Quantifying lottery tickets under label noise: accuracy, calibration, and complexity

Viplove Arora[1] **Daniele Irto**[2] **Sebastian Goldt**[1] **Guido Sanguinetti**[1]

[1]Theoretical and Scientific Data Science, SISSA, Trieste, Italy
[2]Data Science and Scientific Computing, University of Trieste, Trieste, Italy

Abstract

Pruning deep neural networks is a widely used strategy to alleviate the computational burden in machine learning. Overwhelming empirical evidence suggests that pruned models retain very high accuracy even with a tiny fraction of parameters. However, relatively little work has gone into characterising the small pruned networks obtained, beyond a measure of their accuracy. In this paper, we use the sparse double descent approach to identify univocally and characterise pruned models associated with classification tasks. We observe empirically that, for a given task, iterative magnitude pruning (IMP) tends to converge to networks of comparable sizes even when starting from full networks with sizes ranging over orders of magnitude. We analyse the best pruned models in a controlled experimental setup and show that their number of parameters reflects task difficulty and that they are much better than full networks at capturing the true conditional probability distribution of the labels. On real data, we similarly observe that pruned models are less prone to overconfident predictions. Our results suggest that pruned models obtained via IMP not only have advantageous computational properties but also provide a better representation of uncertainty in learning.

1 INTRODUCTION

Conventional statistical wisdom suggests that increasing the size of a model leads to an initial improvement in generalisation performance, followed by dramatic overfitting and degradation of accuracy in highly overparameterised models. In reality, the implicit regularisation of large models leads to a double descent which, under suitable conditions, leads to overparameterised models with even stronger generalisation performance [Vallet et al., 1989, Opper et al., 1990, Geman et al., 1992, Belkin et al., 2019, Nakkiran et al., 2021, Loog et al., 2020]. This phenomenon is well understood theoretically in simple cases and has been replicated in practice in a variety of modern deep neural networks architectures [Belkin et al., 2019, Nakkiran et al., 2021, Arpit et al., 2017, Neyshabur et al., 2019, Belkin et al., 2020].

However, decades of research on pruning neural networks [LeCun et al., 1989, Hassibi and Stork, 1992, Gale et al., 2019, Hoefler et al., 2021] has also shown that a large number of parameters (weights) in these overparameterized models can be removed without compromising on the generalisation error. He et al. [2022] recently reported that iterative pruning leads to a converse phenomenon to double descent, a *sparse double descent*: generalisation performance initially degrades upon pruning, and then improves to reach an optimum frequently providing even better performance than the original full model. The sparse double descent allows us to identify an optimal model by plotting the generalisation error as a function of the number of parameters during the iterative pruning process. How the architectural bias induced by pruning enables the networks to reverse the double descent curve is not understood theoretically, and is relatively unexplored empirically.

Here we seek to address this knowledge gap in the iterative magnitude pruning (IMP) framework [Han et al., 2015], a simple and successful approach for pruning deep neural networks [Blalock et al., 2020, Renda et al., 2020] which is pivotal to finding "lottery tickets" [Frankle and Carbin, 2019, Frankle et al., 2019, 2020], i.e. sparse subnetworks that can be trained in isolation without compromising on accuracy. We start with a remarkable empirical observation. Figure 1 shows the double descent curve (thick line) and the sparse double descent curves (thin lines) for fully connected networks trained on MNIST (left) and a ResNet-18 trained on CIFAR-10 (right). Naturally, the starting point of the double descent curve is fixed at the smallest possible model. He et al. [2022] considered sparse double descent starting from a single, large model. Here, we consider sparse double

Accepted for the 39th *Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

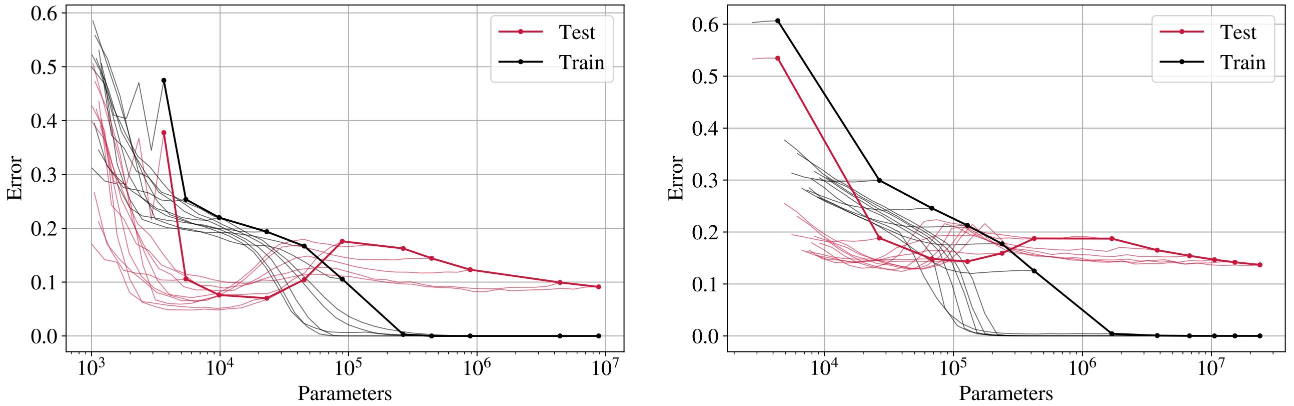


Figure 1: Pruning models along the double descent curve (dark red) shows that sparse double descent curves (light red) from different models coincide at the minima. Results are shown for three-layer FCNs on MNIST (left) and ResNet-18 on CIFAR-10 (right) with 20% label noise averaged over three replicates.

descent curves starting from neural networks with a large range of parameters. We show in fig. 1 that irrespective of the initial full model size (or indeed test accuracy), all sparse double descent curves tend to achieve their minimum (best pruned model) within a very narrow range of model sizes, which is systematically smaller than the trough of the first descent (best small full model). What is special about such sparse models? What enables the improvement in generalisation by pruning?

To answer this question quantitatively, we investigate the behaviour of the best pruned models in a controlled binary classification scenario where data is generated from distributions of differing complexity. We find that the size of the best pruned model correlates well with task complexity. Moreover, best pruned models are considerably better at capturing the underlying true conditional label distribution than either full models (which tend to be significantly overconfident) or best small full models. These empirical observations are replicated in our analysis of real data, suggesting that IMP indeed captures important aspects of the uncertainty in the learning problem, as well as producing computationally more tractable models.

We see a key contribution of our empirical study in suggesting a way to estimate the number of "effective" parameters that a trained model contains, and that are required for solving a classification task. Quantifying the effective number of parameters in trained neural networks has been a central question for the theory of neural networks for a long time, see for example Breiman [1995], yet it remains an open challenge. In this manuscript, we focus on the empirical findings and leave the theoretical analysis for future work. Our main contributions are:

1. We find that IMP prunes models of different sizes to produce sparse double descent curves that coincide at the minima of the test error. We define the *effective number of parameters* required for a given classification task and model/architecture. We demonstrate this phenomenon in fully-connected and convolutional neural networks trained on MNIST and CIFAR-10. Additionally, the effective number of parameters is comparable across architectures, suggesting that our procedure identifies an intrinsic property of the classification task.
2. By training neural networks on binary Gaussian mixture classification tasks of increasing difficulty, we show that the effective number of parameters in a model correlates with task difficulty.
3. We finally study the calibration of pruned models and show that pruned models capture the true distribution of the synthetic data models better than their unpruned counterparts. As a consequence, we show that the best pruned models are better at capturing uncertainty in their predictions.

2 RELATED WORK

Dating back to 1998, Poppi and Massart [1998] used pruning to find the optimal neural network architecture. They found that pruning could improve generalisation and even recover the optimal model in a linear dataset. Kuhn et al. [2021] introduces a new complexity measure for neural networks, namely the fraction of weights that can be pruned from the network without affecting its performance. They found that the fraction of prunable weights increases with network width for a ResNet trained on CIFAR-10. Li et al. [2018] uses the idea of subspace training to estimate the number of parameters (or intrinsic dimension) needed to achieve good performance using neural networks. They found that many problems have smaller intrinsic dimensions than one might suspect, and the intrinsic dimension for a given dataset varies little across a family of models with vastly different sizes.

While these conclusions align with our findings, their results rely on the usage of random subspace solutions with a performance $\approx 90\%$ of the baseline to define the intrinsic dimension. Our approach instead uses sparsification to find the effective number of parameters using models that generalise better than the baseline.

Venkatesh et al. [2020] used different calibration strategies on overparameterised models to study the impact on the resulting lottery tickets. Lei et al. [2023] propose a new sparse training method that improves the reliability of pruned models. From a theoretical perspective, Sakamoto and Sato [2022] used PAC-Bayesian theory to understand the generalisation behaviour of pruned networks. Zhang et al. [2021] characterise the performance of training a pruned neural network to show that pruning enlarges the convex region near a desirable model with guaranteed generalisation. Yang et al. [2023] used a controlled setting under random pruning to determine pruning fractions that can improve generalisation performance. A series of theoretical works [Malach et al., 2020, Orseau et al., 2020, Pensia et al., 2020, da Cunha et al., 2022, Burkholz, 2022] have shown the existence of a winning ticket inside larger (deeper and wider) networks of different sizes, thus providing some intuition on the amount of overparameterisation required.

To understand how IMP can improve the generalisation of neural networks by acting as a regulariser, Jin et al. [2022] studied the loss of influential samples in the optimally pruned models. Paul et al. [2022] use the geometry of the error landscape at each level of pruning to understand the principles behind the success of IMP (with rewinding) without label noise. Our focus is primarily on the properties (accuracy, number of parameters, and calibration) of the best pruned models. Ankner et al. [2022] used the framework of Pope et al. [2021], which uses GANs to generate images with known intrinsic dimensions, to find that the intrinsic dimensionality of data correlates with the prunability of neural networks. Certain efforts have also been made to understand the masks learned using pruning [Paganini and Forde, 2020, Pellegrini and Biroli, 2021].

Since the rediscovery of the double descent behaviour in deep neural networks [Belkin et al., 2019, Nakkiran et al., 2021], characterising double descent curves from simple models to deep networks has become a very active area of research, see Hastie et al. [2022], Spigler et al. [2019], Advani et al. [2020], Belkin et al. [2019], d'Ascoli et al. [2020a], Lin and Dobriban [2021], d'Ascoli et al. [2020b], Mei and Montanari [2022] for a small sample. Here, our focus is not on this double descent phenomenon – instead, our goal is to provide a univocal procedure to associate a pruned model with a large model using the sparse double descent.

3 EXPERIMENTAL SETUP

Datasets: We used MNIST [LeCun et al., 1998] and CIFAR-10 [Krizhevsky et al., 2009] for our experiments with real data. Additional experiments were also performed using the Fashion-MNIST dataset [Xiao et al., 2017] (see fig. S3 for results). We also perform experiments in a controlled setting, where we consider binary mixture classification with inputs sampled from a Gaussian mixture in $D = 100$ dimensions. This simple setting for the data model allows us to precisely compute the true conditional class probability of data and characterise the actual decision boundary. This allows us to closely investigate the architectural bias induced by pruning. We consider two settings for mixture classification. The *linear* dataset consists of two clusters that can be separated using a linear classifier. The two clusters have means $\mu_1 \neq \mu_2$, but same covariance matrix $\Sigma_1 = \Sigma_2$. The *XOR* dataset was created such that the resulting Gaussians (that have the same covariance) are placed like the graphic representation of the XOR logical function. The training and test sets contain $10\,000$ and 5000 samples respectively. See appendix A for more details.

A crucial aspect of our experiments is adding symmetric label noise by randomly permuting the labels for a fraction of the training data. Adding label noise to training data provides a straightforward way to produce double descent [Nakkiran et al., 2021] and sparse double descent [He et al., 2022] in deep neural networks. Note that using other kinds of label noise could provide different conclusions, but is not the focus of the current study and hence is out of scope for this paper. While adding label noise seems unrealistic, Northcutt et al. [2021] found that labelling errors are pervasive in several benchmark machine learning datasets and lower capacity models may be more practical. Building on this observation, we show that pruning can potentially provide a way of finding this capacity.

Models: A two-layer fully-connected network (FCN) was used for our experiments on the Gaussian mixture classification tasks with five replicates for each model (see details in appendix C). The width of the hidden layer was varied to obtain models of varying sizes and produce the double descent curve. For MNIST, we used two and three-layer FCNs while varying the width of the first hidden layer, and ResNet-6 with convolutional filters of different widths. For CIFAR-10, we used ResNet-18 [He et al., 2016] and varied the width of the convolutional filter to obtain the double descent curve. Each experiment was replicated three times for real-world datasets. The hyperparameters used in these experiments are described in appendix C. The cross-entropy loss function was used to train all the models. For CIFAR-10, we also successively removed one class at a time and then trained and pruned a ResNet-18 model with a fixed width of the convolutional filter. All our models were trained long enough to ensure that they were not in the epoch-wise double descent regime [Nakkiran et al., 2021]. The `OpenLTH` library[1] was used for our experimental evaluations. The code for replicating our experiments is available on GitHub[2].

Network pruning: IMP with 20% weights removal at each iteration (lottery ticket rewinding was used when required, see appendix B) was used for our experiments as it provides an effective procedure to find subnetworks with nontrivial sparsities that have low test error [Frankle and Carbin, 2019, Frankle et al., 2020, Blalock et al., 2020, Renda et al., 2020]. It should be noted that other pruning techniques like random pruning and gradient-based pruning also produce the sparse double descent curve [He et al., 2022] and can be used instead.

Computational costs: Our analysis requires us to work in the double descent paradigm, which allows us to properly define a best pruned model. Unfortunately, this implies significant computational costs as we have to perform the full sparse double descent analysis. For example, figs. 1, 2 and S2 required us to train approximately 300 different models and close to 1000 models when including replicates. Nevertheless, we believe that our extensive investigations using the mixture classification task and different architectures for MNIST and CIFAR-10 lay a solid foundation for the phenomenon described in the paper.

Terminology: The results in fig. 1 show that repeated pruning of different-sized models produces sparse double descent curves that achieve the lowest test error within a small range of parameters. To facilitate further discussion on this phenomenon, we define *best pruned model* and *effective number of parameters*.

Definition 1 (Best pruned model) *Let $E(\cdot)$ denote the test error (or 0-1 loss) of a model. For a trained model f_w, test dataset $\mathbb{S}_{\text{test}}$, and pruning method ϕ, the best pruned model f_w^{bp} is defined as*

$$f_w^{\text{bp}} = \underset{w}{\arg\min}\ E(\phi(f_w, \mathbb{S}_{\text{test}})).$$

These best pruned models have low generalisation error but unlike their overparameterized parent models, they have a non-zero error on the training data.

Definition 2 (Effective number of parameters) *The number of non-zero weights in the best pruned models f_w^{bp} obtained by pruning an overparameterised model.*

By pruning overparameterised models of different sizes we can obtain best pruned models that have similar accuracy but a slightly different effective number of parameters.

4 RESULTS

We first confirm that fully-connected neural networks exhibit both traditional double descent and sparse double descent on the mixture classification tasks: fig. 2 shows the same trend in the effective number of parameters as we observed in fig. 1. As expected, the test error decreases with the distance between cluster means ν (see fig. S1).

Next, we considered the effective number of parameters in the best pruned models (cf. definition 1). Focusing on the linear dataset in fig. 2a, we see that the pruned models (thin red lines) generally achieve the lowest test error with roughly 500 parameters. This number of parameters is optimal for starting networks of different size, which we use to determine the effective number of parameters for that dataset (see definitions 1 and 2). For networks trained on the linear dataset, we find that the effective number of parameters, i.e. the number of parameters in the best pruned models, ranges from ~ 200 to ~ 500 for different starting models and for different values of ν (see appendix D for the full distribution). We find the same behaviour for the XOR dataset, fig. 2b, but with a higher number of parameters (between 250 and 1000) than for the linear dataset with the same ν.

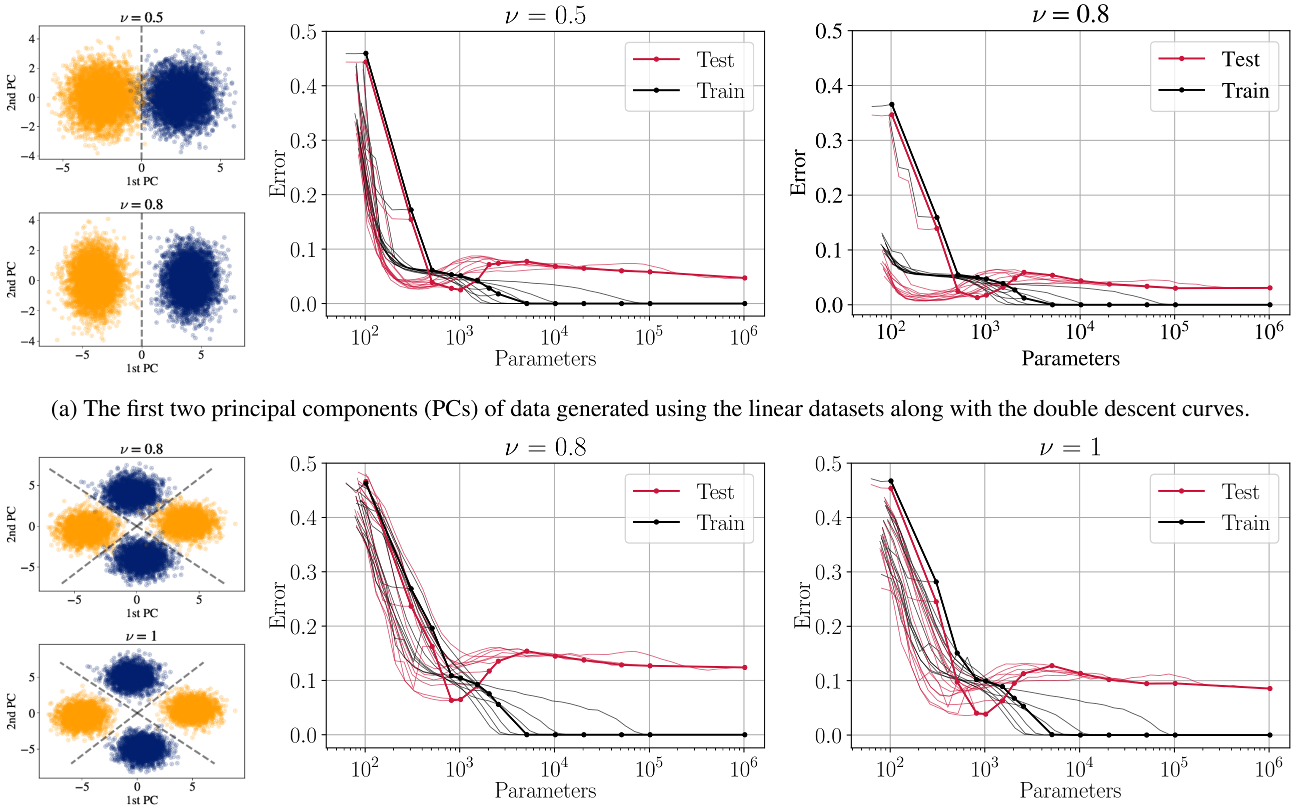
The double descent and sparse double descent curves for MNIST and CIFAR-10 datasets can be seen in figs. 1 and S2. Tables 1 and S2 show the number of parameters and test errors for the full/unpruned models and the corresponding best pruned models. Notice, importantly, that the best pruned models do not interpolate, so they exhibit a significant improvement in generalisation gap. Focusing on the number of parameters, we observe that a 200× increase for the full models results in only a $\sim 3.5\times$ increase in the number of parameters for the best pruned models. Comparing the test errors, a surprising finding is that sparse subnetworks with low test error exist inside the unpruned models that lie at the interpolation regime of the double descent curve. This means that even though the unpruned models have high test errors, they can be pruned using IMP to achieve much lower test errors. Pruning thus acts as a type of regularisation in this case, which is known to mitigate the peak of the test accuracy at the interpolation threshold [Belkin et al., 2019].

Our findings for CIFAR-10 similarly show that the best pruned models have better generalisation than the unpruned counterparts (see table 1). Similar to MNIST, we observe that the number of parameters in the best pruned models increases with the size of the original model but at a much slower rate. Comparing the effective number of parameters for MNIST and CIFAR-10, one can easily conclude that more parameters are required for CIFAR-10.

The effective number of parameters correlates with task difficulty: As illustrated in the data plots of fig. 2, optimally separating the linear dataset requires a plane, whereas

[1] `https://github.com/facebookresearch/open_lth`

[2] `https://github.com/viplovearora/noisy_lottery_tickets`



(a) The first two principal components (PCs) of data generated using the linear datasets along with the double descent curves.

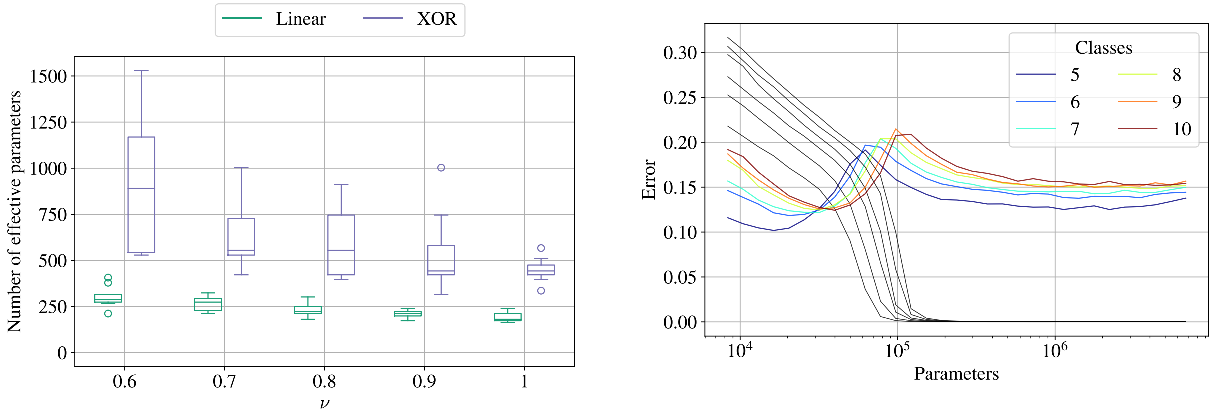
(b) The first two principal components (PCs) of data generated using the XOR datasets along with the double descent curves.

Figure 2: Average over 5 replicates of train and test error for models with different sizes, after numerous iterations of pruning, on the (a) linear, and (b) XOR datasets as the distance between clusters ν is varied. The red bold line with dots shows the traditional double descent curve, while the thinner lines represent the test error reached after pruning iterations of the initial models (sparse double descent). Similarly, black lines show train error both for the full and pruned models.

Table 1: Number of parameters and test error for unpruned (full) and best pruned models for MNIST and CIFAR-10. Average values over 3 replicates are reported. We observe that a 200× increase for the full models results in only a $\sim 3.5\times$ increase in the number of parameters for the best pruned models. Notice also that the error achieved by pruned models appears insensitive to the error rate of the original full model, i.e. even models with poor generalisation can be rescued by pruning.

MNIST (3 layer FC)				CIFAR-10 (ResNet-18)			
Parameters		Test error		Parameters		Test error	
Full	Pruned	Full	Pruned	Full	Pruned	Full	Pruned
45 200	6061	0.105	0.048	128 271	17 211	0.143	0.141
89 400	4908	0.176	0.059	238 155	16 359	0.159	0.143
266 200	5987	0.163	0.071	238 155	16 359	0.159	0.143
443 000	6377	0.144	0.064	1 689 080	19 647	0.187	0.130
885 000	10 196	0.123	0.084	3 798 420	35 028	0.165	0.128
4 421 000	10 683	0.099	0.089	10 546 700	49 799	0.146	0.128
8 841 000	17 094	0.091	0.097	23 725 050	57 356	0.137	0.129

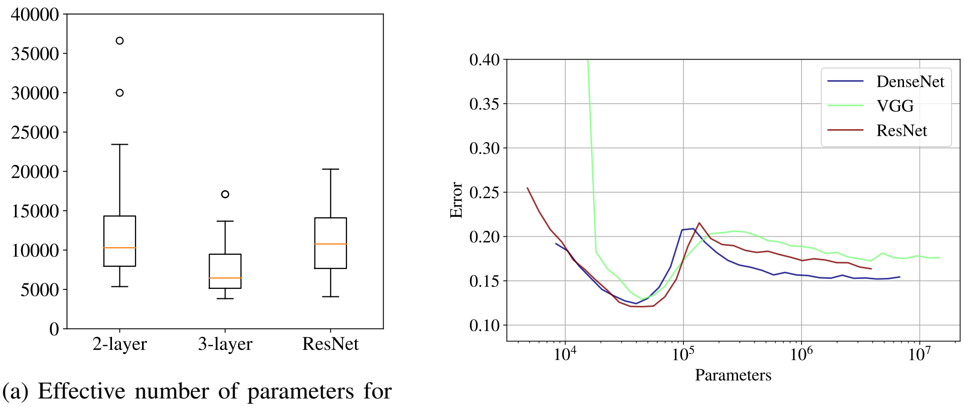
two planes are needed for separating the classes in the XOR dataset. Thus, one would expect that the effective number of parameters would double going from the linear to XOR datasets. Interestingly, this is what we observe in fig. 3a, even though the best pruned models are obtained from full models of the same size. The pruning approach thus suggests a way to quantify the effective number of parameters in deep, over-parameterised neural networks. Can we extend this approach to measuring task difficulty to more realistic datasets and convolutional networks? The results in fig. 3b and table S3 show that for a fixed ResNet model, as we decrease the number of classes in the CIFAR-10 dataset (or



(a) Comparing the effective number of parameters in the two mixture classification datasets for different values of ν.

(b) Sparse double descent curves on subsets of CIFAR-10 dataset show that effective number of parameters is related to task difficulty.

Figure 3: Our analysis shows that the effective number of parameters correlates with task difficulty: (a) the effective number of parameters required for the XOR dataset is approximately twice that of the linear dataset, which reflects the higher complexity of the XOR task, and (b) decreasing the number of classes in the CIFAR-10 dataset allows IMP to find models with fewer effective number of parameters.



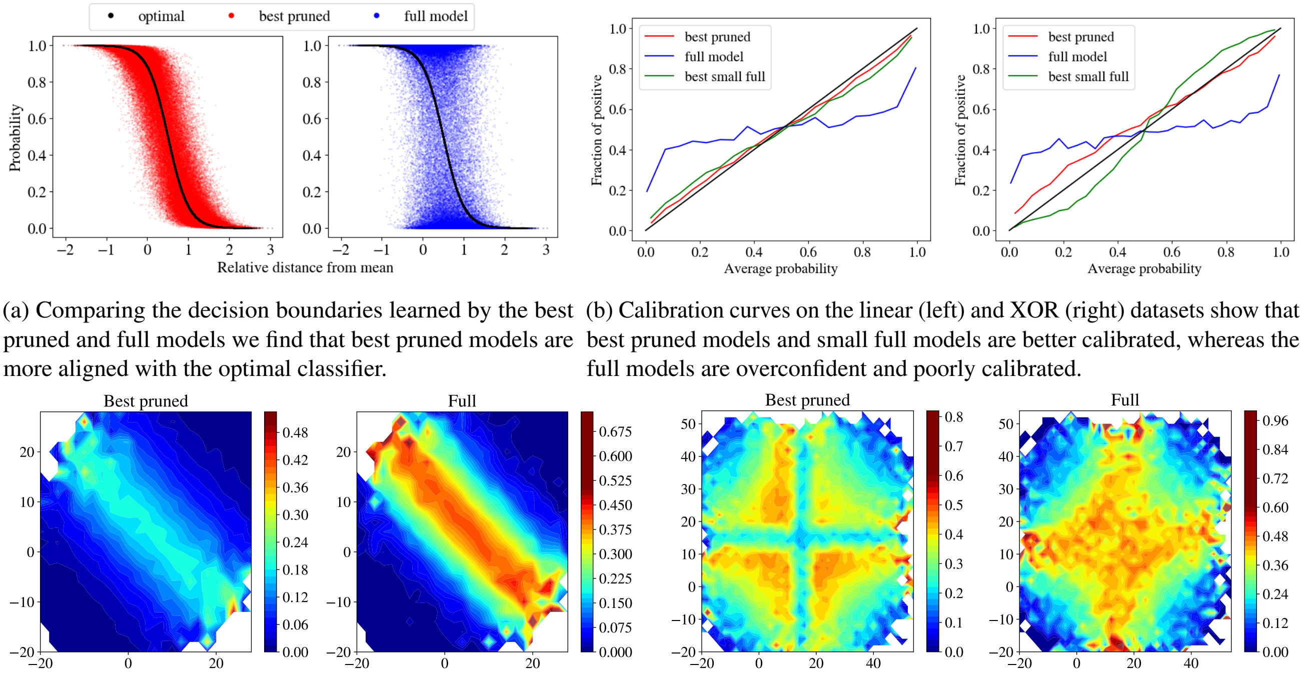
(a) Effective number of parameters for different models trained on the MNIST dataset.

(b) Sparse double descent curves obtained on pruning different convolutional neural networks for CIFAR-10 dataset.

Figure 4: Comparing the effective number of parameters across different neural network architecture for MNIST and CIFAR-10 datasets.

equivalently make the task easier for an overparameterised model), the effective number of parameters reduces for a smaller classification task. This further affirms our observation on the mixture classification task that the effective number of parameters is related to the difficulty of the task. Interestingly, the effective number of parameters does not decrease linearly with the number of classes.

The effective number of parameters is comparable across architectures: How does the choice of neural network architecture impact the effective number of parameters for a given dataset? We compare the size of best pruned models obtained for the MNIST dataset using three different neural network architectures: two-layer FCN, three-layer FCN, and ResNet-6, in fig. 4a. We find that although the test error of the best pruned models for the different architectures is different (see tables 1 and S2), the effective number of parameters are fairly comparable. This provides additional empirical evidence that our procedure might be invariant to the neural network architecture and reflects a notion of the 'difficulty' of the classification task, something that we already noted in the mixture classification task. This observation also carries forward to different convolutional neural network architectures on CIFAR-10. Figure 4b shows the sparse double descent curves obtained for DenseNet, VGG, and ResNet. The effective number of parameters is similar across the three different architectures: 44 590 for DenseNet, 44 468 for VGG, and 39 483 for ResNet.



(a) Comparing the decision boundaries learned by the best pruned and full models we find that best pruned models are more aligned with the optimal classifier.

(b) Calibration curves on the linear (left) and XOR (right) datasets show that best pruned models and small full models are better calibrated, whereas the full models are overconfident and poorly calibrated.

(c) Visualising the absolute difference between the predicted class probabilities and the optimal probabilities for the linear (left) and XOR (right) datasets projected on a 2D plane. Lower values (darker blue colours) are preferred.

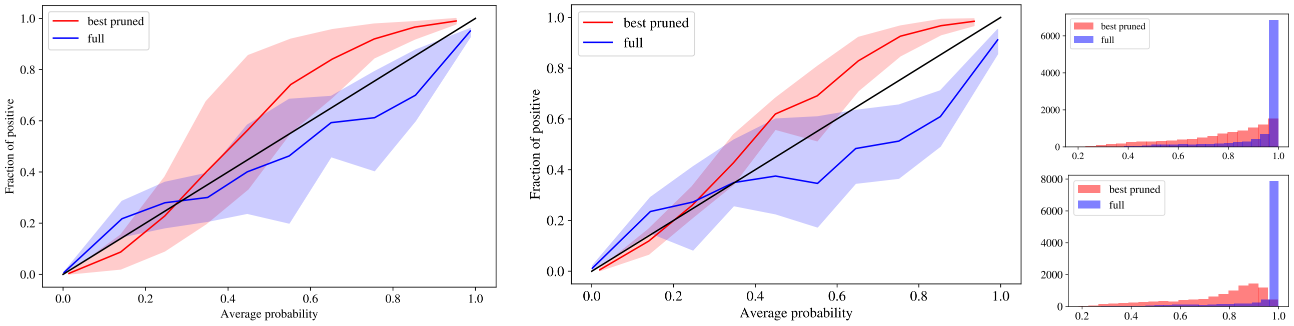
Figure 5: Analysis of prediction uncertainty in the mixture classification task. We find that the best pruned models are better at capturing uncertainty, whereas the overparameterised models tend to be overconfident in their predictions.

4.1 PRUNED MODELS BETTER CAPTURE UNCERTAINTY

We first use the mixture classification task to understand the architectural bias induced by pruning. For the results presented in fig. 5, we evaluate the models with 500 neurons in the hidden layer for the linear ($\nu = 0.2$) and XOR ($\nu = 0.6$) datasets. Similar results are obtained for different starting models and other values of ν. To understand what differentiates the pruned models we analyse the decision boundary learned by these models. We sample $100\,000$ points from the original data distribution and compute the corresponding probabilities for the models of interest along with the conditional class probabilities. In fig. 5a, we plot the probability of data belonging to class $y = 0$ as a function of the relative distance between the centres of the two clusters (located at $x = 0$ and $x = 1$) for the linear dataset. As expected, the optimal probabilities are given by a logistic function shown using the black dots. Comparing the best pruned and full models, we can conclude that the function learnt by the best pruned model is much closer to the true class conditional probability used to generate the data, whereas the full overparameterised model is overconfident in predicting either class.

In fig. 5c, we visualise the absolute difference between the predicted probabilities and the conditional class probabilities relative to the position of each point projected on a 2D plane. A value close to 0 on the colour scale in fig. 5c implies that the model has correctly estimated the conditional probability. One would expect that a model would confidently predict the correct class farther from the boundary and make errors near the boundary. This intuition is clearly illustrated using the results for the linear dataset. Strikingly, the full model makes confident predictions near the boundary, thus resulting in a value of ≈ 0.45 shown using the red colour in fig. 5c. In the XOR dataset, the best pruned model is more uncertain in its predictions, especially near the boundary, but is better than the full model. This is further illustrated using the calibration curves in fig. 5b where we see a near-perfect calibration in the best pruned models for both datasets, whereas the full models are overconfident and poorly calibrated. We also observe that the calibration of the best small full models is comparable to the best pruned models.

The calibration curves for MNIST and CIFAR-10 in fig. 6 show that the curves for the best pruned models are usually above the black line, which means that they make the correct prediction while being uncertain. Figure 6c illustrates this using the probabilities corresponding to the predicted class. This implies that even though the pruned models are accurate, they tend to be underconfident in their predictions while the full models are overconfident. This is at odds with our observation in the mixture classification task. On further analysis, we found that the best pruned models are



(a) MNIST with two-layer FCN. ECE for pruned and full models are 0.114 and 0.052, respectively.

(b) CIFAR-10 with ResNet-18. ECE for pruned and full models are 0.122 and 0.101, respectively.

(c) Probabilities of the predicted class.

Figure 6: Class-averaged calibration curves for the best pruned and full models on (a) MNIST and (b) CIFAR-10 datasets show that the pruned models are underconfident while the full models are overconfident. The highlighted areas signify deviation between classes. (c) Probabilities of the predicted class for the MNIST (top) and CIFAR-10 (bottom) datasets.

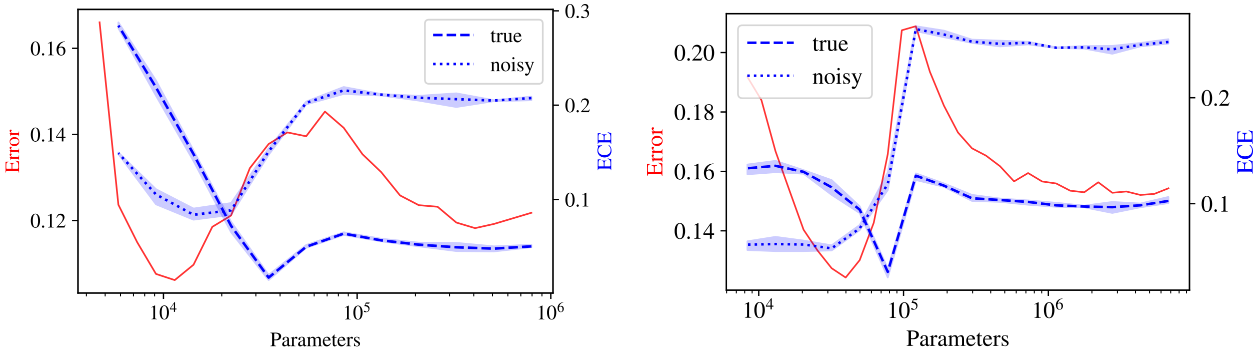


Figure 7: Sparse double descent (red) and expected calibration error (blue) on pruning two-layer FCN on MNIST (left) and ResNet-18 for CIFAR-10 (right). Highlighted area shows the deviation across three replicates. Comparing the calibration error for true and noisy labels we find that the best pruned models are optimally calibrated to noisy data.

calibrated to test data with label noise (see fig. S4). This is a reasonable outcome since the model was trained with label noise, which could introduce significant bias in the classification learned by the model.

To understand how iterative pruning impacts calibration, in fig. 7 we plot the test error and expected calibration error [Guo et al., 2017] (ECE) for a fixed model. For both noisy and true labels, we observe that the ECE initially increases slightly on iterative pruning but then drops significantly. The major difference between the two scenarios is the location of the minimum. For true labels, the lowest ECE is observed for models that have high test error, whereas the best pruned models are better calibrated to data with noisy labels. This strange behaviour requires further investigation to understand the impacts of pruning on model calibration. Perhaps, a trade-off between test error and calibration error can be used to choose models that have low error and are well-calibrated.

5 CONCLUSIONS

The existence and identifiability of small pruned networks, which can still generalise as well as large, over-parameterised models, remains an empirically well-supported fact that still lacks a satisfactory theoretical explanation. In this paper, we provided a number of empirical observations which unveiled two intriguing characteristics of small pruned networks: (1) their number of parameters correlate with task difficulty and provide a measure of the effective number of parameters in large models, and (2) pruned models are less prone to overconfident predictions.

Working in the IMP framework, we consistently reproduce the sparse double descent phenomenon first observed by He et al. [2022] in a number of synthetic and real datasets. While He et al. [2022] focused on the sparse double descent of the large model, we observe that irrespective of the size (and test accuracy) of the model from which we start the sparse double descent, the resulting optimal pruned models all achieve a similar generalisation error and have comparable sizes. The size of the best pruned models (defined by the number of retained parameters) appears to correlate well with natural notions of task complexity, both in real and simulated datasets.

We then analysed more in-depth the functions learned by these small networks: on simulated data, where we know the actual class conditional probability functions, we observe that pruned models capture much better the true underlying function, whereas full models tend to be overconfident. On real data, we again verify that full models are more confident than pruned models, which in this case tend to be slightly under-confident. A possible explanation is that, as customary in double descent studies, our models are trained on data with label noise, which impacts their calibration on data with true labels. Indeed, we find that pruned models are nearly optimally calibrated to data with label noise, while full models are once again over-confident.

Our results are somewhat at odds with a recent claim by Yang et al. [2023], which showed evidence of over-confidence of pruned models in one example (ResNet-50 on CIFAR-100). A possible reason for this discrepancy is that the pruning algorithm employed in [Yang et al., 2023] is greedy and does not require re-training from initialisation. Additionally, their setup did not involve training on noisy labels; both of these observations might explain the different conclusions reached by that study. Indeed, these considerations point to a non-trivial interaction between pruning algorithm, training procedure and statistical characteristics of the resulting pruned models. Exploring such questions, both theoretically and empirically, might finally shed light on the unreasonable success of pruning large neural networks.

In the future, it would be interesting to see how other pruning procedures impact our findings and extend our study to larger datasets like ImageNet. It would be interesting to see if these findings apply to other architectures like auto-encoders and recurrent neural networks. Finally, our findings highlight the importance of the sparse double descent thus encouraging theoretical work to understand the phenomenon.

Author Contributions

Viplove Arora, Sebastian Goldt, and Guido Sanguinetti conceived the idea and wrote the paper. Daniele Irto performed the experiments for the mixture classification task. Viplove Arora performed the rest of the experimental analysis.

Acknowledgements

We acknowledge co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR "Future Artificial Intelligence Research". This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. Viplove Arora acknowledges partial support from Assicurazioni Generali through a liberal grant.

References

Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428 – 446, 2020.

Zachary Ankner, Alex Renda, Gintare Karolina Dziugaite, Jonathan Frankle, and Tian Jin. The effect of data dimensionality on neural network prunability. *arXiv preprint arXiv:2212.00291*, 2022.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

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

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

Leo Breiman. Reflections after refereeing papers for nips. *The Mathematics of Generalization*, pages 11–15, 1995.

Rebekka Burkholz. Most activation functions can win the lottery without excessive depth. In *Thirty-sixth Conference on Neural Information Processing Systems*, 2022.

Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent : Bias and variance(s) in the lazy regime. In *ICML*, 2020a.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where &amp; why do they appear? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3058–3069. Curran Associates, Inc., 2020b.

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

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Zheng He, Zeke Xie, Quanzhi Zhu, and Zengchang Qin. Sparse double descent: Where network pruning aggravates overfitting. In *International Conference on Machine Learning*, pages 8635–8659. PMLR, 2022.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.*, 22(241):1–124, 2021.

Tian Jin, Michael Carbin, Daniel M. Roy, Jonathan Frankle, and Gintare Karolina Dziugaite. Pruning's effect on generalization through the lens of training and regularization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Lorenz Kuhn, Clare Lyle, Aidan N Gomez, Jonas Rothfuss, and Yarin Gal. Robustness to pruning predicts generalization in deep neural networks. *arXiv preprint arXiv:2103.06002*, 2021.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Bowen Lei, Ruqi Zhang, Dongkuan Xu, and Bani Mallick. Calibrating the rigged lottery: Making all tickets reliable. In *International Conference on Learning Representations*, 2023.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *The Journal of Machine Learning Research*, 22(1):6925–7006, 2021.

Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117 (20):10625–10626, 2020.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021 (12):124003, 2021.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

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

Michela Paganini and Jessica Forde. On iterative neural network pruning, reinitialization, and the similarity of masks. *arXiv preprint arXiv:2001.05050*, 2020.

Mansheej Paul, Feng Chen, Brett W Larsen, Jonathan Frankle, Surya Ganguli, and Gintare Karolina Dziugaite. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? *arXiv preprint arXiv:2210.03044*, 2022.

Franco Pellegrini and Giulio Biroli. Sifting out the features by pruning: Are convolutional networks the winning lottery ticket of fully connected ones? *arXiv preprint arXiv:2104.13343*, 2021.

Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic overparameterization is sufficient. *Advances in neural information processing systems*, 33:2599–2610, 2020.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

RJ Poppi and DL Massart. The optimal brain surgeon for pruning neural network architecture applied to multivariate calibration. *Analytica chimica acta*, 375(1-2):187–195, 1998.

Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020.

Keitaro Sakamoto and Issei Sato. Analyzing lottery ticket hypothesis from PAC-bayesian theory perspective. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52 (47):474001, 2019.

F Vallet, J-G Cailton, and Ph Refregier. Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *EPL (Europhysics Letters)*, 9(4):315, 1989.

Bindya Venkatesh, Jayaraman J Thiagarajan, Kowshik Thopalli, and Prasanna Sattigeri. Calibrate and prune: Improving reliability of lottery tickets through prediction calibration. *arXiv preprint arXiv:2002.03875*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Hongru Yang, Yingbin Liang, Xiaojie Guo, Lingfei Wu, and Zhangyang Wang. Theoretical characterization of how neural network pruning affects its generalization. *arXiv preprint arXiv:2301.00335*, 2023.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks. *Advances in Neural Information Processing Systems*, 34: 2707–2720, 2021.