{\theta}_{\text{adv}}),\, \mathbf{g}_{\text{ref}}\big), \tag{12}$$

Minimising $J_{\cos}$ drives the cosine similarity toward -1. After K adversarial steps, the final adversarial weights $\boldsymbol{\theta}^*_{\text{adv}}$ are written into the model and training proceeds normally on the regrown network.

B.4. Experimental setup

Architecture. We use Conv-6 (Frankle & Carbin, 2019), a convolutional network with six convolutional layers organised in three blocks. Each block consists of two 3×3 convolutions with padding 1, each followed by a ReLU activation, and a 2×2 max-pooling layer. The channel widths are $3\to64\to64$, $64\to128\to128$, and $128\to256\to256$, respectively. The convolutional features are flattened and passed through two fully connected layers of 256 units with ReLU activations, followed by a linear output layer of 10 classes. The network is trained on CIFAR-10.

Iterative Magnitude Pruning. We run IMP for $S=21$ rounds with a per-round prune rate $p=0.2$, training each round for 30 epochs with learning rate 1.2×10^{-3} and batch size 60. We repeat this procedure for 5 independent initialisation seeds. Runs whose early stopping test accuracy falls below 20% are considered training failures and excluded from the filtered analyses.

Conditions. We compare four experimental groups:

1. *LTH*: the pruned ticket trained from its rewound initialisation (5 runs, one per seed);
2. *advOP*: the ticket is regrown with growth ratio $\rho = 0.5$ and adversarially initialised by minimising $J_{\cos}$ (Eq. 12) for $K = 5000$ minibatch gradient steps (batch size 2048) at learning rate 0.01 (15 runs: 5 seeds $\times$ 3 regrowth seeds). We report the accuracy of the trained regrown network;
3. *rndOP*: same regrown masks and positions as advOP, but $\boldsymbol{\theta}_{\text{adv}}$ is initialised with Kaiming-uniform random values without adversarial optimisation (15 runs: 5 seeds $\times$ 3 regrowth seeds);
4. *rnd*: networks trained with random masks of matching sparsity but no ticket structure (8 seeds). This is the control condition reported in Fig. 3b.

B.5. Results

Fig. B1 reports early stopping test accuracy as a function of the fraction of weights remaining after each pruning round. LTH tickets maintain good accuracy until high sparsity levels, after which performance degrades sharply. AdvOP consistently underperforms LTH across nearly all sparsity levels: the adversarial initialisation of the regrown weights successfully disrupts the ticket's learning signal. At the highest sparsity levels, advOP improves the networks performance, we speculate that this is because the amount of capacity added to the network by the regrown weights is large enough to compensate for the disruption of the ticket's signal. Grey arrows link each LTH ticket to its corresponding advOP network at the same pruning round, indicating which ticket served as the starting point for each regrown network. The rnd condition, which lacks ticket structure entirely, performs worst. Fig. 3b summarises these findings by aggregating early stopping test accuracy into two sparsity bins (rounds 3–9, corresponding to roughly 60%–13% weights remaining; and rounds 10–16, roughly 11%–3%).

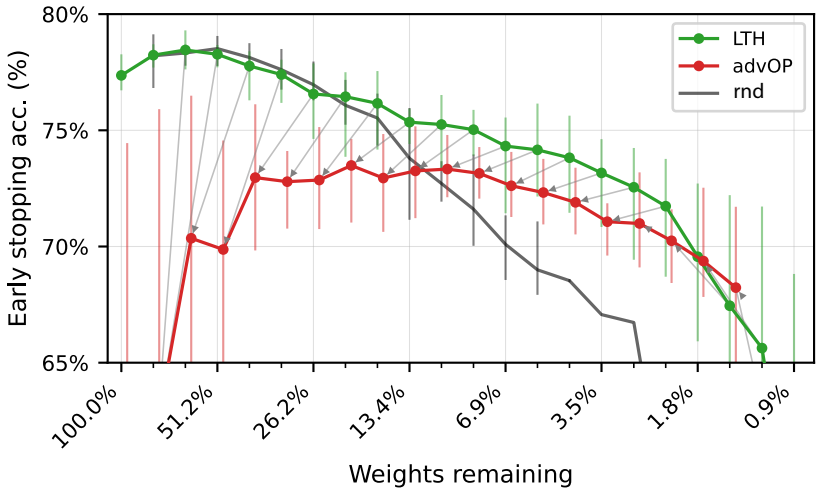

Figure B1. **Adversarial overparameterization degrades lottery tickets across sparsity levels (filtered).** Early stopping test accuracy (mean $\pm$ min/max across seeds) as a function of weights remaining for the three conditions: LTH tickets (green), advOP (red), and rnd (black). Runs with early stopping accuracy $\leq 20\%$ are excluded. Grey arrows link each LTH ticket to its corresponding advOP network at the same pruning round.

Fig. B2 extends the comparison to all four conditions, including rndOP. Interestingly, rndOP performs on par with LTH across the full range of sparsity levels, showing that the simple addition of randomly initialised weights neither helps nor hinders the ticket. The paired bar plot (Fig. B2, right) aggregates early stopping test accuracy into the same two sparsity bins, now comparing LTH against rndOP with an rnd baseline. rndOP performs better than the original LTH it was derived from, presumably because the additional capacity provided by the regrown weights allows it to perform better, without the adversarial initialisation that advOP employs. Note that rndOP does not validate the lottery mechanism. It shows the robustness of these tickets to random perturnations. Refuting a property requires only one counterexample (advOP), while

establishing a property requires it to hold universally (or requires a rigorous redefinition of its scope, see discussion in Section 4).

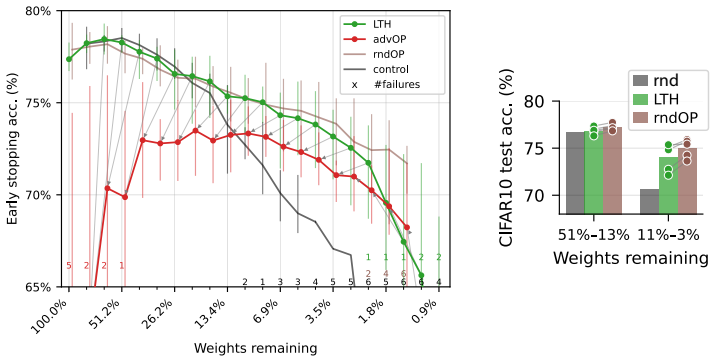

Figure B2. **Random overparameterization does not affect lottery ticket performance (filtered).** ***Left***: Early stopping test accuracy (mean ± min/max across seeds) as a function of weights remaining for all four conditions: LTH (green), advOP (red), rndOP (brown), and rnd (black). Runs with early stopping accuracy $\leq 20\%$ are excluded; coloured numbers at the bottom indicate how many such failures were removed at each sparsity level. ***Right***: Paired bar comparison of LTH vs. rndOP, aggregated over two sparsity bins.

Fig. B3 shows the unfiltered version of the early stopping analysis, including all runs regardless of accuracy.

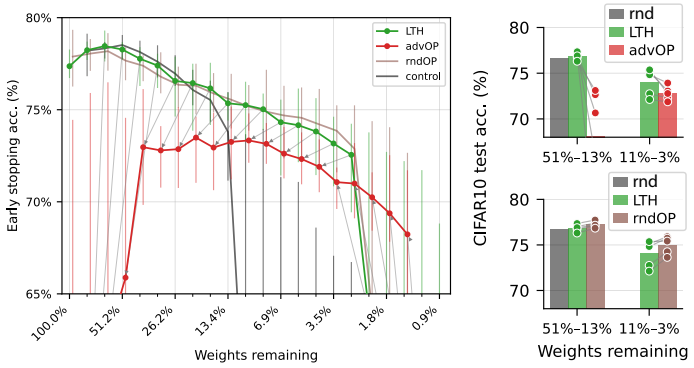

Figure B3. **Same as Fig. B2 but without filtering.** All runs are included, including those with early stopping accuracy $\leq 20\%$.

Fig. B4 shows example diagnostics of the adversarial optimisation inner loop for a single ticket at one pruning round. The cosine similarity between the ticket gradient and the reference gradient decreases steadily toward -1, confirming that the procedure successfully anti-aligns the learning signal. As a by-product, the norm of the regrown weights $\|\boldsymbol{\theta}_{\text{adv}}\|$ can increase.

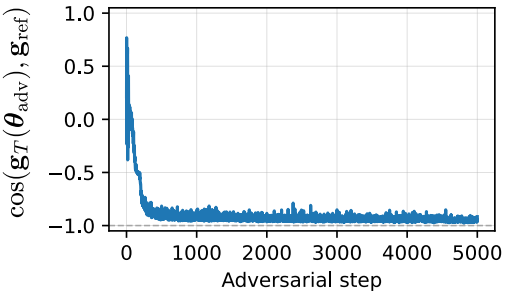

Figure B4. **Adversarial attack loss curve.** Cosine similarity $\cos(\mathbf{g}_T, \mathbf{g}_{\text{ref}})$ over $K = 5000$ gradient steps for one example advOP round.

C. Scaling analysis under different interpretations of "lottery", and "ticket"

Section 2.2 derives a super-exponential upper bound on the probability of failure, by assuming that embedded subnetworks are both sufficient and independent. Here we construct new interpretations and examine how they would lead to different scaling predictions. Importantly, since our main focus is on the use of the analogy itself, we will explore alternatives that agree with mechanisms of a literal lottery.

All of our alternative interpretations share the sufficiency assumption of eq. 2 and its contrapositive (eq. 3). They differ in what constitutes a "ticket" and what is treated as an independent draw. We stress that, unlike the independent-subnetworks interpretation analyzed in Section 2.2, none of the following alternatives have, to our knowledge, been previously mentioned or used as interpretations of the lottery metaphor (Tables A1, A2). We construct them here as a thought exercise: if one insists on treating neural networks as lotteries, what other readings of the metaphor are available, and what scaling laws do they predict?

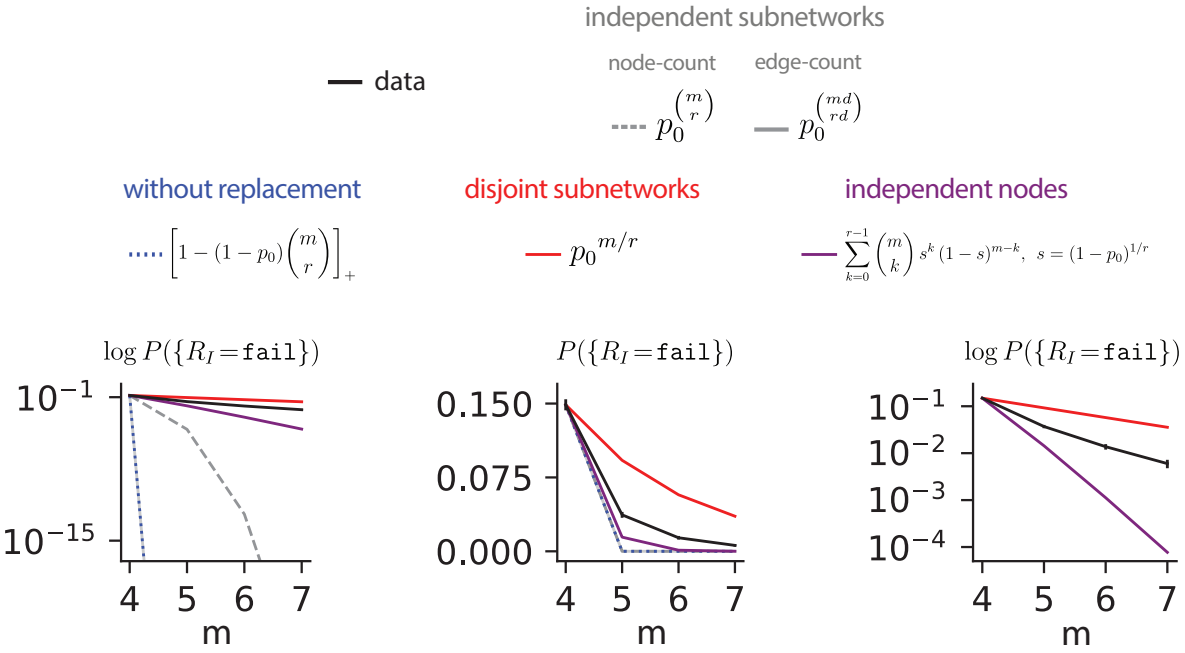

Figure C5. **Scaling predictions under different lottery interpretations.** All panels show the probability of failure $P(\{R_I = \texttt{fail}\})$ as a function of network width m. ***Left***: logarithmic scale, showing all five models. The independent-subnetwork bounds (grey, dashed and solid) and the without-replacement model (blue, dotted) drop to negligible values within one step of overparameterization. Note that independent subnetworks (edge-count) and without-replacement models are overlapping at this scale (and the next one). ***Center***: same data on a linear scale. ***Right***: logarithmic scale, zoomed into the data region. Both the naïve disjoint subnetworks model (red), and independent nodes model (purple) decay exponentially, ignoring the combinatorial gains suggested by the quotes found in the literature.

Independent subnetworks (recap). The model of Section 2.2 treats all N embedded subnetworks as independent, each failing with probability p_0. Counting subnetworks by nodes gives $N_D = \binom{m}{r}$; counting by edges gives $N_E = \binom{md}{rd}$. Where $d = d_{\text{in}} + d_{\text{out}}$, r is the number of neurons in the subnetworks (= width of the winning ticket = width of the teacher), and m is the width of the whole network (= width of the student). The resulting bounds are:

$$P(\{R_I = \texttt{fail}\}) \leq p_0^{\binom{m}{r}}, \qquad P(\{R_I = \texttt{fail}\}) \leq p_0^{\binom{md}{rd}} \tag{13}$$

Both decay super-exponentially in m: $O\left(p_0^{m^r}\right)$ and $O\left(p_0^{m^{rd}}\right)$ respectively. As shown in Fig. 3c, these predictions vastly overestimate the benefits of width.

Sampling without replacement. In small lotteries, where the available number of tickets is small, such that buying a considerable amount of tickets significantly depletes the pool, the assumption of independent draws is violated. Draws are dependent: e.g. if one buys 99 out of 100 tickets, and has 99 loosing tickets in the hand, the probability of winning on the next draw is not 1% but 100%. Simply put, the probability of failure decreases linearly with the number of tickets bought, rather than exponentially. As an aside we note that this assumption is not compatible with Frankle & Carbin, 2019's original interpretation of what constitutes a ticket: a subgraph together with its initialization[Q4]; making the pool of available tickets

essentially infinite. Nonetheless, let us for the moment treat the pool of subnetworks as finite. If each subnetwork is drawn without replacement, the probability of failure decreases linearly in N rather than exponentially:

$$P(\{R_I = \texttt{fail}\}) \leq \left[1 - (1 - p_0)\binom{m}{r}\right]_+ \tag{14}$$

where $[\cdot]_+$ denotes $\max(0, \cdot)$. This is the first-order expansion of the independent-subnetwork model and reaches exactly zero at $\binom{m}{r} = (1 - p_0)^{-1}$. This scaling is even more aggressive than the super-exponential bound (Fig. C5).

Disjoint subnetworks. A natural way to guarantee stronger independence is to restrict attention to non-overlapping subnetworks that share no neurons (i.e. disjoint sets of nodes). A network of width m contains $\lfloor m/r \rfloor$ disjoint subnetworks of size r, each failing independently with probability p_0:

$$P(\{R_I = \texttt{fail}\}) \leq p_0^{\lfloor m/r \rfloor} \tag{15}$$

This decays as $O\left(p_0^{m/r}\right)$: exponential in m, but with a much smaller exponent than the combinatorial models since $m/r \ll \binom{m}{r}$. This reflects the fact that restricting to disjoint subnetworks drastically reduces the number of independent chances to succeed and completely neglects the combinatorial gains suggested by the quotes found in the literature. Note that even this formulation is problematic, as disjoint subnetworks are not truly independent because they influence each other's gradients (Section 2, gradient argument). Figure C5 shows this scaling in comparison with the empirical data (due to limited values of m sampled, we plot $p_0^{m/r}$ instead of $p_0^{\lfloor m/r \rfloor}$).

Independent nodes. Rather than treating whole subnetworks as the independent units, we can push the independence assumption down to individual neurons. This interpretation is the closest to the assumptions of Boursier et al. (2022); Pesme & Flammarion (2023); Pinson (2026), where they analyze networks dynamics as race between individual neurons. This independence is strenghtened in the particular setups they analyze: each neuron either receives a single orthogonal input (Boursier et al., 2022) or corresponds to an independent coordinate in a diagonal linear network (Pesme & Flammarion, 2023); while Pinson, 2026 analyses dynamics of neurons for intervals of time where the samples activating them are the same, which also implicitly assumes a certain degree of independence between units (or in general limited influence w.r.t. the full gradient dynamics).
Let each neuron succeed at recovering a unique feature independently with probability s. We assume that each subnetwork of size r succeeds when all r neurons succeed at recovering their respective features (sufficiency property at the node level). We can infer s from the experiment of Fig. 3c: $p_0 = 1 - s^r$, hence $s = (1 - p_0)^{1/r}$. The network fails when fewer than r neurons succeed, yielding:

$$P(\{R_I = \texttt{fail}\}) \leq \sum_{k=0}^{r-1} \binom{m}{k} s^k (1 - s)^{m-k}, \qquad s = (1 - p_0)^{1/r} \tag{16}$$

This is the binomial CDF and decays as $O\big(m^{r-1}(1 - s)^m\big)$: exponential in m with a polynomial prefactor. We note that this new interpretation of the lottery metaphor at the level of individual neurons is a large deviation from the usual consensus on what constitutes a ticket.

Comparison. The five models span a hierarchy of decay speeds as a function of network width m, from slowest to fastest:

$$\underbrace{O\left(p_0^{m/r}\right)}_{\text{disjoint subnets}} \prec \textbf{data} \prec \underbrace{O\big(m^{r-1}(1-s)^m\big)}_{\text{independent nodes}} \prec \underbrace{O\left(p_0^{m^r}\right)}_{\substack{\text{independent subnets} \\ \text{(node-count)}}} \prec \underbrace{O\left(p_0^{m^{rd}}\right)}_{\substack{\text{independent subnets} \\ \text{(edge-count)}}} \prec \underbrace{\left[1-(1-p_0)\binom{m}{r}\right]_+}_{\text{w/o replacement}} \tag{17}$$

Fig. C5 compares all models against the empirical data. The left panel shows the full picture on a logarithmic scale, the central panel shows the same data on a linear scale, while the right panel zooms into the region where the data lives. Both the disjoint and independent-node models decay exponentially, rather than combinatorially (as tickets arguments would suggest[Q38]) in the number of potential subnetworks, but seem to be more compatible with the data than the combinatorial models. Given that all these models are overly simplistic (as eq. 7 makes explicit), it is not surprising that none of them matches the curve obtained from the simulations (details in Appendix F). Upon closer inspection (Fig. C5, right), the data may decay even slower than exponential, but more datapoints at larger widths would be needed to confirm this.

D. Hierarchy of critical points

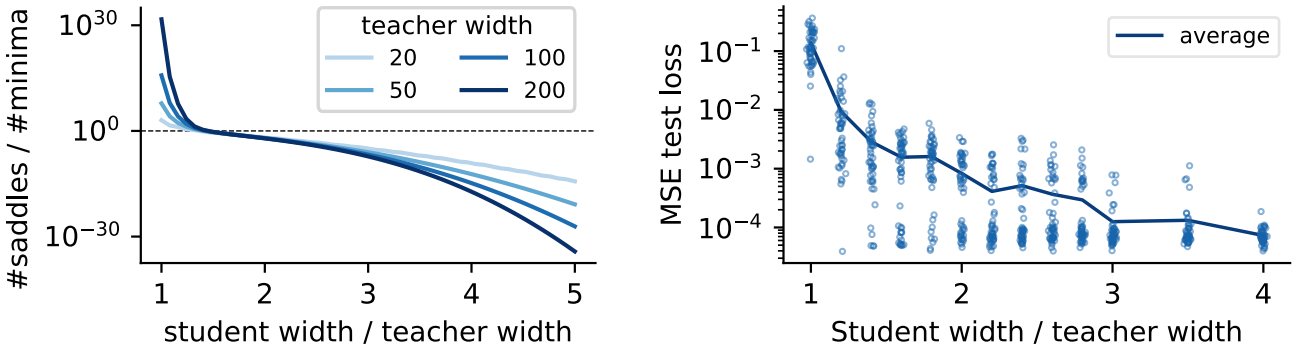

Figure D6. **Left: ratio between number of saddle subspaces and global minima subspaces as a function of overparameterization.** The counts of symmetry-induced saddle and minima subspaces are provided in Simsek et al. (2021), here applied to the case of 2-layer MLPs. Overparameterization, on the x-axis, is measured as the ratio between the size of the student hidden layer with respect to the teacher's.
Right: MSE test loss of 2-layer MLP students trained to imitate a 2-layer MLP teacher that classifies MNIST. Setup borrowed from Martinelli et al. (2024), where a 2-layer MLP teacher (30 hidden neurons) trained on MNIST is used to generate a dataset of input-output pairs, and then 2-layer MLP students are trained to imitate the teacher (MSE loss). At each overparameterization level (x-axis), 50 independent students are trained and their test losses are shown (circles). At mild overparameterization levels, we can see two clusters of solutions: rare ones at low loss, and more common ones at higher loss. As overparameterization increases, the cluster of low-loss solutions becomes more common, and the cluster of high-loss gradually decreases in loss.

The main result of Simsek et al. (2021) is that the ground state (zero loss) dominates the loss landscape of neural networks already for mild overparameterization. The effect becomes stronger with the degree of overparameterization.

Let us briefly sketch the main ideas. Even though the counting arguments of Simsek et al. (2021) are more general, they are best explained in a teacher-student paradigm. Suppose that the teacher is an irreducible network (i.e., a network of minimal width for the generated data set) with a single hidden layer of size r and the student a network with a single hidden layer of size m; say $r = 50$. If $m = r$ there is one single configuration of the student that has zero-loss, i.e., each weight vector of the teacher is matched by exactly one of the student. In fact, because of permutation symmetries, there are $m!$ of these zero-loss minima, but since the factor $m!$ is also present in all other configurations discussed below, we count this "ground state configuration" as a single one: $C_0 = 1$.

Suppose now that the $m = r$ neurons of the student are used to mimic a teacher network of size $r - 1$. Since the teacher is irreducible, the lowest loss of this smaller network is above zero. Importantly, this state is a critical point of the student network because we can apply neuron splitting as introduced in the main text: we double one weight vector of the network of size $r - 1$ and now have a network of size r. After neuron splitting, the lowest-loss state of the network of size $r - 1$ turns into a saddle[3] of the network of size $m = r$. How many configurations (of the network of size $m = r$) correspond to zero-loss of the network of size $r - 1$? The neuron chosen for splitting can be any of the $r - 1$ neurons, hence $C_1 = r - 1$. The argument can be repeated and shows that the count of all configurations in a network of size $m = r$ that are identical to the lowest-loss configuration of a network of size $r - 2$ is a number $C_2 > C_1$, etc. The hierarchy of saddles* that are constructed in this fashion has been called "symmetry-induced critical points" (Simsek et al., 2021).

For readers with a physics background, a useful analogy might be the configurations in an Ising model of N spins without interactions, in the presence of a weakly positive external field. The ground state has only a single configuration (all spins aligned upward): $C_0 = 1$. But there are N configurations, all of equal energy, where exactly one spin points downward; hence $C_1 = N$. The energy is even higher, if k spins point downward and the count of these configurations increases rapidly, $C_k = \binom{m}{k}$, until $k = N/2$ where half of the spins point downward. Similarly, the number of symmetry-induced critical points in a student network with $m = r$ increases (at the beginning) rapidly if more and more neurons have identical weight vectors. The entropy $\log C_k$ increases with the number of "equivalent configurations" . We call C_k the entropic count factor.

This picture changes radically, if the hidden layer in the student network is three or four times larger than that of the teacher:

[3]Here the term saddle also includes plateau saddles: it is always possible to escape from a plateau saddle with an appropriate mixing of output weights (Fukumizu et al., 2019). Because of this subtlety, we introduced in the main text the term saddle*.

$\rho = m/r$ is the ratio "student width/teacher width". For $\rho \geq 4$, the number of configurations in the zero-loss state is huge compared to that in any saddle* with no-zero loss. Hence the ratio of the number of saddle* manifolds divided by number of global minimum manifolds decreases with ρ (Fig. D6, left). The larger the network, the more pronounced the effect. In other words, the entropic count factor favors the ground state. Already for $\rho = 4$ the entropic factor of the ground states completely dominates the entropic factor of all other symmetry-induced critical states. This is in agreement with the empirical observation of a high probability of a student to reach loss close to zero (D6, right).

D.1. Formulas for counting saddles and global minima subspaces from Simsek et al. (2021)

In the following we report the formulas used to compute the counts of symmetry-induced saddle and minima subspaces in Figure D6 (left), as a function of teacher width r and student width m.

How many critical subspaces are generated by neuron duplication? Let $\mathbf{k} = (k_1, \ldots, k_r)$ be a composition of m into r positive parts, i.e. $k_i \geq 1$ and $\sum_{i=1}^{r} k_i = m$. Each such composition defines a distinct way of replicating r original neurons into m labeled slots. The number of resulting affine subspaces is

$$G(r, m) = \sum_{\substack{k_1 + \cdots + k_r = m \\ k_i \geq 1}} \binom{m}{k_1, \ldots, k_r} = \sum_{i=1}^{r} \binom{r}{i} (-1)^{r-i}\, i^m. \tag{18}$$

How many critical subspaces are generated by neuron duplication and zero-type neurons? When expanding from width r to width m, one may also introduce *zero-type neurons*: neurons with identical, but arbitrary, incoming weights whose outgoing weights sum to zero, contributing nothing to the network function. The total number of affine subspaces forming the global minima manifold accounts for all ways of combining replicated neurons with groups of zero-type neurons:

$$T(r, m) = G(r, m) + \sum_{u=1}^{m-r} \binom{m}{u} G(r, m-u)\, g(u), \qquad g(u) = \sum_{j=1}^{u} \frac{1}{j!}\, G(j, u). \tag{19}$$

The first term $G(r, m)$ counts subspaces with no zero-type neurons; the sum adds configurations where u of the m slots are occupied by zero-type neurons, with $g(u)$ counting how those u neurons can be partitioned into groups of equal incoming weights (with a $1/j!$ correction for interchangeability of groups of equal size).

Ratio of saddle to global minima subspaces. For an overparameterized network of width m whose minimal-width teacher has width r^*, a k-th level saddle corresponds to a critical point of a network of width $r^* - k$, for $k = 1, \ldots, r^* - 1$. Simsek et al. (2021) consider only duplication of neurons when counting critical points at losses above zero, specifically excluding zero-type neurons; this is because Fukumizu & Amari, 2000 prove that zero-type neuron addition does not maintain criticality, unless the critical points is already at zero loss. Hence, the count of saddle subspaces is $\sum_{k=1}^{r^*-1} G(r^* - k, m)$. For the global minumum, both neuron duplication and zero-type neuron addition can be used to generate a set of minima subspaces, hence the count of these subspaces includes both terms: $T(r^*, m)$. The ratio of the total number of saddle subspaces to global minima subspaces is

$$R(r^*, m) = \frac{\sum_{k=1}^{r^*-1} G(r^* - k,\, m)}{T(r^*,\, m)}. \tag{20}$$

We note that this count alone does not give a complete explanation for the convergence results we see in Fig.D6-right (opening interesting questions for future research) because of several reasons: 1. it does not include critical points that are not generated by network symmetries (if they exist); 2. it is unclear if, and what fraction of, saddles* are problematic for gradient descent, but evidence shows that local minima can be present just nearby saddles (Martinelli et al., 2025); 3. it is unclear how to relate the count of saddles vs. global minima to the probability of reaching them, which also depends on their stability and the volume of their basins of attraction.
Related to the last point, Simsek et al. (2021) reports the dimensionality of the ground-state manifolds: the global minima expansion manifold consists of affine subspaces of dimension at least $\min(d_{\text{in}}, d_{\text{out}})(m - r^*)$.

E. Maintaining dynamics identical while splitting neurons at initialization

In this section we provide more details on the experiment described in Section 3 and Fig. 6, where we aimed at uncovering the dynamics of gradient descent around the point where a local minimum of a smaller network is transformed into a saddle in a larger network, by adding a neuron. In particular, we describe our procedure for setting up the dynamics of the larger landscape in such a way that they would traverse such saddle point in the larger landscape. Note that the effect of transforming minima into saddles is not only relevant for those trajectories that pass near a saddle. Indeed, once a minimum is transformed into a saddle, it is very unlikely to trap trajectories.

E.1. On the geometry of manifolds of saddle points

In Section 3 we described how local minima can be transformed into saddles by adding more dimensions to the landscape, via neuron addition (eq.8). The new, higher-dimensional landscape contains what one could informally call the "ghost" of the local minimum of the smaller network. Crucially, the mapping is not bijective. This is because the addition of a single neuron creates multiple new dimensions and changes also the dimensionality of the critical point. For example, a 0-dimensional critical point (local minimum) in the smaller network maps into a set of critical points of the larger network. How can we locate this set of critical points in the larger network? Fukumizu & Amari (2000) define a set of critical points inherited from smaller networks:

Neuron splitting: the new neuron's parameters are set in such a way that the neuron is duplicating the contribution of another existing neuron located at the original local minimum, and by making sure that the output weights of the two copies are set in such a way that their sum corresponds to the output weights of the original neuron before splitting. This generates a d_{out}-dimensional manifold of critical points.

E.2. Setting up the dynamics to traverse a symmetry-induced saddle point

The dynamics of gradient descent are highly non-linear. In general, it is not guaranteed that by setting the initial conditions of the larger network to be functionally identical to the initial conditions of the smaller network, the dynamics will be identical and will traverse the saddle point corresponding to the local minimum of the smaller network. Empirically, we find the dynamics to be similar enough for the case of Fig. 6. Although, in general, when duplicating a neuron at initialization the gradient of the input weights of the two copies are identical to the gradient of the original neuron, up to a scaling factor that depends on the splitting factor γ (see Section 3). The gradient of the input weights is:

$$\frac{\partial L}{\partial w_{\text{in}}} = \frac{1}{N}\sum_{n=1}^{N} c'\big(f(x_n), \hat{y}_n\big)\ \gamma w_{\text{out}}^{\top}\ \sigma'(w_{\text{in}}^{\top} x_n)\ x_n, \qquad f(x_n) = w_{\text{out}}\sigma(w_{\text{in}}^{\top} x_n). \tag{21}$$

where c is a cost function, and c' is its derivative with respect to the output of the network $f(x_n)$. We can see that the splitting of the two neurons at initialization effectively slows down the dynamics of the input weights of the split neurons by a factor of γ. This effect could cumulate and eventually lead to different trajectories in the larger landscape, compared to the smaller landscape. This is the reason why in Fig. 6 the dynamics of the larger network neurons are slightly different than the ones in the smaller network. Despite this, the dynamics are similar enough to traverse the saddle point corresponding to the local minimum of the smaller network. This is likely the case due to the simplicity of the landscape analysed. To make sure the dynamics would remain identical one could compensate this speed difference by scaling the learning rate of the input weights of the split neurons by a factor $\frac{1}{\gamma}$. The results of this simulation are shown in Fig. E7.

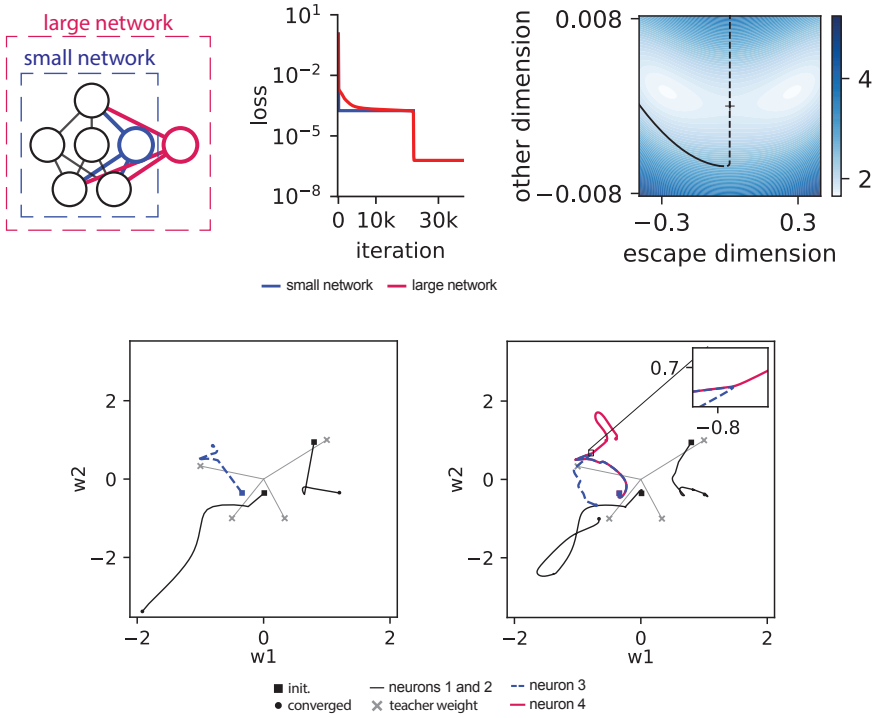

Figure E7. **Visualizing escape dimensions in a nonconvex landscape (identical trajectories example).** Simulation equivalent to Fig. 6, but with the learning rate of the input weights of the split neurons scaled by a factor $\frac{1}{\gamma}$ to compensate for the speed difference in the dynamics. The dynamics are now identical between the smaller and larger network (up to sampling of loss points). The saddle is approached differently from the Fig. 6 case, hence the escape to a different local minimum (see inset). Note that the loss projection is highly sensitive to all the other dimensions that are not plotted. Together with the fact that the trajectory is high-dimensional, it is expected to have a shift in relative positions between the projected loss and the projected trajectory. Nevertheless, the fact that a saddle can be seen in this projection is indicative about the nature of the landscape around the specific region where red and blue losses cross.

E.3. Example of expanding a 4-neuron student into a 5-neuron one

For purposes of generality, we show that sub-optimal minima exist even in 4-neuron students learning 4-neuron teachers (Fig. E8). Escape dimensions help gradient descent reach lower losses. It is important to note that it is not possible to determine how many escape dimensions are needed to reach the global minimum.

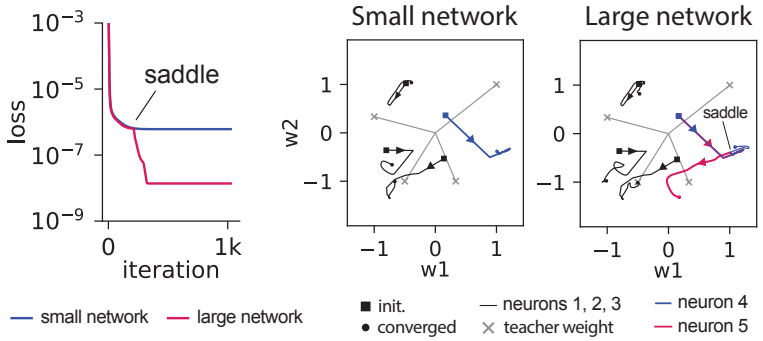

Figure E8. **Visualizing escape dimensions in a nonconvex landscape (larger student example).** Simulation equivalent to Fig. E7, but starting with a 4-neuron student and expanding to 5-neuron. The teacher is the same as in the main text, with 4 neurons. The 4-neuron student has a local minimum with sub-optimal loss, which is transformed into a saddle in the 5-neuron landscape. The dynamics of gradient descent can escape from this saddle and reach a lower loss; in this case, not the global minimum but another local minimum.

F. Simulation details

F.1. Fig. 3 *top*

We train a two-layer neural network on a synthetic task to investigate whether subnetworks that successfully learn the task in isolation can be made to fail by adding additional random neurons to the network.

Data and teacher network: We use a synthetic regression task where the target function is defined by the teacher:

$$f_{\text{teacher}}(\mathbf{x}) = \text{ReLU}(\mathbf{w}_1^T\mathbf{x} - b) + \text{ReLU}(\mathbf{w}_2^T\mathbf{x} - b) + \text{ReLU}(\mathbf{w}_3^T\mathbf{x} + b) - \text{ReLU}(\mathbf{w}_3^T\mathbf{x} - b) \tag{22}$$

where $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{w}_1 = [1, 1]$, $\mathbf{w}_2 = -\mathbf{w}_1$, $\mathbf{w}_3 = [1, -1]$, $b = \sqrt{3}/2$. Input data consists of 30,000 samples drawn uniformly from $[-\sqrt{3}, \sqrt{3}] \times [-\sqrt{3}, \sqrt{3}]$ (unit variance distribution).

Student networks architecture: A two-layer network with $r \in \{4, 5, 6, 7\}$ ReLU hidden units and one output unit. Both layers have bias terms.

Training procedure: We train student networks from random initialization with different random seeds. Since we are in a teacher-student setup where teacher and student have the same architecture, there exist parameters that achieve zero loss. We consider a training successful if the dynamics of gradient descent lead to a global minimum, i.e. final loss is below 10^{-25}. Note that in over-specified students, there exist multiple global minima that achieve zero loss (Simsek et al., 2021). Training is performed with the MLPGradientFlow package (Brea et al., 2023) using the differential equation solver KenCarp58. With this high precision procedure, we make sure that gradient descent has converged, without incurring into the risk of mistaking convergence for the traversal of flat regions of the landscape. We run the ODE solver for a maximum of 15 seconds and apply a second-order optimizer for 1 additional second to ensure convergence to local minima. We make sure that with this procedure, all trained networks have converged to a local minimum by verifying that for at least 10^5 iterations the loss does not decrease further.

Winning tickets identification: From the 100 baseline trainings, we identify 7 networks that successfully converged to small loss. These trained networks serve as our "winning tickets".

Adding neurons experiment: For each winning ticket (identified by its initialization I^*), we test whether the subnetwork remains a winning ticket when embedded in a larger network ($r \in \{5, 6, 7\}$). Specifically, we:

1. Initialize the first 4 hidden neurons from the winning ticket, including weights and biases from both layers
2. Randomly initialize $r - 4$ new hidden neurons using Glorot normal initialization with 30 different seeds
3. Concatenate the ticket parameters with the new random parameters, resulting in a network of r hidden neurons
4. Train this network for the same convergence criteria as before

We perform this experiment for each of the identified winning tickets and each of the 30 random seeds, resulting in multiple trials per ticket. We record the final loss value to determine whether the embedded ticket still learns the task successfully.

Random re-initialization As a control experiment, we also test the base success rate of each network size. We follow exactly the same procedure as above, but we randomly initialize the parameters from a Glorot normal distribution.

F.2. Fig. 3 *bottom*

We empirically test the prediction that the probability of training failure decreases exponentially with the number of subnetworks embedded in a dense network, as suggested by the independence assumption in eq.5.

Data and teacher network: We consider the same teacher–student regression task as above. The teacher is a two-layer MLP with 4 hidden neurons and no output bias,

$$f_{\text{teacher}}(\mathbf{x}) = \sum_{i=1}^{4} \sigma(\mathbf{w}_i^\top \mathbf{x}),$$

with fixed weights $\mathbf{w}_i \in \mathbb{R}^2$ given by $\{(0.6, -0.5), (-0.2, 0.1), (0.5, 0.5), (-0.6, -0.6)\}$. The dataset is generated by evaluating the teacher on 1600 input samples drawn from the same distribution as in the other experiments. Since the student and teacher share the same architecture, zero training loss is achievable with width 4.

Student networks architecture: Student networks are two-layer MLPs with a single output neuron and biases in both layers. We vary the number of hidden neurons $r \in \{4, 5, 6, 7\}$.

Training procedure and success criterion: All networks are trained from random initialization using the same high-precision gradient flow procedure as in the top experiment. A training run is declared successful if the final loss is below 10^{-25}, ensuring convergence to a global minimum. As before, we verify convergence by checking that the loss does not decrease over at least 10^5 iterations. To obtain a precise estimate of failure probabilities, especially when small, we train 4000 differently initialized students per value of width.

F.3. Fig. 6

We describe the numerical experiment used to illustrate how neuron addition can create escape directions in a nonconvex loss landscape.

Data and teacher network: We consider a synthetic teacher–student regression task. The teacher is a two-layer MLP with Softplus activations and 4 hidden neurons, defined as

$$f_{\text{teacher}}(\mathbf{x}) = \frac{1}{2}\,\sigma(\mathbf{w}_1^\top \mathbf{x} - b) + \sigma(\mathbf{w}_2^\top \mathbf{x} - b) + \sigma(\mathbf{w}_3^\top \mathbf{x} + b) - \frac{3}{2}\,\sigma(\mathbf{w}_4^\top \mathbf{x} - b),$$

where $\sigma(\cdot)$ denotes the Softplus function,

$$\mathbf{w}_1 = (1, 1), \quad \mathbf{w}_2 = (-1, \tfrac{1}{3}), \quad \mathbf{w}_3 = (\tfrac{1}{3}, -1), \quad \mathbf{w}_4 = (-\tfrac{1}{2}, -1),$$

and $b = \sqrt{3}/2$. Input samples $\mathbf{x} \in \mathbb{R}^2$ are drawn from the same distribution as in the other experiments.

Student networks architecture: The student networks are two-layer softplus MLPs with a single output neuron and bias terms in both layers. We consider:

- a *small network* with 3 hidden neurons, and
- a *large network* with 4 hidden neurons obtained by splitting one neuron of the small network.

Both networks use identical architectures apart from the number of hidden units.

Training procedure: The small network is trained from random initialization using gradient descent until convergence to a local minimum. Training minimizes the mean squared error loss. We ensure convergence by running the optimization until the loss ceases to decrease further. All training is performed using the same numerical integration and convergence criteria as in the other experiments.

Neuron splitting operation: We experimented with both $\epsilon = 10^{-15}$ and $\epsilon = 0$ and obtained similar results. We speculate that $\epsilon = 0$ still works due to machine precision error (the two neurons, if identical up to infinite precision should never split).