# Iterative Magnitude Pruning as a Renormalisation Group: A Study in The Context of The Lottery Ticket Hypothesis

**Abu-Al Hassan**
**CID : 01495449**

Supervised by: Akshunna S. Dogra, Jeroen S.W. Lamb

Imperial College London

June 25, 2023

# Abstract

This thesis delves into the intricate world of Deep Neural Networks (DNNs), focusing on the exciting concept of the Lottery Ticket Hypothesis (LTH). The LTH posits that within extensive DNNs, smaller, trainable subnetworks — termed "winning tickets" — can achieve performance comparable to the full model. A key process in LTH, Iterative Magnitude Pruning (IMP), incrementally eliminates minimal weights, emulating stepwise learning in DNNs. Once we identify these winning tickets, we further investigate their "universality" - that is, we check if a winning ticket that works well for one specific problem could also work well for other, similar problems. We also bridge the divide between the IMP and the Renormalisation Group (RG) theory in physics, promoting a more rigorous understanding of IMP.



# Acknowledgements

I would like to express my deepest gratitude to Akshunna S. Dogra for their continuous support and supervision throughout my project, as well as for sharing their valuable insights on transferability and the Lottery Ticket Hypothesis. I am also grateful to Jeroen S.W. Lamb for providing guidance on Renormalisation Group theory and offering helpful advice on structuring the thesis. Additionally, I would like to extend my appreciation to William T. Redman for sharing his expertise on Iterative Magnitude Pruning.


# Declaration

I, Abu-Al Hassan, declare that this thesis, titled "Iterative Magnitude Pruning as a Renormalisation Group: A Study in The Context of The Lottery Ticket Hypothesis", is my own work, except where cited.

# Contents

# 1 Introduction

Deep Neural Networks (DNNs), despite their impressive capabilities, often entail a considerable computational overhead due to the sheer magnitude of parameters — typically between $10^6$ to $10^{11}$ [15] [20]. This vastness tends to decelerate the training process. One potent strategy to combat this computational bottleneck is pruning — eliminating superfluous connections or neurons, which reduces the number of computationally expensive parameters, thereby accelerating prediction times.

Our thesis is devoted to an in-depth exploration of the Lottery Ticket Hypothesis (LTH), a groundbreaking idea in the realm of deep learning [11] [12] [37]. According to LTH, there exist "winning tickets" — smaller subnetworks embedded within DNNs — that can be trained to match or even outperform the full models. This thesis seeks to uncover these winning tickets and test their universality, i.e., a winning ticket that is successful for one task (the specific problem a neural network is designed to solve) may also be efficacious for other tasks within the same class (similar problem types) [29] [27] [37]. This principle could potentially expedite training times and expedite the development of high-performing models across a range of tasks within the same class.

An integral component of our investigation is the technique of Iterative Magnitude Pruning (IMP) [10] [24], which facilitates the discovery of winning tickets by gradually eliminating the least significant weights. In a more theoretical vein, our thesis draws connections between IMP and the Renormalisation Group (RG) theory, a powerful mathematical framework in physics [14]. RG theory provides insights into how transformations in a physical system unfold at various scales. Similarly, in IMP, we view different parameter densities as variations in scale. We show that IMP is an RG scheme [29]. By applying methodologies from the Renormalisation Group theory to IMP, we aspire to foster a more rigorous and generalisable understanding of IMP, which is currently bereft of an effective theory [28, p. 1]. This endeavour could potentially amplify the efficacy and reliability of IMP and, in turn, substantially impact the field of deep learning.

# 2 Introduction to Iterative Magnitude Pruning

Iterative Magnitude Pruning (IMP) initiates by fully training a neural network and subsequently discarding a fraction of the least significant weights. Following this pruning step, the residual weights revert to their initial configuration. This cycle of training, pruning, and resetting transpires iteratively, progressively unearthing an efficient subnetwork, or "winning ticket". This methodology, grounded in the Lottery Ticket Hypothesis, provides a systematic means to discover these winning tickets within larger, often unwieldy neural networks.

While IMP is favoured in practice due to its conceptual simplicity, ease of implementation, and efficacy, theoretical explanations of its effectiveness have been limited [2]. Balwani and Krzyston's study [2] shows that IMP preferentially retains weights that maintain network

topology, providing unique insights into the extent of pruning possible without affecting zeroth order topological features.

---

**Algorithm:** Iterative Magnitude Pruning (1)

---

**Input:** Loss function $L : \mathbb{R}^p \to \mathbb{R}$, training time $T \in \mathbb{R}_+$, initialisation $\boldsymbol{w}^{\text{init}} \in \mathbb{R}^p$, iterations of pruning $q < p$.

**Output:** $\boldsymbol{w}^{(q)}(T)$

---

**Method for IMP at x% level:** (adaptation of [10, p. 2])

Set $M^0 = \mathbb{I}_p$

for $k = 0$ to $q$ do:

Initialise $\boldsymbol{w}^{(k)}(0) = M^k \boldsymbol{w}^{\text{init}}$

Train $\dot{\boldsymbol{w}}^{(k)}(t) = -M^k \nabla L\left(\boldsymbol{w}^{(k)}(t)\right)$ for $t \in [0, T]$

Prune the smallest $x\%$ of $\left\{\left|w_j^{(k)}(T)\right| : M_{jj}^k = 1\right\}$ and set corresponding $M_{ii}^k = 0$

Set $M^{k+1} = M^k$

return $\boldsymbol{w}^{(q)}(T)$

---

Where $\boldsymbol{w}^{\text{init}} \in \mathbb{R}^p$ is the initial set of weights for the neural network.

$q$ represents the number of pruning steps. It must be less than the number of parameters in the network, $p$.

$M$ is called the mask as the diagonal entries of this matrix will indicate which weights are active (1) and which are pruned (0).

Train $\dot{\boldsymbol{w}}^{(k)}(t) = -M \nabla L\left(\boldsymbol{w}^{(k)}(t)\right)$ for $t \in [0, T]$: This line trains the network by following the negative gradient of the loss function, with the modification that pruned weights don't get updated.

While technically, biases in a neural network are also considered weights, it's important to understand their unique role in model learning and performance. Biases are employed to shift the activation function to the left or right, which can prove crucial in successful learning and adaptation to data. Pruning biases might, therefore, impose a greater negative impact on a network's performance than pruning weights.

Moreover, the count of bias terms in a neural network is usually significantly less than that of

weights. Therefore, retaining biases does not substantially contribute to model complexity. This consideration, coupled with the potential for enhanced accuracy, justifies the exclusion of biases from the pruning process. This approach is consistent with the methodology proposed in the original Lottery Ticket Hypothesis paper by Frankle and Carbin [11].

For a theoretical understanding of neural network pruning and the effects of pruning on the properties and capabilities of the neural network, you can read [35] [9].

# 3 Lottery Ticket Hypothesis

The Lottery Ticket Hypothesis (LTH), an intriguing concept in the field of deep learning, was proposed by Frankle and Carbin in 2019 [11]. This hypothesis posits that randomly initialized, dense neural networks embed subnetworks – known as "winning tickets". These subnetworks, when trained in isolation, are capable of achieving comparable accuracy to the original, dense network, but with less computational time. Winning tickets are identified using the previously discussed method of Iterative Magnitude Pruning (see section 2). The LTH derives its name from the analogy of unearthing these high-performing subnetworks within the complex web of a dense network as akin to finding a winning ticket in a lottery. By facilitating the discovery of these "winning tickets", the LTH offers a strategy to drastically reduce the computational resources required for training DNNs, without diminishing their performance – thereby boosting efficiency.

The size of a winning ticket, or the pruned network, hinges upon several factors: the specific task at hand, the model's architectural design, the optimization algorithm, and the pruning strategy employed – whether Magnitude-based Pruning, Sensitivity-based Pruning, or Random Pruning etc [6] [21] [11]. However, there is no strict size constraint that defines a subnetwork as a winning ticket. Empirical evidence from experiments conducted by Frankle and Carbin in 2019 suggests that winning tickets containing less than 10-20% of the parameters of the original network can operate without sacrificing accuracy [11, p. 1].

## 3.1 Identifying winning tickets from Iterative Magnitude Pruning

At each iteration of IMP, a subnetwork is derived. Assume we execute $q$ iterations of IMP with initial weights denoted as $\boldsymbol{w}^{\text{init}}$ , following the procedure outlined in (1). We conclude with a subnetwork characterized by weights $\boldsymbol{w}^{(q)}(T)$ and its corresponding mask matrix $M^q$. Subsequently, we initialize the same neural network with weights $\boldsymbol{w} = M^q \boldsymbol{w}^{\text{init}}$ . If the model, when trained from $\boldsymbol{w}$, attains an accuracy comparable to the full model or better (model trained from $\boldsymbol{w}^{\text{init}}$ ), we classify $\boldsymbol{w} = M^q \boldsymbol{w}^{\text{init}}$ as a winning ticket for the model.

## 3.2 Universality of winning tickets

The Lottery Ticket Hypothesis posits that winning tickets $\boldsymbol{w}$ are task-specific, suggesting that a winning ticket identified for one task might not exhibit high performance on a differ-

ent task [11]. The inherent structure and connections of a winning ticket are optimized for a specific task and may not generalize well to others.

Nonetheless, recent studies have unearthed evidence that winning tickets can exhibit transferability to related tasks [29] [27] [31]. For instance, if two tasks share close relations, a winning ticket $\boldsymbol{w}$ discovered for one task might still exhibit commendable performance on the other task, albeit not optimal.

Additionally, research [5] demonstrates that winning tickets can be transferred between disparate architectures. This signifies the capability to utilize a winning ticket identified in one neural network architecture as an initialization point for training in a different architecture.

These observations imply the feasibility of employing winning tickets to investigate the similarities between "tasks" and "architectures".

# 4 Renormalisation Group theory

The central concept in Renormalisation Group theory is the concept of "scaling." Many systems exhibit behaviour that is "scale invariant," meaning that their properties remain the same under a change in scale [33] [19]. In such systems, physical quantities often follow a "power law" behaviour, with quantities scaling as some power of the scale factor.

The Renormalisation Group theory is particularly useful for understanding the behaviour of systems near "critical points," where the system undergoes a phase transition. For instance, near the critical temperature, many systems exhibit power-law scaling behaviour. In such cases, the power-law exponent, known as the "critical exponent," can provide significant insights into the behaviour of the system.

The theory also provides a way to classify systems into "universality classes." Systems within the same universality class have the same critical exponents, indicating that they behave in the same way near their critical points, regardless of the specifics of their microscopic interactions. This has significant implications for the study of complex systems, as it allows for the prediction of macroscopic behaviour from a limited understanding of microscopic interactions.

## 4.1 Block Spins argument by L.P. Kadanoff

The 'Block Spins' argument introduced by Leo P. Kadanoff [18] is one of the seminal concepts that formed the basis of modern Renormalization Group (RG) theory. By introducing the concept of block spins and coarse-graining, Kadanoff established a connection between microscopic details and macroscopic behaviour. His approach led to the development of the renormalization group theory. Kandanoff introduces the concept of scaling. The idea of scaling arises from the observation that, near the critical point, systems exhibit similar

behaviour on different length scales. Kadanoff's approach to coarse-graining and block spins helps to demonstrate that the free energy of the original system and the block spin system have the same functional form, with the block spin system effectively scaling the original system.

This scaling behaviour leads to the development of scaling laws and critical exponents, which are crucial for understanding the properties of systems near critical points.

Consider a d-dimensional grid with spacing $a$, where the system has the Hamiltonian function (energy function)[14, pp. 230–235]:

$$\begin{aligned} \beta H_\Omega &= -\beta J \sum_{\langle ij\rangle=1}^{N} S_i S_j - \beta H \sum_i S_i \\ &\equiv -K \sum_{\langle ij\rangle} S_i S_j - h \sum_{i=1}^{N} S_i \end{aligned} \tag{2}$$

with

$$\begin{aligned} K &\equiv \beta J \\ h &= \beta H. \end{aligned} \tag{3}$$

On the grid, we can replace a block of side $\ell a$ that contains $\ell^d$ with a single 'block spin'. Then the total number of block spins is $N\ell^{-d}$ for some $N$.

We can define the spin of block I to be $S_I$:

$$S_I \equiv \frac{1}{|\overline{m}_\ell|} \frac{1}{\ell^d} \sum_{i \in I} S_i \tag{4}$$

Where $\bar{m}_\ell$ is the average magnetization of block we defined as:

$$\bar{m}_\ell \equiv \frac{1}{\ell^d} \sum_{i \in I} \langle S_i \rangle \tag{5}$$

After this normalisation the magnitudes of the spins are the same as the original system (non-coarsed):

$$\langle S_I \rangle = \pm 1 \tag{6}$$

Block spin renormalisation theory has two assumptions:
**The first assumption** is that since in the original system, spins interact with only nearest-neighbour spins and the external field, we can assume that the new blocks also interact with the nearest neighbour block spins and an effective external field.

As a result of this assumption, we need to define new coupling constants between the block spins and the effective external field. We can write these as $K_\ell$ and $h_\ell$, where for the original

system $\ell = 1$. With this the Hamiltonian becomes:

$$-\beta H_\ell = K_\ell \sum_{\langle IJ \rangle}^{N\ell^{-d}} S_I S_J + h_\ell \sum_{I=1}^{N\ell^{-d}} S_I, \tag{7}$$

Notice this has the same form as the original Hamiltonian except. This new system has fewer spins than the original system.

The system with Hamiltonian $H_\ell$ is further away from criticality than the original system $H_\Omega$. It also has a new effective reduced temperature, $t_\ell$ and an effective magnetization field $h_\ell$. The reduced temperature, t, measures how far the temperature of the system is from the critical temperature ($T_c$).

$$t \equiv \frac{T - T_c}{T_c} \tag{8}$$

The relationship between effective reduced temperature and effective magnetization field is:

$$h_\ell = h\bar{m}_\ell l^d. \tag{9}$$

Since $H_\ell$ has the same form as $H_\Omega$, which means that the free energy of the block spin system will also have a similar form as the original spin system but with $t_\ell$ and $h_\ell$ instead of $t$ and $h$.
The free energy per spin (or block spin) of the original system is related to that of the block spin system by:

$$N\ell^{-d} f_s(t_\ell, h_\ell) = N f_s(t, h) \tag{10}$$

Leading to the functional form of the free energy per spin changes under the block spin transformation:

$$f_s(t_\ell, h_\ell) = \ell^d * f_s(t, h) \tag{11}$$

**The second assumption** is about how the reduced temperature ($t_\ell$) and external field ($h_\ell$) change during the transformation. We assume that:

$$\begin{aligned} t_\ell &= t\ell^{y_t} \quad y_t > 0 \\ h_\ell &= h\ell^{y_h} \quad y_h > 0. \end{aligned} \tag{12}$$

They are both dependent on the block size $\ell$ and we don't know $y_t$ and $y_h$ yet but assume they are positive.

Now we can write the relationship between the free energy per spin of the original system and the block spin system:

$$f_s(t, h) = \ell^{-d} f_s\left(t\ell^{y_t}, h\ell^{y_h}\right) \tag{13}$$

Kadanoff's block spin argument helps us understand the form of scaling relations, but it

doesn't give us the exponents $y_t$ and $y_h$. While Kadanoff's block spin idea provided a heuristic way to understand how systems change under coarse-graining, it didn't offer a precise, quantitative method for predicting these changes.

## 4.2 Critical Phenomena and Renormalisation Group

Having introduced 'Block spin' argument by Kadanoff, we build on this by introducing Wilson's RG approach [34]. Wilson extended the idea of block spins to field theories, providing a mathematical framework that is applicable to a broad range of systems beyond Ising-type spin models. He answered the question of how and why systems' properties change under transformations of scale and provides a way to understand the "fixed points" of this flow, which correspond to scale-invariant phases of the system. Kadanoff's original block spin argument didn't include these key concepts.

We will now detail coarse-graining transformation by examining the characteristics of the block spin transformations [14, pp. 236–239]. The key idea is that after performing a block spin transformation, the distance between block spins is a. If we rescale the lengths so that the new distance between block spins is the same as the original distance between microscopic spins, the system appears similar to the original system but with a different Hamiltonian. Repeating these steps produces a series of Hamiltonians, each describing systems that are further from criticality.

Let's consider a Hamiltonian described as:

$$\mathcal{H} \equiv -\beta H_\Omega = \sum_n K_n \Theta_n\{S\} \tag{14}$$

Here, $K_n$ are the coupling constants, and $\Theta_n S$ are the local operators, which are functionals of the degrees of freedom $S$.
I will now describe how the coupling constants change when we "zoom out" and look at the system at a larger scale.
The equation:

$$[K'] \equiv R_\ell[K] \quad \ell > 1 \tag{15}$$

tells us that when we apply the RGT, the original set of coupling constants $K$ transforms into a new set $K'$.

To calculate the RGT, we first define the partition function $Z_N[K]$, which is a mathematical tool used to analyze the statistical properties of a system, and a quantity $g[K]$ that is related to the free energy per degree of freedom:

$$Z_N[K] = \mathrm{Tr}\, e^{\mathcal{H}}$$

$$g[K] \equiv \frac{1}{N} \log Z_N[K]$$

The partition function is crucial because many thermodynamic properties, such as internal energy, entropy, and free energy, can be derived from it. It is a measure that encodes the statistical properties of a system in equilibrium.

The RGT reduces degrees of freedom by $\ell^d$, creating a new effective Hamiltonian for "block variables" $S'_I$ by taking a partial trace over original degrees of freedom $S_i$, while block degrees of freedom are fixed.

$$\begin{aligned} e^{\mathcal{H}'_N\{[K'],S'_I\}} &= \mathrm{Tr}_{\{S_i\}}\, e^{\mathcal{H}_N\{[K],S_i\}} \\ &= \mathrm{Tr}\,\{S_i\}\, P\left(S_i, S'_I\right) e^{\mathcal{H}_N\{[K],S_i\}} \end{aligned}$$

$P\left(S_i, S'_I\right)$ is a projection operator used to allow unrestricted trace in the equation. It ensures that the range of values for coarse-grained degrees of freedom $S'_I$ matches that of the original degrees of freedom $S_i$.

Let's work with the Ising spins on a square lattice. We define an RGT transformation using blocks of linear dimension $(2\ell + 1)a$, where a is the lattice spacing. The block spin $S'_I$ is given by:

$$S'_I = \mathrm{sign}\left(\sum_{i \in I} S_i\right) = \pm 1.$$

This means we assign the block spin value based on the sign of the sum of spins within the block. The associated projection operator is defined as:

$$P\left(S_i, S'_I\right) = \prod_I \delta\left(S'_I - \mathrm{sign}\left[\sum_{i \in I} S_i\right]\right)$$

The projection operator must satisfy three requirements:
(i) $P\left(S_i, S'_I\right) \geq 0$
(ii) $P\left(S_i, S'_1\right)$ reflects the symmetries of the system;
(iii) $\sum_{\{S'_1\}} P\left(S_i, S'_I\right) = 1$

Condition (i) ensures that the exponential term of the transformed Hamiltonian is non-negative, allowing us to identify the effective Hamiltonian for the new degrees of freedom, $S'_I$.

In condition (ii) by symmetry, it is meant that the proposed operator does not introduce any new or unauthorised couplings that were not possible in the original, non-coarse-grained system.

For example, if the original Hamiltonian has the form:

$$\mathcal{H}_N = NK_0 + h\sum_i S_i + K_1\sum_{ij} S_iS_j + K_2\sum_{ijk} S_iS_jS_k + \dots$$

The transformed Hamiltonian will have the same form, but with new, transformed coupling constants:

$$\mathcal{H}'_{N'} = N'K'_0 + h'\sum_I S'_I + K'_1\sum_{IJ} S'_IS'_J + K'_2\sum_{IJK} S'_IS'_JS'_K + \dots$$

Condition (iii) ensures a well-defined projection operator that has a clear, one-to-one mapping between the original and new parameters, and if probabilistic, the sum of all probabilities equals 1. Probabilistic Renormalisation Group operators are often used to study complex or disordered systems where deterministic Renormalisation Group operators are not well-suited.

As a result, the partition function is invariant under the Renormalisation Group transformation thus preserving the statistical properties and thermodynamic behaviour of the original system.

$$\begin{aligned} Z_{N'}\left[K'\right] &\equiv \mathrm{Tr}\left\{S'_I\right\} e^{\mathcal{H'}_N\left\{\left\{K'\right],S'_I\right\}} \\ &= \mathrm{Tr}\left\{S'_I\right\}\mathrm{Tr}\left\{S_I\right\} P\left(S_i,S'_I\right) e^{\mathcal{H}_N\{[K],S_i\}} \\ &= \mathrm{Tr}\left\{S_i\right\} e^{\mathcal{H}_N\{[K],S_i\}}\cdot 1 \\ &= Z_N[K] \end{aligned}$$

Which gives,

$$\begin{aligned} \frac{1}{N}\log Z_N[K] &= \frac{\ell^d}{\ell^d N}\log Z_{N'}\left[K'\right] \\ &= \ell^{-d}\frac{1}{N'}\log Z_{N'}\left[K'\right] \end{aligned}$$

leading to,

$$g[K] = \ell^{-d} g\left[K'\right]$$

Thus the free energy per degree of freedom is related to the transformed free energy by a factor of $\ell^{-d}$.

## 4.3 Fixed points[14, pp. 242–246]

In order to understand the behaviour of systems undergoing repeated applications of renormalization group (RG) transformations, we observe the evolution or "flow" of parameters. These parameters, originating from a wide range of initial values, constitute what is termed the renormalization group flow.

Interestingly, an aspect of this flow is its common tendency to gravitate towards specific points, known as fixed points. Fixed points in the space of coupling constants by definition

don't change under Renormalisation Group transformations. In the vicinity of these fixed points, systems tend to exhibit a characteristic scaling behaviour. This behaviour implies that these systems manifest consistent patterns or characteristics, irrespective of their scale or size, showcasing the universality of such patterns.

Mathematically a fixed point is represented by:

$$[K^*] = R_\ell \, [K^*]$$

Let's consider a Hamiltonian close to the fixed point Hamiltonian. We write it as:

$$\mathcal{H} = \mathcal{H} \, [K^*] \equiv \mathcal{H}^*$$

$$\mathcal{H} = \mathcal{H}^* + \delta\mathcal{H}$$

.
After performing an Renormalisation Group transformation: $[K'] = R_\ell[K]$, the new coupling constants, denoted by $K'_n$, can be written as:

$$K'_n = K'_n[K] \equiv K^*_n + \delta K'_n$$

By Taylors theorem,

$$K'_n \left\{K^*_1 + \delta K_1, K^*_2 + \delta K_2, \ldots\right\} = K^*_n + \sum_m \left.\frac{\partial K'_n}{\partial K_m}\right|_{K_m = K^*_m} \cdot \delta K_m + O\left((\delta K'\right)$$

Allowing us to express the change in coupling constants after the transformation, $\delta K'_n$, in terms of the original changes, $\delta K_n$.

$$\delta K'_m = \sum_m M_{nm} \delta K_m$$

Where $M_{nm}$ is the partial derivative of the new coupling constants with respect to the original ones, evaluated at the fixed point values. The matrix $M$ is the linearized Renormalisation Group transformation near the fixed point. For simplicity, we can assume that $M$ is symmetric.

To examine Renormalisation Group flows near the fixed point, we employ the linearised Renormalisation Group transformation denoted by $M^{(\ell)}$ and then study the eigenvalues and eigenvectors. "Flow" refers to the evolving system properties and interactions between regimes as the scale changes.

The eigenvalues and eigenvectors of $M^{(\ell)}$ are denoted as $\Lambda^{(\sigma)}_\ell$ and $e^{(\sigma)}_n$, where $\sigma$ identifies the eigenvalues and $n$ refers to the vector components. Employing the Einstein summation convention, we get:

$$M^{(\ell)}_{nm} e^{(\sigma)}_m = \Lambda^{(\sigma)} e^{(\sigma)}_n$$

Using the associativity of matrices,

$$\mathbf{M}^{(\ell)}\mathbf{M}^{(\ell')} = \mathbf{M}^{(\ell\ell')}$$

giving us,

$$\Lambda^{(\sigma)}_{\ell}\Lambda^{(\sigma)}_{\ell'} = \Lambda^{(\sigma)}_{\ell\ell'}$$

By differentiating with respect to $\ell'$, setting $\ell' = 1$, and solving the obtained differential equation, we derive:

$$\Lambda^{(\sigma)}_{(\ell)} = \ell^{y_\sigma}$$

Here, $y_\sigma$ is a number to be determined, but it is independent of $\ell$.

This shows that the eigenvalues can be expressed as a power of the scale factor, $\ell$. This information can help us understand how the system behaves near fixed points as the scale changes.

I will now explore how the changes in coupling constants, denoted by $\delta K$, transform under the linearized Renormalisation Group transformation $M$. We can express $\delta K$ in terms of the eigenvectors of $M$:

$$\delta\mathbf{K} = \sum_\sigma a^{(\sigma)}\mathbf{e}^{(\sigma)}$$

In this case, we express $[K]$ as a vector $\mathrm{K} = (K_1, K_2, \ldots)$. The orthonormality of eigenvectors is assumed to determine the coefficients $a^{(\sigma)}$:
When we apply the linearized Renormalisation Group transformation $\mathbf{M}$, we get:

$$\begin{aligned}\delta\mathbf{K}' &= \mathbf{M}\delta\mathbf{K} \\ &= \mathbf{M}\sum_\sigma a^{(\sigma)}\mathbf{e}^{(\sigma)} \\ &= \sum_\sigma a^{(\sigma)}\Lambda^{(\sigma)}\mathbf{e}^{(\sigma)} \equiv \sum_\sigma a^{(\sigma)\prime}\mathbf{e}^{(\sigma)},\end{aligned}$$

Here, we define $a^{(\sigma)\prime}$ as the projection of $\delta K'$ in the direction $\mathrm{e}^{(\sigma)}$. This equation is crucial because it tells us that some components of $\delta\mathrm{K}$ grow under $M^{(\ell)}$ while others shrink. If we arrange the eigenvalues by their absolute value,

The eigenvalue absolute values follow the order:

$$|\Lambda_1| \geq |\Lambda_2| \geq |\Lambda_3|$$

We consider three cases:

a) $\left|\Lambda^{(\sigma)}\right| > 1$ or $y^\sigma > 0$: $a^{(\sigma)\prime}$ grows with increasing $\ell$.

b) $\left|\Lambda^{(\sigma)}\right| < 1$ or $y^\sigma < 0$: $a^{(\sigma)}$ diminishes with increasing $\ell$.

c) $\left|\Lambda^{(\sigma)}\right| = 1$ or $y^\sigma = 0$: $a^{(\sigma)\prime}$ remains constant as $\ell$ increases.

After applying $\mathrm{M}^{(\ell)}$ multiple times, only components of $\delta K$ along directions $\mathrm{e}^{(\sigma)}$ for the first case (a) are significant. Projections of $\delta\mathbf{K}$ in other directions will either shrink or stay fixed. These cases are named:

a) Relevant eigenvalues/directions/eigenvectors.

b) Irrelevant eigenvalues/directions/eigenvectors.

c) Marginal eigenvalues/directions/eigenvectors.

## 4.4 Universality in Renormalisation Group theory

The concept of universality signifies that disparate systems, despite possessing distinct microscopic details, may exhibit analogous behaviour near critical points–points at which a system undergoes a phase transition. Systems that demonstrate such behaviour are classified into the same 'universality class.'

The 'universality class' of a system is dictated by the count and the nature of its pertinent parameters (or directions). Various microscopic models may possess unique parameter coupling constants corresponding to a given critical phenomenon. However, only a handful of these parameter combinations influence the systems' behaviour around critical points. The parameters associated with these influential directions are deemed as relevant. Systems that share an identical count and nature of relevant parameters at a fixed point will exhibit matching critical exponents, thereby categorizing them into the same universality class [29, p. 4].
Parameters classified as irrelevant do not influence the universality class, as their impact recedes over large scales or at lower energies, a concept encapsulated by the energy-length duality. Irrelevant directions exhibit a negative critical exponent, denoted as $y$. Consequently, an increase in the scale factor ($\ell$) diminishes the contribution of these operators, given that they are effectively multiplied by $\ell^y$, where $y < 0$ [14].

## 4.5 RG Flow [29, pp. 3–4]

The Renormalisation Group operator, $\mathcal{R}$, simplifies complex systems by replacing local variables with their composite values. The way to formally study $\mathcal{R}$ is by analyzing its effect on the energy function (Hamiltonian). For classical spin systems, the Hamiltonian has a specific form:

$$\mathcal{H}(\mathbf{s}, \mathbf{k}) = -\sum_i k_1 s_i - \sum_{\langle i,j \rangle} k_2 s_i s_j - \ldots, \tag{16}$$

Where $s_i$ are the spins, $\langle \cdot, \cdot \rangle$ represents nearest neighbor sites on the lattice, and $k_i$ are the strengths of the different coupling constants.

The Renormalisation Group operator combines spins, creating a new spin system with different coupling constants. The new system is represented by a new Hamiltonian,

$$\mathcal{R}\mathcal{H}(\mathbf{s}, \mathbf{k}) = \mathcal{H}\left(\mathbf{s}', \mathcal{T}\mathbf{k}\right) = \mathcal{H}\left(\mathbf{s}', \mathbf{k}'\right), \tag{17}$$

Which is derived from the original Hamiltonian, $\mathcal{H}(\mathbf{s}', \mathbf{k})$, using the transformed spins, $\mathbf{s}'$, and couplings, $\mathbf{k}'$, determined by the operator $\mathcal{T} : \mathbb{R}^K \to \mathbb{R}^K$.

The Renormalisation Group operator, $\mathcal{R}$, can be used to change the Hamiltonian of a spin system and its set of couplings. When $\mathcal{R}$ is applied repeatedly, it generates a flow in the function space of Hamiltonians and the space of coupling constants, which is referred to as Renormalisation Group flow. The Renormalisation Group flow changes depending on the eigenvectors of the linearised operator $\mathcal{T}$ near fixed points, with the eigenvectors defined as relevant ($\lambda_i > 1$) or irrelevant($\lambda_i < 1$) based on their eigenvalues.

### 4.6 Power-law scaling in Renormalisation Group theory[14, pp. 252–253]

Having gained an understanding of Renormalisation Group flows, let's examine how the Renormalisation Group explains scaling behaviour. The phenomenon of power-law scaling is commonly observed in the critical phenomena studied using Renormalisation Group such as phase transitions in Physics. Phase transitions occur when changing one control parameter within a range leads to the divergence of another parameter called the order parameter or its derivative.

For example, near the critical temperature of a ferromagnetic phase transition (where metal goes from being non-magnetic to magnetic as it is cooled), the magnetization display power-law scaling for t such that $t_L < t < t_C$. The scaling is characterised by the equation $m \sim (t_C - t)^{-\beta} = \Delta t^{-\beta}$, where m is magnetization, $\Delta t$ represents the difference between the critical temperature and the temperature of the system. Here $\beta$ is called the critical exponent.

Working with the Ising model, we have two important factors that can change, which are labelled 't' and 'h'. We start with a formula for the free energy density. This is a measure of the energy in the system that could potentially be used to do work.

$$f(t, h) = \ell^{-d} f\left(t', h'\right) \tag{18}$$

The new parameters $t'$ and $h'$ are generated by transformations:

$$\begin{aligned} T' &= R_\ell^T(T,H) \\ H' &= R_\ell^H(T,H) \end{aligned} \tag{19}$$

Where $R_\ell^T$ and $R_\ell^H$ are coarse-graining functions. We then consider the neighbourhood of a fixed point (T*, H*) where

$$\begin{aligned} T^* &= R_\ell^T(T^*,H^*) \\ H^* &= R_\ell^T(T^*,H^*). \end{aligned} \tag{20}$$

Next, we consider small deviations from the fixed point (T*, H*), called $\Delta T$ and $\Delta H$:

$$\begin{aligned} \Delta T &= T - T^* \\ \Delta H &= H - H^* \end{aligned} \tag{21}$$

$$\begin{pmatrix} \Delta T' \\ \Delta H' \end{pmatrix} = \mathbf{M} \begin{pmatrix} \Delta T \\ \Delta H \end{pmatrix} \tag{22}$$

with

$$\mathrm{M} = \begin{pmatrix} \partial R_\ell^T/\partial T & \partial R_\ell^T/\partial H \\ \partial R_\ell^H/\partial T & \partial R_\ell^H/\partial H \end{pmatrix}_{\substack{T=T^* \\ H=H^*}} \tag{23}$$

$M$ represents how small changes in T and H affect the transformations. The eigenvectors of M are special directions in which the transformations act simply by stretching or shrinking, and they are combinations of T and H. Often, M is diagonal and does not mix T and H, which simplifies things. For now, we assume this is the case.

We write the eigenvalues of M (these are numbers associated with each eigenvector that tell us how much stretching or shrinking occurs in that direction) as:

$$\begin{aligned} \Lambda_\ell^t &= \ell^{y_t}; \\ \Lambda_\ell^h &= \ell^{y_h}, \end{aligned} \tag{24}$$

the Renormalisation Group transformation becomes

$$\begin{pmatrix} t' \\ h' \end{pmatrix} = \begin{pmatrix} \Lambda_\ell^t & 0 \\ 0 & \Lambda_\ell^h \end{pmatrix} \begin{pmatrix} t \\ h \end{pmatrix}. \tag{25}$$

Before proceeding further with the derivation of power-law scaling. We will define the correlation length $\xi$ in Renormalisation Group theory. This gives a sense of how far apart two points can be while still influencing each other.

Mathematically, the correlation function typically decays exponentially for distances larger than the correlation length:

$$C(r) \approx exp(\frac{-r}{\xi}) \tag{26}$$

where C(r) is the correlation function and r is the distance between two points.

Suppose under Renormalisation Group transformations, we zoom out by a factor of $\ell$, the correlation length effectively gets smaller by the same factor ($\xi \to \frac{\xi}{\ell}$).

Hence if we apply the Renormalisation Group transformation n times, the correlation length transforms as:

$$\xi(t,h) = \ell^n \xi\left(\ell^{n y_t} t, \ell^{n y_h} h\right) \tag{27}$$

The singular part of the free energy density then transforms according to:

$$\begin{aligned} f(t,h) &= \ell^{-d} f\left(t', h'\right) = \ell^{-nd} f\left(t^{(n)}, h^{(n)}\right) \\ &= \ell^{-nd} f\left(\ell^{n y_t} t, \ell^{n y_h} h\right), \end{aligned} \tag{28}$$

This equation looks like the result we expected from the Kadanoff block spin argument, 13. Next, if we choose $\ell^n = b \cdot t^{-\frac{1}{y_t}}$, we obtain:

$$f(t,h) = t^{d/y_t} b^{-d} f\left(b, h/t^{y_h/y_t}\right) \tag{29}$$

The scaling behaviour of the singular part of the free energy density is typically written as:

$$f_s(t,h) = |t|^{2-\alpha} F_f\left(h/|t|^{\Delta}\right) \tag{30}$$

with

$$F_f(x) \equiv f_s(1,x) \tag{31}$$

This is a key result known as the static scaling hypothesis, which is a central assumption in the Renormalisation Group analysis of critical phenomena. Where $\alpha$ is the critical exponent associated with the heat capacity. More specifically, $2-\alpha$ is the scaling dimension of the free energy. $\Delta$ describes how the magnetic field $h$ scales with the reduced temperature $t$ in the critical region.

We then find:

$$\begin{aligned} 2 - \alpha = d\nu &= \frac{d}{y_t} \\ \Delta &= y_h / y_t. \end{aligned} \tag{32}$$

We now have a way to calculate the exponents $y_t$ and $y_h$, at least approximately, from the Renormalisation Group recursion relations.

# 5 Understanding the Dynamics of Iterative Magnitude Pruning [29, p. 4]

The IMP process can be represented by the equation

$$\mathcal{I}\mathcal{L}(\mathbf{a}, \boldsymbol{\theta}) = \mathcal{L}\left(\mathbf{a}', \mathcal{T}\boldsymbol{\theta}\right) = \mathcal{L}\left(\mathbf{a}', \boldsymbol{\theta}'\right) \tag{33}$$

Where the new set of parameters, $\boldsymbol{\theta}'$, are given by an operator $\mathcal{T}$ and result in a new set of activations, $\mathbf{a}'$.

In IMP, $\mathcal{T}$ is a combination of two operators: a masking operator $\mathcal{M}$ and a refining operator $\mathcal{F}$. This means that $\mathcal{T} = \mathcal{F} \circ \mathcal{M} : \mathbb{R}^N \to \mathbb{R}^N$, where $N$ is the number of parameters in the DNN. The pruning procedure used determines the definition of $\mathcal{M}$ (e.g. magnitude pruning), while the refinement procedure used defines $\mathcal{F}$.

For $n$ iterations of $\mathcal{I}$, the final DNN is given by

$$\mathcal{I}^n\mathcal{L}(\mathbf{a}^{(0)}, \boldsymbol{\theta}^{(0)}) = \mathcal{L}(\mathbf{a}^{(n-1)}, \mathcal{T}^n\boldsymbol{\theta}^{(0)}) = \mathcal{L}(\mathbf{a}^{(n-1)}, \boldsymbol{\theta}^{(n-1)}) \tag{34}$$

This creates a path in parameter space, $\boldsymbol{\theta}^{(0)} \to \boldsymbol{\theta}^{(1)} \to \ldots \to \boldsymbol{\theta}^{(n-1)}$, called the IMP flow, determined by the eigenvectors of $\mathcal{T}$ and their eigenvalues.

## 5.1 Power-law scaling in IMP

Rosenfeld et al. (2020) [30] recently discovered a similarity between universality in renormalization group (RG) and Lottery Ticket Hypothesis theories when they analysed the pruning of a DNN using IMP. They found that when the density (the percentage of remaining parameters) of the DNN falls within a certain range, $d_L < d < d_C$, the error of the DNN follows a power-law relationship, which can be expressed as $e \sim (d_C - d)^{-\gamma} = \Delta d^{-\gamma}$, where $\gamma$ is the critical exponent.

## 5.2 IMP flow

To understand how IMP flow works, we need to find the eigenfunctions of the aforementioned operator $\mathcal{T}$. By examining the eigenvalues associated with these eigenfunctions, we can determine which directions are relevant and irrelevant. Note this is analogous to the section on fixed point analysis in Renormalisation Group theory.

For spin systems, Renormalisation Group flow is studied in the space of coupling constants. Where the coupling constants are assumed the same for all spins, which greatly reduces the dimensionality of the parameter space. However NNs don't set the parameters of the same type to the same value, hence we need to study the IMP flow in directly by estimating the relative "influence" the parameters of a given layer have on the full NN. This can be done by considering the total remaining parameter magnitude that remains in layer $i$ after

$n$ applications of IMP [29, p. 6]:

$$M_i(n) = \frac{\sum_{j=1}^{N^{(i)}} \left| m_j^{(i)}(n) \cdot \theta_j^{(i)}(n) \right|}{\sum_{k=1}^{N} |m_k(n) \cdot \theta_k(n)|} \tag{35}$$

Here $N^(i)$ is the number of parameters in layer $i$ and $m^{(i)} \in \{0,1\}^{N^{(i)}}$ is the pruning mask. The dot product of the parameters with the pruning mask makes sure only the non-pruned weights are considered.

If we are to consider $M_i(n)$ as eigenfunctions of the IMP operator they should scale exponentially with respect to the number of IMP iterations. As $M_i(n+1) = \mathcal{T}M_i(n) = \lambda_i M_i(n) = \lambda_i^{n+1} M(0)$. We can drive $\lambda_i$ as:

$$\lambda_i = \frac{M_i(n+1)}{M_i(n)} \tag{36}$$

In Redman et al, they found that $M_i(n+1)$ is appropriate to be considered as an eigenfunction due to its well-defined nature and the standard error of the mean of $\lambda_i$ is less than 5%. The degree of coarse-graining ($x \in (0,1)$) at each iteration of IMP affects the magnitude of the eigenvalues. Therefore we are interested in the quantity $\sigma$:

$$\lambda_i \sim c^{\sigma_i} \tag{37}$$

Where $\sigma$ is invariant to the choice of c and taking $log_c(\lambda_i)$ gives $\sigma_i$. Here c is defined as $\frac{1}{1-x}$. We want to compare across models that prune using different c values, we will report $\sigma_i$. Using this we have:

1. Relevant directions, which have $\lambda_i > 1$, have $\sigma_i > 0$.
2. Irrelevant directions, which have $\lambda_i < 1$, have $\sigma_i < 0$.

The Lottery Ticket Hypothesis suggests that DNNs can be reduced to a low-dimensional subspace of important parameters during training, and this subspace can be used to transfer winning tickets between different DNN models. The Renormalisation Group theory provides tools for finding these subspaces and comparing them between models, by analyzing the eigenvectors and eigenvalues of the transformation matrix. The models that have the same eigenvectors with eigenvalues greater than 1 are said to have the same relevant parameters, making it possible to know if winning tickets can be transferred between them without additional experiments. However, having distinct relevant directions does not mean that the tickets cannot be transferred, but it suggests that the models have different properties and will be differently affected by the RG.

# 6 Connection between Renormalisation Group and IMP

The IMP and Renormalisation Group operator are techniques for simplifying complex physical systems. IMP "sparsifies" neural networks, while Renormalisation Group carries out

coarse-graining on spin systems. By reducing the number of degrees of freedom in the original system, the coarse-grained spin system becomes simpler and more manageable. Both techniques show power-law scaling and have unique flows that allow us to understand the relevant components of the system that determine certain macroscopic behaviour.

Table 1. Showing analogous quantities in Renormalisation Group and IMP theory [29, p. 2].

| **RG** | IMP |
|---|---|
| Spins ($s_i$) | Unit activations ($a_i$) |
| Coupling constants ($k_i$) | Parameters ($\theta_i$) |
| Hamiltonian ($\mathcal{H}[\mathbf{s}, \mathbf{k}]$) | Loss function ($\mathcal{L}[\mathbf{a}, \boldsymbol{\theta}]$) |

(38)

Prior to the Redman et al 2021 [29], there was no connection made between renormalisation group and IMP. However previous research had made a connection between renormalisation group theory and deep learning [26] [22] [3]. One important such paper was Mehta et al [26], which is based on the idea of hierarchical organization, where understanding at one level is built upon the understanding at a lower level. In deep learning, each layer of the network captures features at a different level of abstraction, with input data at the bottom and the final output at the top. This is analogous to the renormalization group, where the behaviour of a system at larger scales (higher levels) is derived from the behaviour at smaller scales (lower levels).
For a different flavour,

## 6.1 Proof: IMP is a Renormalisation Group scheme

In this section, we meticulously unpack and refine the proof sketch provided by Redman [29, p. 5], incorporating additional clarity and rectifying minor inaccuracies. Our primary aim is to demonstrate that IMP aligns with the conceptual underpinnings of a Renormalisation Group scheme. To this end, we scrutinize a single application of the IMP procedure as encapsulated by the $\mathcal{I}$ operator. Subsequently, we endeavour to validate that $\mathcal{I}$ fulfils all prerequisites necessary to qualify as a Renormalisation Group projection operator, thereby substantiating the claim that IMP can indeed be categorized as a Renormalisation Group operator.

We start by establishing a concrete correlation between IMP and the Renormalisation Group theory by showing that IMP satisfies the specifications of a Renormalisation Group scheme. For this purpose, we direct our attention to the projection operator, $\mathcal{P}$, tied with the Renormalisation Group operator. This projection operator, $\mathcal{P}$, facilitates the mapping of spins within a classical spin system, denoted as $s_i$, onto a coarser spin system, $s'_I$. The operation is characterized by the following equation:

$$\mathrm{Tr}_{\{s_i\}} \mathcal{P}\left(s_i, s'_I\right) \exp\left[\mathcal{H}\left(s_i, \mathbf{k}\right)\right] = \exp\left[\mathcal{H}\left(s'_I, \mathbf{k}'\right)\right], \tag{39}$$

Where $\mathrm{Tr}_{\{s_i\}}$ is the trace operator over the values that the $s_i$ can take (e.g. $\pm 1$ ).

The projection operator must have three properties:
1) $\mathcal{P}\left(s_i, s'_I\right) \geq 0$
2) $\mathcal{P}\left(s_i, s'_I\right)$ respects the symmetry of the system .
3) $\sum_{\{s'_I\}} \mathcal{P}\left(s_i, s'_I\right) = 1$.

Our primary objective is to identify the projection operator linked to $\mathcal{I}$. To initiate this, we map the activations of all units, denoted as 'a', before and after the application of IMP. This approach is based on the similarity between the activations 'a' and the quantity $\mathbf{s}$. Concentrating on unit $j$ in layer $i$, we represent its activation with the following equation:

$$a_j^{(i)} = h\left[\sum_k g_k(\mathbf{a}, \boldsymbol{\theta})\right], \tag{40}$$

This equation signifies that the activation of unit $j$ in layer $i$, symbolized as $a_j^{(i)}$, is equivalent to the activation function $h$ applied to the sum of functions $g_k$. The functions $g_k$ delineate the influence of different parameters and activations of other units on $a_j^{(i)}$.

In a feedforward DNN, $g_0$ represents the impact of the bias of unit $j$ in layer $i$, which is given by $g_0 = \theta_j^{(i)}$, and the weighted input from the previous layer is given by $g_1 = \sum_{k=1}^{N^{(i-1)}} \theta_{jk}^{(i)} a_k^{(i-1)}$. Here $N^{(i-1)}$ is the number of units in layer $i-1$.

The IMP method modifies the parameters $\theta$ of a deep neural network using the operator $\mathcal{T}$, which is a combination of two other operators $\mathcal{M}$ and $\mathcal{F}$. The resulting activation of unit $j$ in layer $i$ after the application of $\mathcal{I}$ is given by:

$$a_j'^{(i)} = h\left[\sum_k g_k(\mathbf{a}', \mathcal{F} \circ \mathcal{M}\boldsymbol{\theta})\right], \tag{41}$$

The projection operator $\mathcal{P}$ associated with $\mathcal{I}$ is defined as:

$$\mathcal{P}\left(a_j^{(i)}, a_j'^{(i)}\right) = \prod_{j=1}^{N} \delta\left\{a_j'^{(i)} - h\left[\sum_k g_k(\mathbf{a}', \mathcal{F} \circ \mathcal{M}\boldsymbol{\theta})\right]\right\} \tag{42}$$

Remarkably, this projection operator fulfills all three properties essential for a Renormalisation Group projection operator:
1) The product of Kronecker delta functions within $\mathcal{P}$ guarantees its non-negativity.
2) The Renormalisation Group projection operator, in preserving the inherent symmetry of the system, avoids introducing any new terms or couplings that did not exist originally. In the context of IMP, the operator merely removes the connections between units instead of introducing new forms of interaction. This preservation ensures the system's behaviour remains unaffected until a layer collapse occurs. In a layer collapse, all the weights interconnecting two layers are nullified. Preserving the original system symmetry is vital as pruning

should not alter the fundamental nature of the model under investigation.
3) In the context of the IMP method, this property can be satisfied by fixing the ordering of test and training samples during each epoch by setting a fixed random seed. This ensures that the masking and refining operations defined in the projection operator $\mathcal{P}\left(a_j^{(i)}, a_j^{\prime(i)}\right)$ are deterministic and produce unique results. Then $\sum_{\{s_I'\}} \mathcal{P}\left(s_i, s_I'\right) = 1.$ .

These observations imply that the Renormalisation Group theory serves as an apt language to examine IMP.

*Note:* Within the domain of artificial neural networks, the term "layer collapse" refers to a scenario where all the weights that connect two layers within the network become null, effectively disrupting the information passage through these layers. Consequently, these two layers virtually merge into one, reducing the network's total layers. When such an event occurs, the loss function and the activations of the units may undergo significant changes, potentially affecting the network's training and performance.

# 7 Identifying Gaps in the Connection Between Renormalisation Group Theory and IMP

While the connection between Renormalisation Group theory and Iterative Magnitude Pruning (IMP) has been established, several gaps remain that need further investigation. The following are a few key areas that require additional exploration:

- **Lack of Correlation Function Analogy:** There is currently no clear analogy for the correlation function as used in Renormalisation Group theory within the IMP framework. We speculated that this may be related to the degree of correlation among the activations of different units for a given data set.
- **Uniform Coupling Constants versus Diverse Parameters:** Renormalisation Group theory assumes that coupling constants are identical for each spin. In contrast, most Deep Neural Networks (DNNs) don't assign all parameters of the same type to the same value. This discrepancy poses a potential issue for drawing direct parallels between the two theories.
- **Absence of Effective IMP Theory:** At present, there is no comprehensive theory that outlines what percentage of IMP is optimal for a model to maintain its performance. Correspondingly, the Renormalisation Group theory doesn't offer a similar concept. The development of such a theoretical framework would provide greater insight into the application and limitations of IMP within DNNs.
- **Adaptability to Various Neural Network Architectures:** While IMP has been applied to different types of DNNs, it would be beneficial to investigate how the connection to Renormalisation Group theory extends to other types of neural network architectures, such as recurrent neural networks or generative adversarial networks.

- **Inclusion of Other Pruning Strategies:** IMP is just one of many pruning strategies for DNNs. It's worth investigating how Renormalisation Group theory could be applied to other pruning strategies and whether the current connection could be generalized.
- **Role of Initialisation:** The initialisation of the neural network weights plays a crucial role in the Lottery Ticket Hypothesis and hence in IMP. However, how this aspect correlates with the Renormalisation Group theory is not entirely understood and warrants further study.

# 8 Objective and Design of the Experiments

We have chosen two Hamiltonian neural networks (HNNs)[16] from "Hamiltonian Neural Networks for Solving Equations of Motion" by M. Mattheakis [25]. These networks have been designed to solve equations of motion for the Non-linear Oscillator system and a chaotic Hénon-Heiles dynamical system. Although both neural networks are composed of three layers, the last layer in the Hénon-Heiles system is twice the size of the final layer in the Non-linear Oscillator system.

Our experiment involves the application of the Iterative Magnitude-based Pruning (IMP) method to each layer of both systems, as well as the systems as a whole. Our primary focus is to investigate whether IMP manifests the power-law scaling predicted by renormalisation group theory and to scrutinise the associated critical exponents. To facilitate this, we will calculate the $\sigma$ (as defined in Eq. 52) for each layer. This will enable us to probe the similarities between the two systems and their architectural design, allowing us to anticipate the potential for transferability of their "winning tickets". Our ultimate objective is to identify these winning tickets for the two systems and evaluate whether their interchangeability aligns with our expectations based on the critical exponents and sigma values.

We anticipate some degree of transferability between the Non-linear Oscillator system and the Hénon-Heiles system. This expectation stems from the fact that both systems are non-linear dynamical systems [32] embodying energy conservation principles, making them ideal for modelling various physical phenomena. Such similarities suggest that these systems could belong to the same equivalence class under the purview of Renormalisation Group theory.

For those interested in a deeper exploration of our study, our codebase and raw data are available at [17]. While the universality of winning tickets among neural networks has been the subject of past studies, our work is novel in its specific application to Hamiltonian neural networks employed to solve differential equations [8] [4].

## 8.1 Introduction to Hamiltonian Neural Networks

Hamiltonian neural networks (HNNs), as proposed in the literature [25] [16] [7], introduce a novel approach to solving differential equations that describe dynamical systems. The Hamiltonian is a function that encapsulates the total energy (kinetic plus potential) of a

physical system. By using the Hamiltonian as a guiding principle, HNNs learn to predict the evolution of a system over time.

The HNN architecture is comprised of neural networks that approximate the kinetic and potential energies of the system, as well as the gradients of these energies with respect to position and momentum variables. It's structured to inherently conserve the Hamiltonian, which is a desirable property in many physics problems where energy conservation [1] is paramount. This conservation property can lead to more accurate and stable simulations compared to traditional neural networks that do not conserve energy.

## 8.2 IMP experiment 1. Nonlinear Oscillator

The experiment is with a one-dimensional nonlinear oscillator with hamiltonian:

$$H(x,p) = \frac{p^2}{2} + \frac{x^2}{2} + \frac{x^4}{2}, \tag{43}$$

We assume that the natural frequency and the mass of the oscillator are considered to be unity. The corresponding equations that govern motion are:

$$\dot{x} = p, \qquad \dot{p} = -(x + x^3) \tag{44}$$

The loss function is the mean squared error:

$$L = \frac{1}{K}\sum_{n=1}^{K}\left[\left(\dot{\hat{x}}^{(n)} - \hat{p}^{(n)}\right)^2 + \left(\dot{\hat{p}}^{(n)} + \hat{x}^{(n)} + \left(\hat{x}^{(n)}\right)^3\right)^2\right] \tag{45}$$

The hyper-parameters of the neural network when getting trained were set the same as in the paper. This neural network has 3 layers. The first and the hidden layer each have 50 neurons, and the output layer has 2 neurons. We use a learning rate of $8 \times 10^{-3}$ and trained for $5 \times 10^4$ epochs.

I ran IMP pruning experiments at 1%, 5% and 10%, where we only pruned certain layers of the neural network and not the whole model. Pruning at different percentages is useful as it helps us investigate what percentage IMP is most effective when we carry out pruning of the entire model later. We start by pruning individual layers rather than the full model to learn about the layer-specific complexity as some layers might contain more redundancy or less critical information than others, which may warrant different pruning strategies.

The graphs below are average across many runs of the experiment until the graph stopped changing significantly.

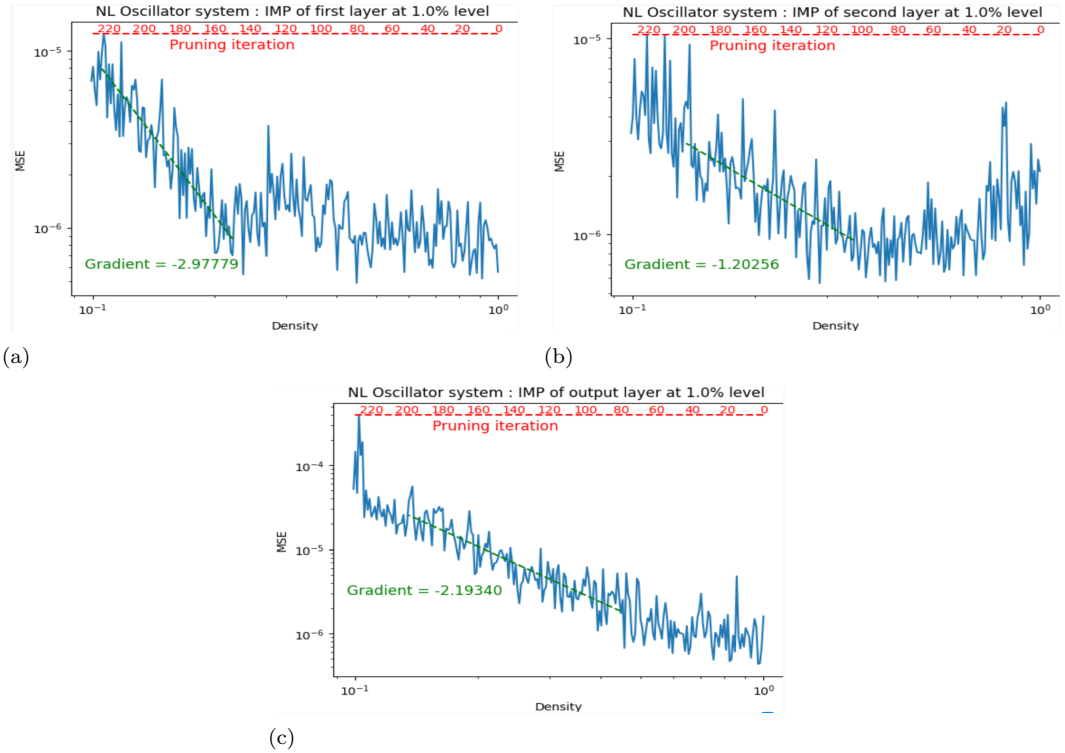


Figure 1: IMP at 1% performed on each layer separately

From Figure 1 (a) and (b) we notice that we can prune 60% of the input layer and hidden layer before we start to see the power-law scaling that we expect from IMP being a renormalisation group. The effectiveness of IMP is very apparent for the hidden layer as the MSE actually consistently decreases while we prune 30% of the hidden layer. This layer-specific behavior aligns with the findings of Zhang et al. [36], who proposed a distinction between 'robust' and 'critical' layers in a deep neural network. Our first and hidden layers may be viewed as 'robust', with redundant representation capability, enabling the remaining neurons to compensate even after significant pruning. Conversely, the output layer, which begins to show power-law scaling after 25% pruning, could be considered more 'critical' [36].

| | Critical exponent at 1% IMP |
|---|---|
| NL Oscillator neural network only input layer pruning | 2.9777 |
| NL Oscillator neural network only hidden layer pruning | 1.2025 |
| NL Oscillator neural network only output layer pruning | 2.1934 |

Table 1: Critical exponents for full model pruning. We work these out by taking the negative of the linear regression on the power-law scaling critical region as explained in section 4.1.

The critical exponent of the input layer and the output layer is significantly larger than that of the hidden layer. This suggests that the input layer is the most important to the model's predictions and the output layer is the second most important to the model predictions. This is because the input layer is the primary feature extractor from the input and the output layer is the decision-making layer.

Now the results from layer-specific pruning for IMP at 5%.

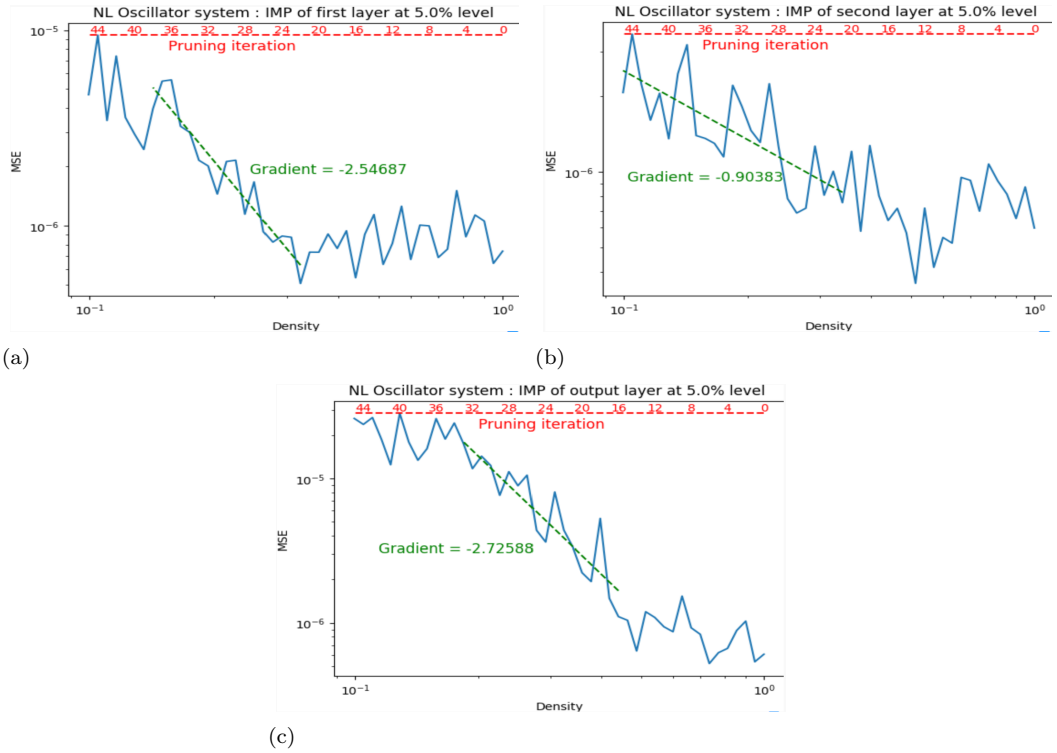


Figure 2: IMP at 5% performed on each layer separately

The critical exponents have changed slightly as now the output layer has the largest critical exponent (2.7258) and the input layer has the second largest critical exponent (0.9038), This however could just be a bad estimation due to the small amount of data points. Overall, we still see a strong power-law scaling and the critical region of power-law scaling exists at similar densities as that of IMP at 1%.

The results for IMP at 10% level:

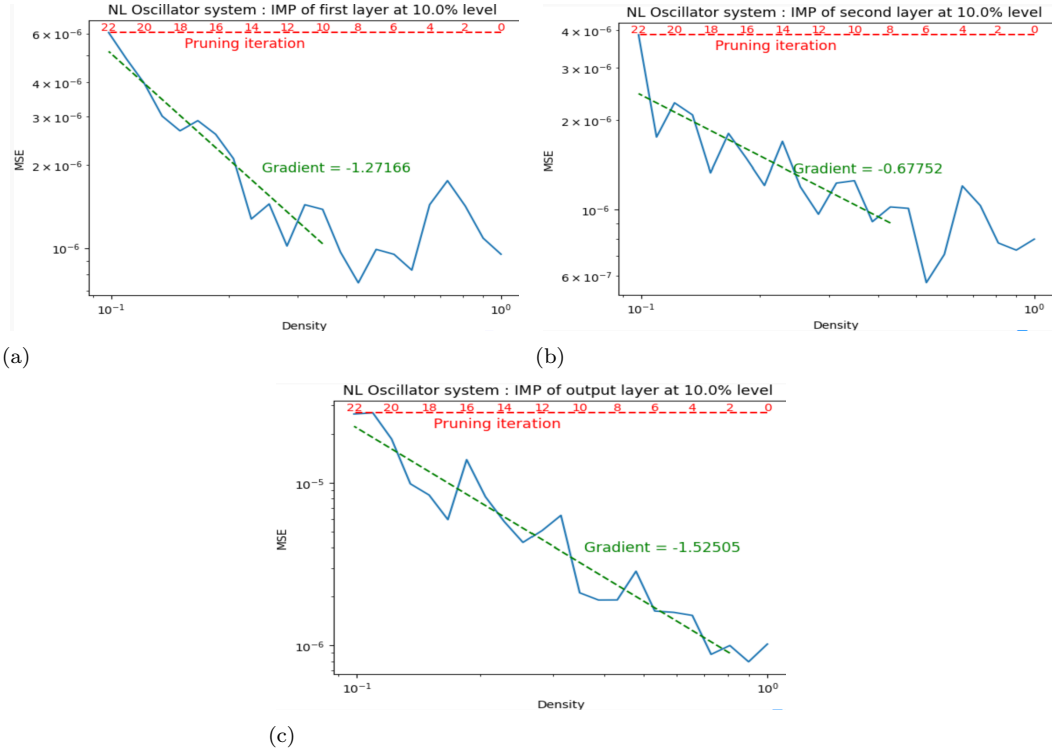


Figure 3: IMP at 10% performed on each layer separately

Our results have provided empirical evidence that power-law scaling in IMP exists at 1%, 5% and 10% levels. The critical exponents change with different levels of IMP but overall the input and output layer is the most important. IMP at 1%, 5% and 10% took 230, 41 and 22 pruning iterations respectively to prune the layers by 90%.

We notice that the critical densities where power-law scaling is observed remain similar across all IMP percentage levels. Furthermore IMP at 1% level helps estimate the critical exponents most accurately due to the large number of data points available.

We decided that to investigate the transferability of winning tickets we will employ IMP at 1% for experiments henceforth as even though this takes roughly 5 times the computational times as IMP at 5% level, it can potentially lead to discovering more winning tickets due to greater pruning iterations.

The pruning of the output layer demonstrated the most stability, evidenced by its mean square error (MSE) line adhering closely to the linear regression line across all pruning percentages. This intriguing finding implies that each layer of a deep neural network may exhibit different levels of tolerance to sparsity. Therefore, a more effective pruning strategy for the discovery of 'winning tickets' may involve pruning the smallest weights uniformly across all layers. This approach capitalizes on Iterative Magnitude Pruning (IMP) to maintain optimal model performance.

One must remember that in Deep Neural Networks (DNNs), layers are not standalone entities. They are intricately interconnected, with alterations to one potentially triggering ripple effects across others. In this regard, pruning the network as a whole, rather than selectively pruning individual layers, effectively accounts for these layer-to-layer interactions, thereby ensuring optimal performance.

When we apply the lens of renormalization group theory to the pruning process of a deep neural network, interesting parallels begin to emerge. If we consider the entirety of the network, this aligns with the 3D Ising model. The depth of the neural network can be likened to the layers of a 3D Ising model, almost as if they were sheets of 2D Ising models stacked upon each other. On the other hand, pruning specific layers in isolation resonates with the 2D Ising model. Given the complex interactivity among layers in a neural network, the 3D Ising model offers a more realistic analogy. Therefore, we posit that a more comprehensive pruning strategy encompassing the entire model would not only be more appropriate but also potentially more successful in uncovering those elusive 'winning tickets'.

Presented below is the average results graph for the pruning of the entire model:

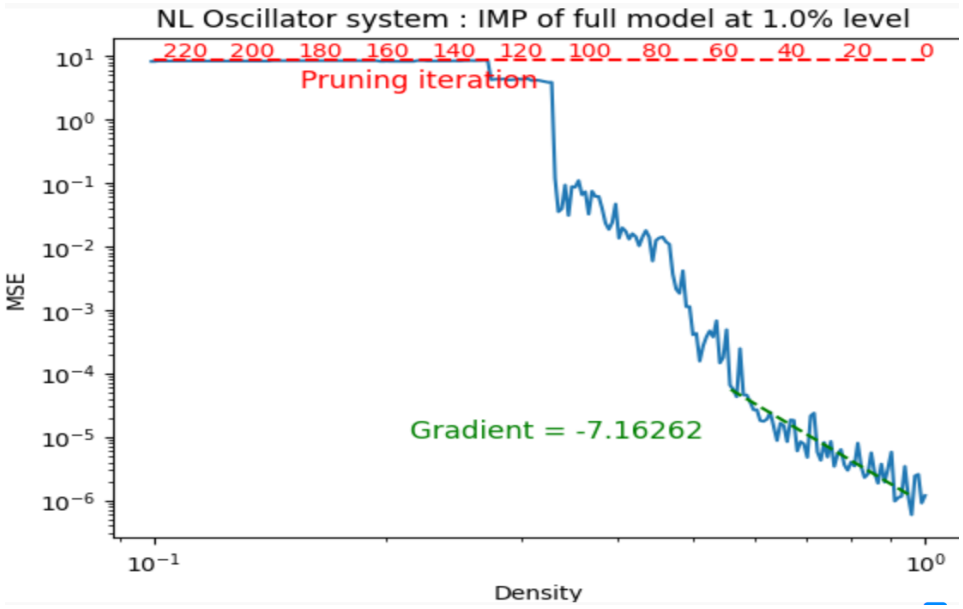


Figure 4: IMP the full NL oscillator model

As observed in Figure 4, we find that for densities below 0.9, the error begins to increase, exhibiting a power-law relationship. By scrutinizing each experiment's outcomes individually, we accumulate compelling empirical evidence to suggest that for densities below 0.9, winning tickets for the Nonlinear Oscillator model are non-existent, regardless of the initial parameters at which the IMP process is commenced.

This prompts us to shift our investigative lens from focusing exclusively on winning tickets

to exploring the transferability of all subnetworks obtained from the original network. This includes subnetworks that may not necessarily show high performance on the original task but may exhibit interesting properties, such as improved performance upon transfer and fine-tuning on a different task as highlighted in the recent work by Fu et al. [13]. Thus, our investigation now expands to not just those tickets that win outright but all tickets with potential, including those that may only reveal their worth across different tasks.

Later in this discussion, we will derive the $\sigma$ for both the Nonlinear Oscillator and the Henon-Heiles (HH) system, as delineated in Section 7.2. This will enable us to ascertain the level of expected transferability between the winning tickets of both the Nonlinear Oscillator and HH system.

### 8.3 IMP experiment 2. Chaotic Hénon-Heiles dynamical system

Now we address the Hénon-Heiles system, which models a star's nonlinear motion around a galactic centre, confined to a plane. This system has four degrees of freedom in phase space, represented as z = (x, y, $p_x$, $p_y$). The system's Hamiltonian is:

$$H(x, y, p_x, p_y) = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(x^2 + y^2) + (x^2 y - \frac{y^3}{3}), \tag{46}$$

Hamilton's equations lead to a system of nonlinear differential equations:

$$\dot{x} = p_x \qquad \dot{y} = p_y \tag{47}$$

$$\dot{p_x} = -(x + 2xy) \qquad \dot{p_y} = -(y + x^2 - y^2) \tag{48}$$

The loss function of the HH system's neural network is:

$$\begin{aligned} L = \frac{1}{K}\sum_{n=0}^{K}\Bigg[&\left(\dot{\hat{x}}^{(n)} - \hat{p}_x^{(n)}\right)^2 + \left(\dot{\hat{y}}^{(n)} - \hat{p}_y^{(n)}\right)^2 + \left(\dot{\hat{p}}_x^{(n)} + \hat{x}^{(n)} + 2\hat{x}^{(n)}\hat{y}^{(n)}\right)^2 \\ &+ \left(\dot{\hat{p}}_y^{(n)} + \hat{y}^{(n)} + \left(\hat{x}^{(n)}\right)^2 - \left(\hat{y}^{(n)}\right)^2\right)^2\Bigg] \end{aligned} \tag{49}$$

The hyper-parameters of the neural network again when getting trained were set the same as in the paper. This neural network has 3 layers. The first and the hidden layers have 50 neurons, and the output layer has 4 neurons. We use a learning rate of $8 \times 10^{-3}$ and trained for $2 \times 10^4$ epochs.
IMP of the HH system at 1% gives:

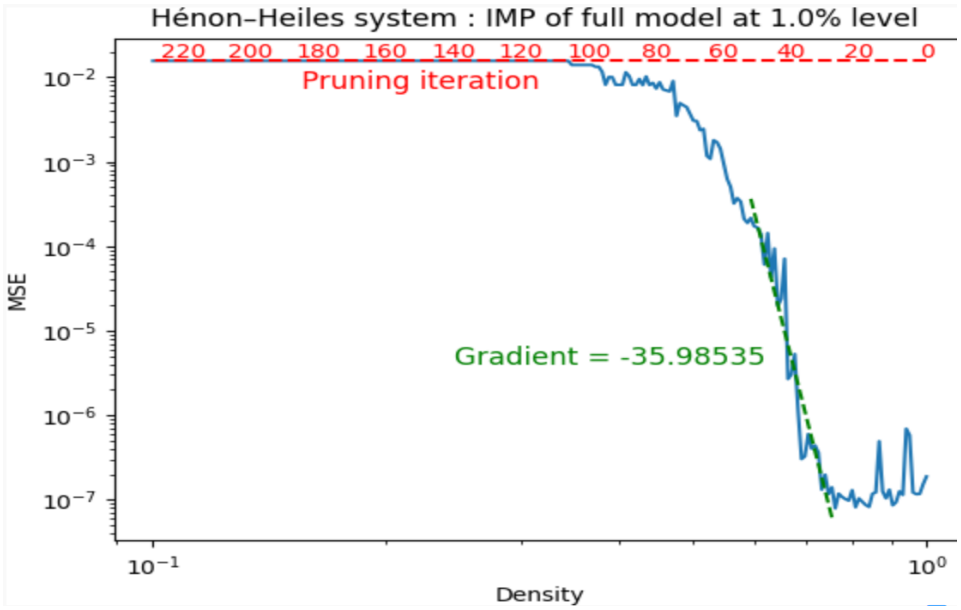


Figure 5: IMP the full HH system model

Figure 5 shows that the HH system is able to produce winning tickets for densities between 1 and 0.75 whereas Figure 4 shows that NL oscillator system is only capable of producing winning tickets for densities 1 to 0.9. The power law scaling is also much steeper for HH system when compared to the NL oscillator (-35.9853 vs -7.1626). The distinctions between the two systems and architectures are beginning to manifest themselves within our graphical representations. Since now we are pruning the entire model, we can determine whether tickets are transferrable between them by comparing the *sigma*s of the two models as described in section 7.2.

# 9 Tranferability of winning ticket between Hénon-Heiles system HNN and Nonlinear Oscillator HNN

In this section, we will be conducting a pair of experiments to investigate the transferability of winning tickets between the Hénon-Heiles (HH) and Nonlinear Oscillator systems. Firstly, we will transfer masks derived from the Iterative Magnitude Pruning (IMP) process of the Nonlinear Oscillator to the HH system, evaluating their performance. We will then reciprocate this procedure by transferring the mask derived from IMP of the NL oscillator HNN to the HH system HNN.

To anticipate the outcomes of the transferability experiments, we will initially calculate the $\sigma$ values, as detailed in Section 7.2.

| | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ |
|---|---|---|---|
| NL Oscillator neural network full model pruning | 2.4778 | -0.1710 | 5.2449 |
| HH system neural network full model pruning | 2.6447 | -0.1205 | 4.4276 |

Table 2: Sigmas: approximating the influence of the parameters of each layer

Observing the table, it is evident that both $\sigma_1$ and $\sigma_3$ are positive for both models, thus corresponding to relevant directions. Conversely, $\sigma_2$ for both models is negative, indicating an irrelevant direction. From the perspective of Renormalisation Group flow, both models share similar relevant and irrelevant directions.

The magnitudes of $\sigma_1$ and $\sigma_2$ for both models exhibit concordance up to one significant figure. However, there is a difference in $\sigma_3$ for the Nonlinear Oscillator and the HH system. This discrepancy suggests potential variations in the scaling behaviour along this direction.

Given the similar relevant and irrelevant directions, we anticipate some degree of transferability between the two models. However, the variations in $\sigma_3$ could introduce unique behaviour during the transferability experiments.

## 9.1 Transfer from HH system to NL Oscillator system in HNNs

Transferring masks between the HH system and the NL oscillator faces a challenge due to the mismatch between the output layers of the two systems, with the NL oscillator having 2 neurons. To address this, we employ scaling, making transferability feasible by duplicating the mask of the NL oscillator when transferring the tickets to the HH system.

The training epochs differ between the NL oscillator neural network ($5 \times 10^4$) and the HH system neural networks ($2 \times 10^4$). This divergence arises from training each system to their respective natural convergence levels, rather than aligning them to a uniform convergence level. The HH system neural network naturally achieves a lower MSE (reaching $10^{-7}$) than the NL oscillator neural network (reaching $10^{-6}$). However, the transferability of the winning ticket is more linked to the network architecture [11] — specifically, the initial structure and weights — rather than the nuances of the training process, such as learning rate and number of epochs. Consequently, differences in the training process do not adversely affect our research.

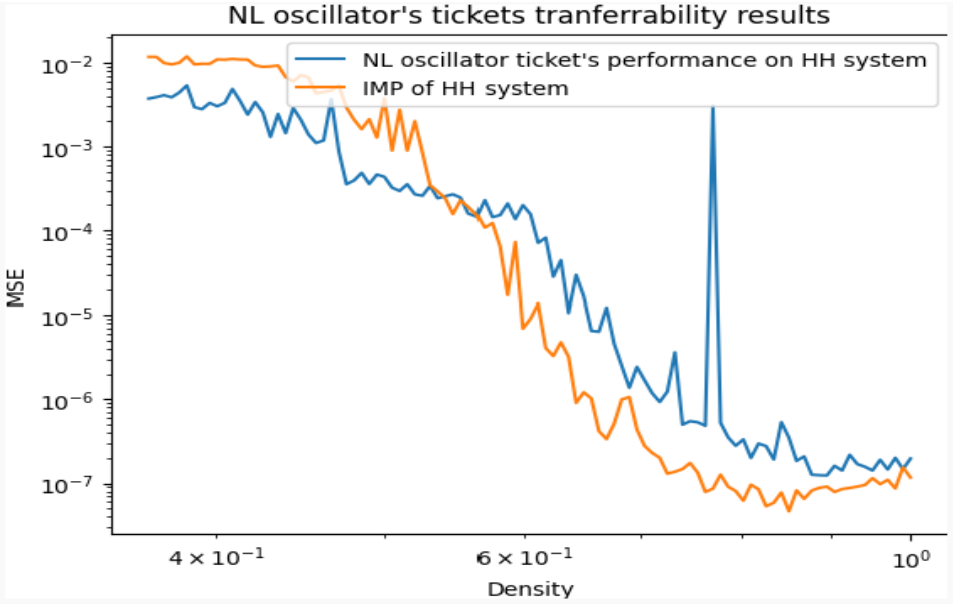


Figure 6: Transferability of winning tickets: The plot showcases the comparative performance of winning ticket transfer from the Nonlinear Oscillator (NL) model to the Henon-Heiles (HH) system (blue line), against the standard performance of the iterative magnitude pruning (IMP) directly applied on the HH system (orange line).

Initially, we notice a slight divergence between the two lines within a density range of 1 to 0.75. This disparity gradually intensifies and remains relatively constant throughout the majority of pruning iterations. Intriguingly, around a density of 0.65, the lines exhibit a brief period of intersection. Subsequently, the lines diverge once again, reverting to a similar spacing as observed earlier. In this latter stage, the transferred tickets showcase superior performance compared to the direct application of IMP on the HH system.

For densities between 1 and 0.75, the transfer of winning tickets proves successful. This outcome is noteworthy since winning tickets have traditionally been identified for high-dimensional problems with millions of parameters [29] [11]. In contrast, we have demonstrated that winning tickets can be transferred for low-dimensional problems, like our two nonlinear dynamical problem Hamiltonian Neural Networks (HNNs), which contain fewer than 3000 parameters.

Interesting phenomena occur for densities below 0.7. For these densities, the MSE is at least tenfold that of the full HH system, rendering the tickets below a density of 0.7 ineligible as winning tickets. Interestingly, for densities lower than 0.65, the transferred tickets perform better on the HH system than direct IMP on the HH system.

## 9.2 Transfer from NL Oscillator System to HH System in HNNs

Addressing the discrepancy in the output layer of the Hénon-Heiles (HH) system, with 4 neurons, and the Nonlinear Oscillator (NL) system, with 2 neurons, is critical for successful

mask transfer. To fit a 4 by 50 weights mask onto a 2 by 50 weight space, we truncate the second and fourth rows before initiating the transfer process.

We present our results on the transferability of masks from the HH system to the NL system in Figure 7.

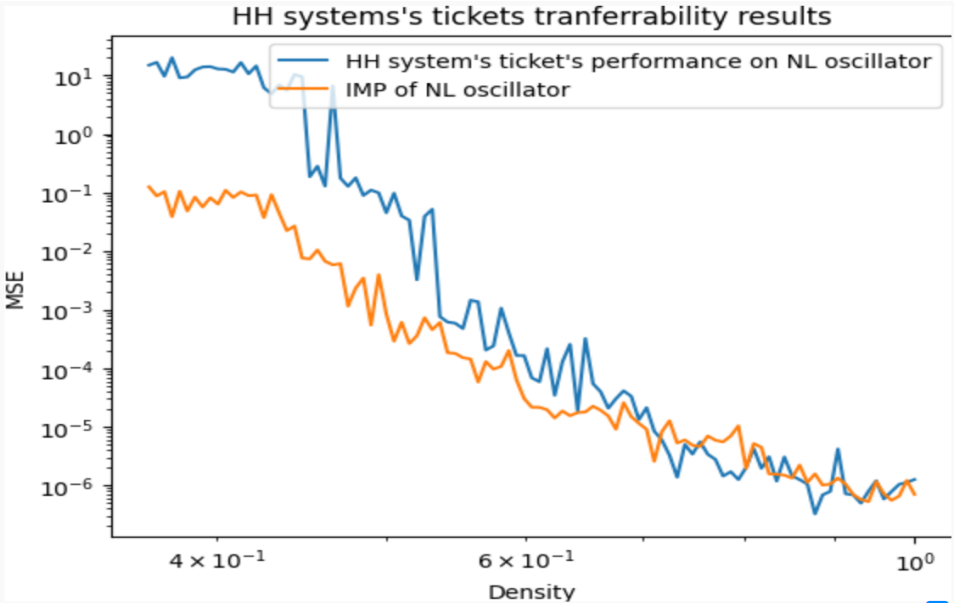


Figure 7: Transferability of winning tickets from HH system to NL system: Performance comparison between direct iterative magnitude pruning (IMP) on the NL system and transferred tickets from HH system.

Interestingly, the transfer of tickets from the HH system to the NL oscillator displays superior efficacy compared to the reciprocal experiment, as indicated by the closer proximity of the performance curves.

Winning tickets are identified for densities ranging from 1 to 0.7. Even for densities below this range, the transferred tickets perform comparably to typical IMP performance on the NL oscillator, down to densities of 0.65. For densities lower than 0.65, the transferred tickets yield an MSE 10 to 20 times larger than the typical IMP of the NL oscillator.

The improved transferability of tickets from the HH system to the NL oscillator could potentially be attributed to the truncation of the output layer mask, which eliminates redundancy. This suggests that the transfer process may favour an absence of redundant information over the repetition of masks.

# 10 Conclusion

In this thesis, we embarked on a journey through the realms of the lottery ticket hypothesis (LTH), iterative magnitude pruning (IMP), and Hamiltonian Neural Networks (HNNs),

within the context of low-dimensional physics problems.

Our work provides solid evidence supporting the existence of winning tickets in smaller neural networks used for low-dimensional problems. This observation, which complements and extends the findings of Frankle and Carbin's seminal work [11], broadens the scope of LTH. The research results also imply that the concept of winning tickets transcends the high-dimensional landscape where it was initially identified, permeating into the realm of lower-dimensional physics problems.

The application of iterative magnitude pruning to two distinctly different systems — the nonlinear oscillator and the Hénon-Heiles system — yielded fascinating insights into the relationship between the initial network architecture and the system it models. We observed that smaller is indeed often better when pruning, suggesting potential avenues for efficient model development and refinement.

Notably, we ventured into the largely unexplored territory of transferring winning tickets across different systems and model architectures. Our findings underscore the potential of such transfers and point to a rich seam of research possibilities.

Furthermore, our work lends new perspectives to understanding the LTH through the lens of renormalisation group theory, suggesting an intriguing connection between these two areas. The parallel drawn between the behaviour of winning tickets and the universality phenomena of renormalisation group theory underscores the underlying order beneath the seeming chaos of deep learning model training.

In conclusion, we hope our exploration inspires further research into the lottery ticket hypothesis, its manifestations in different domains, and its intriguing ties with renormalisation group theory. As we continue to unravel the mysteries of neural networks, we move closer to the creation of more efficient, effective, and transformative artificial intelligence technologies.

# 11 Directions for Future Research

In light of the findings and conclusions reached within this study, we propose the following promising directions for future research:

- **Optimizing Pruning Percentage:** While our study relied on a predetermined pruning rate, future research could explore methodologies for determining the optimal pruning percentage for a given model. The question arises whether smaller pruning percentages always yield better performance. As such, a systematic investigation into the effects of various pruning rates on the lottery ticket hypothesis and the transferability of tickets would be valuable.
- **Adaptability of Masks Across Architectures:** We managed to transfer masks between models of differing architectures by duplicating the mask of the smaller ar-

chitecture. However, the question remains whether other strategies may yield superior results. For instance, when transferring from a larger to a smaller network, would it be advantageous to truncate the mask, average it, or find a maximum? Alternatively, could we develop a mapping function to make this transfer more effective? And how should we handle the reverse scenario, transferring from a smaller to a larger system?

- **Bridging the Gap between IMP and RG Theory:** As we mentioned in Section 7, our current work has only just begun to draw parallels between the lottery ticket hypothesis and the Renormalization Group (RG) theory. Additional research is required to further elucidate the connections and commonalities between these two paradigms.
- **Restricting Pruning to Hidden Layers:** Our pruning strategy included the input and output layers. A worthwhile avenue for exploration is whether constraining pruning to just the hidden layers might yield superior performance in the identification of winning tickets.
- **Exploring Different Architectures:** Our work to date has primarily focused on Hamiltonian Neural Networks (HNNs). Future work could expand this focus to include other architectures. In particular, the Deep Operator Network (DeepONet) architecture, introduced by Lu et al. [23], appears particularly promising. With its capacity for learning both explicit and implicit operators, DeepONet may offer substantial improvements in the efficiency and accuracy of solutions to the equations of motion in our HNN models.

We study the IMP flow in directly by estimating the relative "influence" the parameters of a given layer have on the full NN. This can be done by considering the total remaining parameter magnitude that remains in layer $i$ after $n$ applications of IMP [29, p. 6]:

$$M_i(n) = \frac{\sum_{j=1}^{N^{(i)}} \left| m_j^{(i)}(n) \cdot \theta_j^{(i)}(n) \right|}{\sum_{k=1}^{N} |m_k(n) \cdot \theta_k(n)|} \tag{50}$$

Here $N^(i)$ is the number of parameters in layer $i$ and $m^{(i)} \in \{0,1\}^{N^{(i)}}$ is the pruning mask. If we are to consider $M_i(n)$ as eigenfunctions of the IMP operator they should scale exponentially with respect to the number of IMP iterations. As $M_i(n+1) = \mathcal{T} M_i(n) = \lambda_i M_i(n) = \lambda_i^{n+1} M(0)$. We can drive $\lambda_i$ as:

$$\lambda_i = \frac{M_i(n+1)}{M_i(n)} \tag{51}$$

The degree of coarse-graining ($x \in (0,1)$) at each iteration of IMP affects the magnitude of the eigenvalues. Therefore we are interested in the quantity $\sigma$:

$$\lambda_i \sim c^{\sigma_i} \tag{52}$$

Where $\sigma$ is invariant to the choice of c and taking $log_c(\lambda_i)$ gives $\sigma_i$.

# References


[1] D. Ba, A. S. Dogra, R. Gambhir, A. Tasissa, and J. Thaler. Shaper: Can you hear the shape of a jet? *Journal of High Energy Physics*. URL: https://arxiv.org/abs/2302.12266.

[2] A. Balwani and J. Krzyston. Zeroth-order topological insights into iterative magnitude pruning, 2022. arXiv:2206.06563.

[3] C. Bény. Deep learning and the renormalization group, 2013. arXiv:1301.3124.

[4] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] X. Chen, Y. Cheng, S. Wang, Z. Gan, J. Liu, and Z. Wang. The elastic lottery ticket hypothesis. *CoRR*, abs/2103.16547, 2021. URL: https://arxiv.org/abs/2103.16547, arXiv:2103.16547.

[6] E. J. Crowley, J. Turner, A. J. Storkey, and M. F. P. O'Boyle. Pruning neural networks: is it time to nip it in the bud? *ArXiv*, abs/1810.04622, 2018.

[7] A. S. Dogra. Dynamical systems and neural networks, 2020. arXiv:2004.11826.

[8] A. S. Dogra, J. B. Lai, and M. Peev. Neural network differential equation solvers allow unsupervised error analysis and correction. *arXiv*, 2023.

[9] A. S. Dogra and W. T. Redman. Optimizing neural networks via koopman operator theory. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[10] B. Elesedy, V. Kanade, and Y. W. Teh. Lottery tickets in linear models: An analysis of iterative magnitude pruning. *CoRR*, abs/2007.08243, 2020. URL: https://arxiv.org/abs/2007.08243, arXiv:2007.08243.

[11] J. Frankle and M. Carbin. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635, 2018. URL: http://arxiv.org/abs/1803.03635, arXiv:1803.03635.

[12] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. The lottery ticket hypothesis at scale. *CoRR*, abs/1903.01611, 2019. URL: http://arxiv.org/abs/1903.01611, arXiv:1903.01611.

[13] Y. Fu, Y. Yuan, S. Wu, J. Yuan, and Y. Lin. Robust tickets can transfer better: Drawing more transferable subnetworks in transfer learning, 2023. arXiv:2304.11834.

[14] N. Goldenfeld. *Lectures on Phase Transitions and the Renormalization Group*. CRC Press, 2018.

[15] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[16] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *CoRR*, abs/1906.01563, 2019. URL: http://arxiv.org/abs/1906.01563, arXiv:1906.01563.

[17] A. A. Hassan. Abu-pruning-hamiltonian-nn. https://github.com/MathePhysics/Abu-pruning-Hamiltonian-NN, 2023.

[18] L. P. Kadanoff. *Statics, Dynamics, and Renormalization*. World Scientific Publishing Co Inc, 2000.

[19] A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Proceedings: Mathematical and Physical Sciences*, 434(1890):9–13, 1991. URL: http://www.jstor.org/stable/51980.

[20] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[21] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient convnets. *CoRR*, abs/1608.08710, 2016. URL: http://arxiv.org/abs/1608.08710, arXiv:1608.08710.

[22] H. W. Lin, M. Tegmark, and D. Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, jul 2017. URL: https://doi.org/10.1007%2Fs10955-017-1836-5, doi:10.1007/s10955-017-1836-5.

[23] L. Lu, P. Jin, and G. E. Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *CoRR*, abs/1910.03193, 2019. URL: http://arxiv.org/abs/1910.03193, arXiv:1910.03193.

[24] J. Maene, M. Li, and M. Moens. Towards understanding iterative magnitude pruning: Why lottery tickets win. *CoRR*, abs/2106.06955, 2021. URL: https://arxiv.org/abs/2106.06955, arXiv:2106.06955.

[25] M. Mattheakis, D. Sondak, A. S. Dogra, and P. Protopapas. Hamiltonian neural networks for solving equations of motion. *Physical Review E*, 105(6), jun 2022. URL: https://doi.org/10.1103%2Fphysreve.105.065305, doi:10.1103/physreve.105.065305.

[26] P. Mehta and D. J. Schwab. An exact mapping between the variational renormalization group and deep learning, 2014. arXiv:1410.3831.

[27] A. S. Morcos, H. Yu, M. Paganini, and Y. Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers, 2019. arXiv:1906.02773.

[28] M. Paul, F. Chen, B. W. Larsen, J. Frankle, S. Ganguli, and G. K. Dziugaite. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? In *The Eleventh International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=xSsW2Am-ukZ.

[29] W. T. Redman, T. Chen, A. S. Dogra, and Z. Wang. Universality of deep neural network lottery tickets: A renormalization group perspective. *CoRR*, abs/2110.03210, 2021. URL: https://arxiv.org/abs/2110.03210, arXiv:2110.03210.

[30] J. S. Rosenfeld, J. Frankle, M. Carbin, and N. Shavit. On the predictability of pruning across scales. *CoRR*, abs/2006.10621, 2020. URL: https://arxiv.org/abs/2006.10621, arXiv:2006.10621.

[31] M. Sabatelli, M. Kestemont, and P. Geurts. On the transferability of winning tickets in non-natural image datasets. *CoRR*, abs/2005.05232, 2020. URL: https://arxiv.org/abs/2005.05232, arXiv:2005.05232.

[32] S. H. Strogatz. Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering. *Studies in Nonlinearity*, 1994.

[33] K. G. Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Phys. Rev. B*, 4:3174–3183, Nov 1971. URL: https://link.aps.org/doi/10.1103/PhysRevB.4.3174, doi:10.1103/PhysRevB.4.3174.

[34] K. G. Wilson. The renormalization group: Critical phenomena and the kondo problem. *Rev. Mod. Phys.*, 47:773–840, Oct 1975. URL: https://link.aps.org/doi/10.1103/RevModPhys.47.773, doi:10.1103/RevModPhys.47.773.

[35] H. Yang and Z. Wang. On the neural tangent kernel analysis of randomly pruned neural networks. 2023. URL: https://arxiv.org/pdf/2203.14328.pdf.

[36] C. Zhang, S. Bengio, and Y. Singer. Are all layers created equal? *Journal of Machine Learning Research*, 2020. URL: https://www.jmlr.org/papers/volume23/20-069/20-069.pdf.

[37] H. Zhou, J. Lan, R. Liu, and J. Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. *CoRR*, abs/1905.01067, 2019. URL: http://arxiv.org/abs/1905.01067, arXiv:1905.01067.