Predicting generalization with degrees of freedom in neural networks

Erin Grant [*1] **Yan Wu** [*2]

Abstract

Model complexity is fundamentally tied to predictive power in the sciences as well as in applications. However, there is a divergence between naive measures of complexity such as parameter count and the generalization performance of over-parameterized machine learning models. Prior empirical approaches to capturing intrinsic complexity in a more sophisticated manner than parameter count are computationally intractable, do not capture the implicitly regularizing effects of the entire machine-learning pipeline, or do not provide a quantitative fit to the double descent behavior of overparameterized models. In this work, we introduce an empirical complexity measure inspired by the classical notion of generalized degrees of freedom in statistics. This measure can be approximated efficiently and is a function of the entire machine learning training pipeline. We demonstrate that this measure correlates with generalization performance in the double-descent regime.

1. Introduction

What makes a machine learning model a good scientific explanation? Classical perspectives call for trading off low *complexity* and high *predictive accuracy* (Forster & Sober, 1994). However, deep learning approaches to scientific modeling (Baraniuk et al., 2020; Bianchini et al., 2020; Raghu & Schmidt, 2020; Hope et al., 2022) come with new challenges in navigating this trade-off due to the difficulty in assessing *both* complexity and predictive accuracy. First, many established measures of complexity do not apply to deep learning models because they are vacuous (Dziugaite & Roy, 2017), or they do not capture the non-monotonic behavior of the evaluation error to which complexity is fundamentally tied (Belkin et al., 2019; Nakkiran et al., 2021). Secondly, over-parameterized deep learning models often obtain zero training error (Zhang et al., 2017), meaning that model selection by predictive accuracy on the training set alone is not possible, while maintaining a test set representative of downstream applications is challenging in practice (Nagarajan et al., 2021). These challenges mean that practitioners lack foundational methods for model selection, hindering the reliable deployment of many machine learning techniques (Amodei et al., 2016; D'Amour et al., 2020; Geirhos et al., 2020) and giving rise to poorly understood pathologies such as susceptibility to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015).

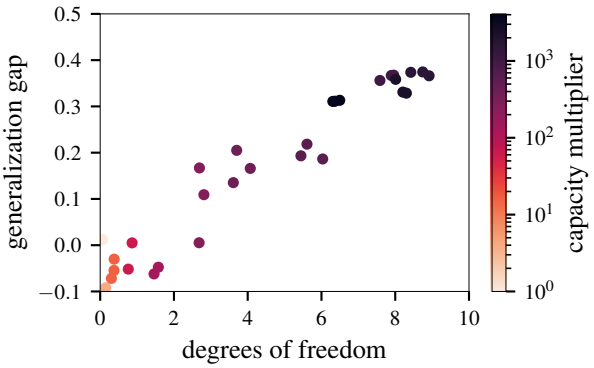

Figure 1. The generalization gap among neural network models that vary by a capacity multiplier (here, kernel parameter multiplier for convolutional neural networks) is monotonically increasing in the complexity measure *generalized degrees of freedom* (Pearson's $r(36) = .88$, $p = 9.9 \times 10^{-11}$).

In this work, we explore the extent to which the *generalized degrees of freedom (GDoF)* framework introduced by Efron (1983) and Ye (1998) can respond to these challenges. GDoF, as its name suggests, generalizes the basic notion of degrees of freedom for non-linear models, and is equivalent to the number of features for linear models. This correspondence provides a natural measurement of the complexity of machine learning models, including deep neural networks. Given the fundamental connection between complexity and generalization, we hypothesize that the GDoF predicts a

[*]Equal contribution [1]Department of EECS, UC Berkeley, Berkeley, USA; work done during an internship at DeepMind. [2]DeepMind, London, UK. Correspondence to: Erin Grant <eringrant@berkeley.edu>, Yan Wu <yanwu@deepmind.com>.

Proceedings of the 2^{nd} Workshop on "AI for Science" (AI4Science) at the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, 2022.

model's generalization performance. GDoF has been explored in various contexts, though Gao & Jojic (2016) was the first to study this measurement for deep neural networks. In this work, we make the following additional contributions:

- We introduce approximations that enable scalable computation of the GDoF during training of deep neural networks (Section (**3**)).
- We reconcile the double-descent phenomenon with the classical trade-off between complexity and predictive accuracy by demonstrating the monotonicity of the generalization gap in GDoF (Fig. (**1**)).
- We use the online nature of the approximation to reveal the evolution of GDoF over training (Fig. (**3**) and Fig. (**5**)).
- We demonstrate that controlling GDoF via subspace training (Li et al., 2018) can improve generalization (Fig. (**4**) and Fig. (**5**)).

2. Background

When reasoning about model complexity, we contrast changes in the *parameter dimension* of a machine learning model—measures proportional to the parameter count of a model—with changes in the *effective dimensionality*, which is indirectly controlled by various components such as the architecture and training procedure. While trivially equivalent re-parametrizations—for instance, replacing a connection weight by the product of two weights, $w = w_1\, w_2$—may result in vastly different parameter dimensions, effective dimensionality captures the intrinsic complexity as a result of the complete system specification. According to classic notions of complexity-fit trade-off, models with low effective dimensionality have low complexity, are less likely to overfit and thus have lower generalization error (Schmidhuber, 1997). The preference for simplicity in model selection is also known as *Occam's Razor* (MacKay, 1991).

However, the classical trade-off between complexity and predictive power, which focuses on the bias and variance trade-off (Geman et al., 1992), has been challenged by the recently re-discovered double descent phenomenon (Belkin et al., 2019): While large models are traditionally understood to be inherently more complex and thus subject to poor generalization as a result of overfitting, modern over-parameterized machine learning models exhibit surprisingly good generalization performance (Zhang et al., 2017). While the origin of the discrepancy between parameter dimension and effective dimensionality is outside of the scope of this work, previous work has shown that it may arise from implicit regularization due to various components of the machine-learning pipeline, including the model architecture (Golubeva et al., 2020), the specific initialization procedure (Kubo et al., 2019; Mehta et al., 2021), gradient-based optimization (Barrett & Dherin, 2021), and overparametrization itself (Neyshabur et al., 2019; Advani et al., 2020).

Various attempts have been made to capture effective dimensionality via model complexity from the perspective of a model's loss landscape (Larsen et al., 2021) and, in particular, its local curvature (Maddox et al., 2020; Loukas et al., 2021). The statistical concept of *degrees of freedom* can also be related to the loss landscape. For linear models, the degree of freedom is equivalent to parameter count, but Efron (1983) and Ye (1998) generalized this concept to non-linear models from the perspectives of *expected optimism* and *average sensitivity* as follows: For a classifier f parameterized by θ, the GDoF is defined from the derivatives of the *model-predicted label* $f_{n,c}$ with respect to the *data label* $y_{n,c}$:

$$\Phi \coloneqq \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\partial f_{n,c}(\theta)}{\partial y_{n,c}} \,, \tag{1}$$

where $n \in \{1, \ldots, N\}$ is the index of the sample in the training set, $c \in \{1, \ldots, C\}$ is the index of the class. When explicit dependence on the training procedure is required, we further expand Eq. (1) and specify the algorithm $\mathcal{A}$, the initial parameter θ_0, and the training step t such that the GDoF can reflect the model complexity at any step t before convergence as

$$\Phi(t) \coloneqq \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\partial f_{n,c}(\theta_t = \mathcal{A}(\theta_0, t))}{\partial y_{n,c}} \,. \tag{2}$$

Naively, computing these derivatives would require back-propagating through the whole training procedure specified by $\mathcal{A}$.

The GDoF measurement Φ in Eq. (2) can be interpreted as the sensitivity of the model f to perturbation in the training label $y_{n,c}$. Gao & Jojic (2016) introduced GDoF to neural networks by approximating the derivatives in Eq. (2) by a specific random perturbation and demonstrated that the GDoF of such over-parameterized models are only a small fraction of the total number of parameters. However, (Gao & Jojic, 2016) computed GDoF by training a perturbed model, as well as the original model, on the complete training dataset until convergence, thus requiring dual training runs for each iteration of computing the GDoF. Here we show that the approximation of Eq. (2) can be broken up in both space—with mini-batches—and in time—by incremental computation. In addition, the latter incremental approximation enables us to reveal the evolution of GDoF during training.

3. Method

In this work, we introduce additional approximations to GDoF, defined as Φ in Eq. (2) for a more scalable estimate

of model complexity. The first straightforward step of approximation is using a Monte-Carlo estimator based on a subset $B < N$ of samples from the training set:

$$\Phi(t) \approx \Phi^{\text{MC}}(t) \coloneqq \frac{N}{B} \sum_{n=1}^{B} \sum_{c=1}^{C} \frac{\partial f_{n,c}\left(\theta_t = \mathcal{A}(\theta_0, t)\right)}{\partial y_{n,c}} \,. \tag{3}$$

The second is to remove the requirement of training the perturbed model to convergence by exploiting the sequential updating of model parameters θ during gradient-based training. The following proposition shows that, given the GDoF of a model trained from initialization to time t, the GDoF at step $t' > t$ can be approximated by simply adding the additional sensitivity brought by the increment $\Delta\theta = \theta_{t'} - \theta_t$ as long as $\Delta\theta$ is small.

Proposition 1. *The GDoF at step t' is related to the GDoF at $t < t'$ via:*

$$\Phi(t') \approx \Phi(t) + \Delta\Phi(t, t') \,, \tag{4}$$

where

$$\Delta\Phi(t, t') = \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\partial f_{n,c}(\theta_{t'} = \mathcal{A}(\theta_t, t' - t))}{\partial y_{n,c}} \tag{5}$$

measures the increment of sensitivity after training f with $\mathcal{A}$ to step t' from θ_t at step t.

Proof. We first use the following shorthand to represent the instance- and class-wise contributions to Φ:

$$\phi_{n,c}(t, t') \coloneqq \frac{\partial f_{n,c}(\theta_{t'} = \mathcal{A}(\theta_t, t' - t))}{\partial y_{n,c}} \,. \tag{6}$$

With this, Eq. (2) can be written as

$$\Phi(t) = \sum_{n=1}^{N} \sum_{c=1}^{C} \phi_{n,c}(0, t) \,. \tag{7}$$

For additional brevity, we drop the example subscript i and the class subscript c when they are irrelevant to the derivation, giving

$$\phi(0, t') = \frac{\partial f(\theta_{t'})}{\partial y} \,. \tag{8}$$

We then construct a Taylor expansion of $f(\theta)$ near $\theta = \theta_t$:

$$f(\theta) = f(\theta_t) + (\theta - \theta_t)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_t} + \mathcal{O}(\|\theta - \theta_t\|^2) \,, \tag{9}$$

which gives, when $\Delta\theta = \theta_{t'} - \theta_t$ is small,

$$f(\theta_{t'}) \approx f(\theta_t) + (\Delta\theta)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_t} \,. \tag{10}$$

Plugging in the approximation Eq. (10) into Eq. (8) gives

$$\begin{aligned} \phi(0, t') &\approx \frac{\partial f(\theta_t)}{\partial y} + \left(\frac{\partial \Delta\theta}{\partial y}\right)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_t} \\ &\quad + (\Delta\theta)^T \left.\frac{\partial^2 f(\theta)}{\partial y\, \partial \theta}\right|_{\theta=\theta_t} \\ &\approx \phi(0, t) + \left(\frac{\partial \Delta\theta}{\partial y}\right)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_t} \,. \end{aligned} \tag{11}$$

Here the approximation comes from dropping the second order term, which is negligible when the update $\Delta\theta$ or the second derivative $\left.\frac{\partial^2 f(\theta)}{\partial y\, \partial \theta}\right|_{\theta=\theta_t}$ is small.

Moreover, since $\phi(t, t')$ depends on y only through the update $\Delta\theta$, we can employ the chain rule to give

$$\begin{aligned} \phi(t, t') &= \frac{\partial f(\theta_{t'} = \mathcal{A}(\theta_t, t' - t))}{\partial y} \\ &= \left(\frac{\partial \Delta\theta}{\partial y}\right)^T \left.\frac{\partial f(\theta_t + \Delta)}{\partial \Delta}\right|_{\Delta=\Delta\theta} \\ &= \left(\frac{\partial \Delta\theta}{\partial y}\right)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_{t'}} \qquad (12) \\ &\approx \left(\frac{\partial \Delta\theta}{\partial y}\right)^T \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta_t} \,, \qquad (13) \end{aligned}$$

where the last line of approximation comes from using the derivatives of $\frac{\partial f(\theta)}{\partial \theta}$ at θ_t for $\theta_{t'}$. In Eq. (12), we used the definition of the derivative, while changing the variable $\theta = \theta_t + \Delta$:

$$\begin{aligned} \frac{\partial f(\theta_t + \Delta)}{\partial \Delta} &= \lim_{\delta \to 0} \frac{f(\theta_t + \Delta + \delta) - f(\theta_t + \Delta)}{\delta} \\ &= \lim_{\delta \to 0} \frac{f(\theta + \delta) - f(\theta)}{\delta} \\ &= \frac{\partial f(\theta)}{\partial \theta} \,. \end{aligned} \tag{14}$$

Combining Eq. (13) with Eq. (11), we have

$$\phi(0, t') \approx \phi(0, t) + \phi(t, t') \,. \tag{15}$$

We can now obtain Proposition (1) from the definitions of ϕ and Φ. □

Although several steps of the approximation require $\|\Delta\theta\|^2$ to be small, in our experiments we found the resulting estimator of GDoF predictive of generalization even when $\Delta\theta$ represents a change of up to 1000 update steps. In addition, we found performing multiple update steps while re-sampling mini-batches (K in Algorithm (1)) improves the estimate of GDoF, as in our experiments in Section (4). We leave detailed analysis of the approximations, such as the error accumulated through the trajectory, for future work.

The next section empirically demonstrates the predictive power of the measurement Eq. (2) and the two approximations we make in Eq. (3) and Eq. (4) for neural networks. Algorithm (**1**) summarizes the algorithm for computing the GDoF from a series of checkpoints saved during training. Notice that taking $T = 0$, and $K = \infty$ (*i.e.,* training the original and perturbed model from initialization until convergence) recovers the algorithm of Gao & Jojic (2016).

Algorithm 1 Algorithm for computing GDoF

Require: saved checkpoints $\mathcal{C}_1, \mathcal{C}_2, \dots \mathcal{C}_T$, the number of update steps K, mini-batch size B, perturbation step ϵ

$\Phi \leftarrow 0$

for $t = 1, \dots T$ **do**

  Load model $f(\theta_t)$ from checkpoint $\mathcal{C}_t$

  for $k = 1, \dots K$ **do**

    Sample a mini-batch of data $\{x_b, y_b\}_{b=1}^{B}$

    Sample perturbation $\{z_b\}_{b=1}^{B} \sim \mathcal{N}(0, 1)$

    Update $f(\theta_{t+1})$ using $\{x_b, y_b\}_{b=1}^{B}$

    Update $f'(\theta_{t+1})$ using $\{x_b, y_b + \epsilon \cdot z_b\}_{b=1}^{B}$

  end for

  Approximate

$$\Delta\Phi(t, t+1) \leftarrow \frac{N}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} z_b(c) \frac{f'_{b,c}(\theta_{t+1}) - f_{b,c}(\theta_{t+1})}{\epsilon}$$

  Accumulate $\Phi(t+1) \leftarrow \Phi(t) + \Delta\Phi(t, t+1)$

end for

4. Experiments

We demonstrate results measuring the GDoF of a convolutional neural network (CNN) trained on the CIFAR-10 dataset. The CNNs follow the architecture from Nakkiran et al. (2021) with three convolutional layers. We vary their capacities via a *capacity multiplier*, the multiple of the number of filter channels. We thus have a batch of models whose first layer has filter sizes $\{1, 2, 4, 8, 12, 16, 20, 24, 32, 40, 48, 64\}$. All models are trained with a batch size of 128 for 3×10^5 training iterations. In Fig. (**2**) we plot the training error and the evaluation error alongside their difference—the generalization gap—and the approximated GDoF for models with different capacity multipliers. The double descent curve (Belkin et al., 2019) can be observed in the evaluation error curve, and the measured GDoF closely tracks the generalization gap.

To further investigate the correlation between the GDoF and the generalization gap, we represent the same data in a scatter plot in Fig. (**1**), which reveals that the generalization gap increased monotonically with the GDoF, as predicted by classical notions of program complexity (Schmidhuber, 1997). To quantify the correlation and verify our observation, we compute Spearman's rank correlation, giving $r(36) = .99$, $p = 2.5 \times 10^{-11}$, and Pearson's correlation coefficient, giving $r(36) = .88, p = 9.9 \times 10^{-11}$.

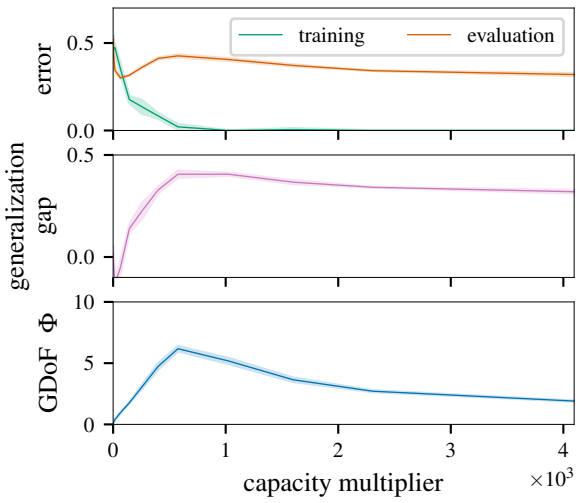

Figure 2. Training and evaluation error **(top)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs. capacity multiplier (here, kernel parameter multiplier) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The evaluation error exhibits the double-descent phenomenon, while the GDoF tracks the generalization gap.

For each capacity multiplier, we also expand the whole training curves into Fig. (**3**) to depict the evolution of the measured quantities over training. We again observe an increase and then decrease in the final generalization gap as the model size increases (curve color darkens), and this trend is reflected in the GDoF: An intermediate capacity multiplier has highest final generalization gap and highest GDoF. Further, the dynamics of the generalization gap and GDoF match over the course of training: For example, the full model (darkest curve) increases in generalization gap as well as GDoF most quickly, but plateaus at intermediate values.

We also explore how the GDoF reacts when model complexity is explicitly controlled. We measure GDoF of the same set of models as in Fig. (**2**), here trained with the subspace method of Li et al. (2018), which was originally proposed as an ingredient of another method to approximate the intrinsic dimensionality of a model. In particular, following Li et al. (2018), we randomly project model parameters to a random, lower-dimensional subspace $\theta_S = R\,\theta$, where R is a random matrix fixed at the start of training, and optimize the lower-dimensional vector θ_S instead of θ to train the model. Intuitively, the model complexity is constrained by the coupling introduced by the random projection.

We choose as the base model the CNN with the highest capacity (first layer filter size 64, ≈ 1.55 million parameters). We then train this base mode using subspace dimensions of

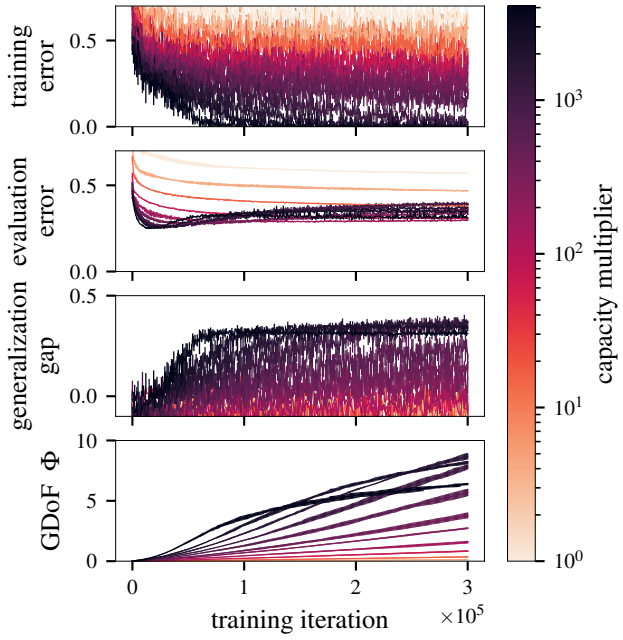

Figure 3. Training and evaluation error **(top two)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs. training iteration for different values of the capacity multiplier (here, kernel parameter multiplier) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The GDoF captures the evolution of the generalization gap over training for the various capacity settings.

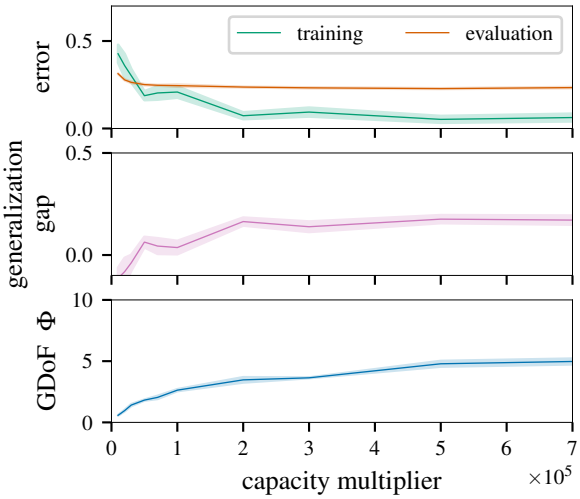

Figure 4. Training and evaluation error **(top)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs capacity multiplier (here, subspace dimension in the algorithm of Li et al. (2018)) for 3 CNNs trained with subspace optimization on the CIFAR-10 dataset for 300 thousand iterations. In contrast to Fig. (**2**), the evaluation error is monotonically decreasing. The GDoF still tracks the generalization gap.

$\{1, 2, 3, 5, 7, 10, 20, 30, 50, 70, 90\} \times 10^4$. Fig. (**4**) demonstrates that the GDoF still tracks the generalization gap when complexity is controlled with subspace optimization, and Fig. (**5**) and Fig. (**6**) further shows that a suitable space dimension can *improve* performance by reducing the generalization gap; that is, intermediate subspace dimensions can lower the final evaluation error below that of the full model.

5. Related Work

Whence generalization? Generalization properties of deep neural networks have received renewed interest from both theoretical and empirical perspectives (Jiang et al., 2021). However, it seems that many properties that correlate with generalization performance are neither necessary nor sufficient to ensure it; for example, sharpness of the loss surface around a minimum (Dinh et al., 2017), or monotonic linear interpolation (Lucas et al., 2021).

Jiang et al. (2021) discuss the results of a recent community competition that targeted complexity measures that reliably capture generalization performance; however, they do not investigate the trend of these complexity measures as the parameter dimension is manipulated. Maddox et al. (2020) resolve the double descent paradox with the introduction of a capacity estimate inspired by posterior contraction in Bayesian learning that renders evaluation performance monotonically increasing.

Capturing complexity. Proposals to measure effective dimensionality have taken many forms, from more classical parameter and representation norms (Bartlett et al., 2017) and sharpness measures (Keskar et al., 2017), to measures motivated specifically by modern deep learning systems, such as measures of gradient noise (Smith & Le, 2018) and optimization speed (Hardt et al., 2016); see Jiang et al. (2020) for a thorough overview and an empirical comparison. The common idea behind all such approaches is to quantify variability and thus complexity in the function represented by a neural network. This functional complexity may result from the interaction of the particular parameterization or architecture in addition to all other aspects of the machine learning pipeline, including the optimization algorithm and other hyper-parameters. Crucially, however, the exact relationship between neural network design decisions and functional complexity is unknown, prompting the study of *implicit regularization* in deep learning (Neyshabur et al., 2015; Neyshabur, 2017; Dherin et al., 2021).

Finding neural networks with minimal complexity. One class of methods directly searches models as programs with low Kolmogorov complexity. For example, Schmid-

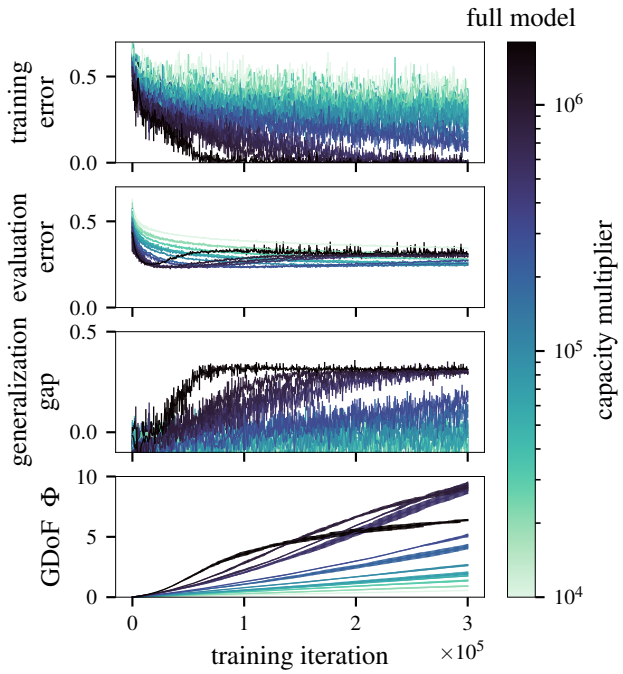

Figure 5. Training and evaluation error **(top two)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs. training iteration for different values of the capacity multiplier (here, subspace dimension in the algorithm of Li et al. (2018)) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The GDoF captures the evolution of the generalization gap over training for the various capacity settings.

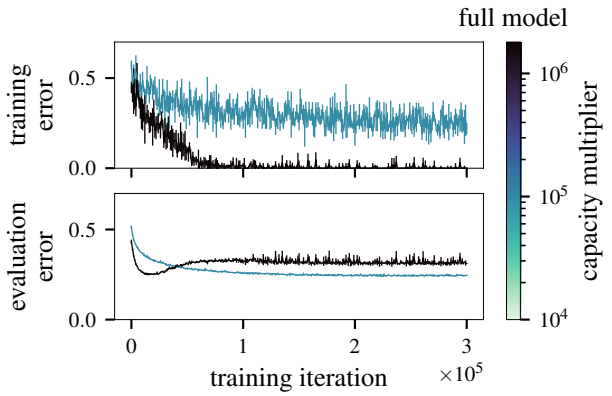

Figure 6. Partial reproduction of Fig. (5), for emphasis. Training **(top)** and validation error **(bottom)** vs. training iteration for the full model as compared to a subspace model of intermediate subspace dimension (here, 10^5). Controlling the complexity via subspace optimization in an intermediate subspace dimension results in a lower final validation error in the absence of early stopping.

huber (1997) proposes a probabilistic search algorithm to discover neural networks with low Kolmogorov complexity, although it has only been demonstrated with toy problems that can be solved with relatively short programs consisting of limited primitives. Another class of methods encourages optimizers to find simpler neural networks by penalizing deviation from a specified *prior* over the weights. The simplest example is "weight-decay" (Hinton, 1987), which can be interpreted as finding the *maximum a posteriori* (MAP) solution given an isotropic Gaussian prior over weights. Nowlan & Hinton (1992) elaborate this idea by introducing a mixture of Gaussians prior over the weights, but still assume that all weights are quantized with the same noise tolerance. Hinton & Van Camp (1993) remove this assumption by employing a variational method to approximate the posterior distribution of *noisy weights* with different precisions.

Bayesian model selection and Occam's Razor. MacKay (1992a) reviews how *model selection* or *comparison* can be achieved in a Bayesian context by comparing the evidence for each model. In particular, given a parameterized model family $\mathcal{H}$ that specifies a *prior* distribution over parameters $\mathbf{w}$, $P(\mathbf{w} \mid \mathcal{H})$ and a *predictive* distribution over data $\mathcal{D}$, $P(\mathcal{D} \mid \mathbf{w}, \mathcal{H})$, the evidence is $P(\mathcal{D} \mid \mathcal{H}) = \int P(\mathcal{D} \mid \mathbf{w}, \mathcal{H}) P(\mathbf{w} \mid \mathcal{H}) \, \mathrm{d}\mathbf{w}$. As a rule of thumb, greater entropy in the prior term $P(\mathbf{w} \mid \mathcal{H})$, corresponding to a more flexible hypothesis class, results in a lesser value for the evidence, a regularizing effect known as *Occam's Razor*. It should be noted that the behavior of Occam's Razor—in particular, what notion of flexibility is penalized—implicitly relies on the model parameterization and is therefore not an *a priori* effect (Wolpert, 1995); for instance, MacKay (1992a) discuss Occam's Razor and Bayesian model comparison for models that admit Gaussian approximation and quadratic regularizers, and MacKay (1992b) study such approximations in the context of neural networks. Nevertheless, Bayesian model selection allows coherent comparison of appropriately related prior distributions (Jefferys & Berger, 1992), and robustness with respect to the selection of the prior distribution can be ensured in some situations (Berger et al., 1994).

Implicit biases in model design. Lineages of machine learning models are often the result of selection on the basis of performance on standardized benchmarks (Dehghani et al., 2021). As a result, such lineages implicitly depend on the types of data to which they have been applied—in particular, such models are likely to have implicit biases that aid performance on such datasets. Analogously, prior selection in Bayesian analysis often implicitly relies on data (Gelman et al., 2017). From a Bayesian perspective, the impact of prior information can be quantified using sensitivity analysis (e.g., Müller 2012) or with measures of prior-data and prior-likelihood conflict (Bousquet, 2008; Nott et al., 2020; Reimherr et al., 2021).

6. Conclusion

In this work, we investigated a complexity measure inspired by the classical notion of degrees of freedom in statistics. the variant of this measure that we propose is scalable and is a function of the entire machine learning training pipeline. We demonstrated that this measure strongly correlates with generalization performance in the double-descent regime, and that, when model complexity is controlled via subspace training, the resulting improvement in the generalization gap is tracked by this measure.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. URL https://doi.org/10.1016/j.neunet.2020.08.022.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Baraniuk, R., Donoho, D., and Gavish, M. The science of deep learning. *Proceedings of the National Academy of Sciences*, 117(48):30029–30032, 2020. URL https://doi.org/10.1073/pnas.2020596117.

Barrett, D. and Dherin, B. Implicit gradient regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.11162.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1706.08498.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. URL https://arxiv.org/abs/1812.11118.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. URL https://doi.org/10.1007/BF02562676.

Bianchini, S., Müller, M., and Pelletier, P. Deep learning in science. *arXiv preprint arXiv:2009.01575*, 2020. URL https://arxiv.org/abs/2009.01575.

Bousquet, N. Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, 35 (9):1011–1029, 2008. URL https://doi.org/10.1080/02664760802192981.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. URL https://arxiv.org/abs/2011.03395.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. URL https://arxiv.org/abs/2107.07002.

Dherin, B., Munn, M., and Barrett, D. G. The geometric Occam's razor implicit in deep learning. In *OPT2021: 13th Annual Workshop on Optimization for Machine Learning at ICML*, 2021. URL https://arxiv.org/abs/2111.15090.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. URL https://arxiv.org/abs/1703.04933.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. URL https://arxiv.org/abs/1703.11008.

Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983. URL https://doi.org/10.1080/01621459.1983.10477973.

Forster, M. and Sober, E. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994. URL https://www.journals.uchicago.edu/doi/10.1093/bjps/45.1.1.

Gao, T. and Jojic, V. Degrees of freedom in deep neural networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016. URL https://arxiv.org/abs/1603.09260.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://arxiv.org/abs/2004.07780.

Gelman, A., Simpson, D., and Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017. URL https://doi.org/10.3390/e19100555.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992. URL https://doi.org/10.1162/neco.1992.4.1.1.

Golubeva, A., Gur-Ari, G., and Neyshabur, B. Are wider nets better given the same number of parameters? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=_zx8Oka09eF.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1225–1234. PMLR, 2016. URL https://arxiv.org/abs/1509.01240.

Hinton, G. E. Learning translation invariant recognition in a massively parallel networks. In *Proceedings of the International Conference on Parallel Architectures and Languages Europe*, pp. 1–13. Springer, 1987. URL https://doi.org/10.1007/3-540-17943-7_117.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory (COLT)*, pp. 5–13, 1993. URL https://doi.org/10.1145/168304.168306.

Hope, T., Downey, D., Etzioni, O., Weld, D. S., and Horvitz, E. A computational inflection for scientific discovery. *arXiv preprint arXiv:2205.02007*, 2022. URL https://arxiv.org/abs/2205.02007.

Jefferys, W. H. and Berger, J. O. Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72, 1992. URL https://www.jstor.org/stable/29774559.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1912.02178.

Jiang, Y., Natekar, P., Sharma, M., Aithal, S. K., Kashyap, D., Subramanyam, N., Lassance, C., Roy, D. M., Dziugaite, G. K., Gunasekar, S., Guyon, I., Foret, P., Yak, S., Mobahi, H., Neyshabur, B., and Bengio, S. Methods and analysis of the first competition in predicting generalization of deep learning. In Escalante, H. J. and Hofmann, K. (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 170–190. PMLR, 2021. URL https://proceedings.mlr.press/v133/jiang21a.html.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.04836.

Kubo, M., Banno, R., Manabe, H., and Minoji, M. Implicit regularization in over-parameterized neural networks. *arXiv preprint arXiv:1903.01997*, 2019. URL https://arxiv.org/abs/1903.01997.

Larsen, B. W., Fort, S., Becker, N., and Ganguli, S. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2107.05802.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=ryup8-WCW.

Loukas, A., Poiitis, M., and Jegelka, S. What training reveals about neural network complexity. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RcjW7p7z8aJ.

Lucas, J., Bae, J., Zhang, M. R., Fort, S., Zemel, R., and Grosse, R. Analyzing monotonic linear interpolation in neural network loss landscapes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2104.11044.

MacKay, D. Bayesian model comparison and backprop nets. *Advances in neural information processing systems*, 4, 1991. URL https://papers.nips.cc/paper/1991/hash/c3c59e5f8b3e9753913f4d435b53c308-Abstract.html.

MacKay, D. J. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992a. URL https://doi.org/10.1162/neco.1992.4.3.415.

MacKay, D. J. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992b. URL https://doi.org/10.1162/neco.1992.4.3.448.

Maddox, W. J., Benton, G., and Wilson, A. G. Rethinking parameter counting: Effective dimensionality revisted. *arXiv preprint arXiv:2003.02139*, 2020. URL https://arxiv.org/abs/2003.02139.

Mehta, H., Cutkosky, A., and Neyshabur, B. Extreme memorization via scale of initialization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2008.13363.

Müller, U. K. Measuring prior sensitivity and prior informativeness in large Bayesian models. *Journal of Monetary Economics*, 59(6):581–597, 2012. URL https://doi.org/10.1016/j.jmoneco.2012.09.003.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.15775.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021. URL https://doi.org/10.1088/1742-5468/ac3a74.

Neyshabur, B. *Implicit regularization in deep learning*. Ph.D. dissertation, Toyota Technological Institute at Chicago, 2017. URL https://arxiv.org/abs/1709.01953.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*, 2015. URL https://arxiv.org/abs/1412.6614.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1805.12076.

Nott, D. J., Wang, X., Evans, M., and Englert, B.-G. Checking for prior-data conflict using prior-to-posterior divergences. *Statistical Science*, 35(2):234–253, 2020. URL https://doi.org/10.1214/19-STS731.

Nowlan, S. J. and Hinton, G. E. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4 (4):473–493, 1992. URL https://doi.org/10.1162/neco.1992.4.4.473.

Raghu, M. and Schmidt, E. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*, 2020. URL https://arxiv.org/abs/2003.11755.

Reimherr, M., Meng, X.-L., and Nicolae, D. L. Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021. URL https://arxiv.org/abs/1406.5958.

Schmidhuber, J. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997. URL https://doi.org/10.1016/S0893-6080(96)00127-X.

Smith, S. L. and Le, Q. V. A Bayesian perspective on generalization and stochastic gradient descent. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.06451.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6199.

Wolpert, D. H. On the Bayesian "occam factors" argument for occam's razor. In *Computational Learning and Natural Learning Systems*, volume 3. MIT Press, 1995. URL https://doi.org/10.1162/neco.1992.4.3.415.

Ye, J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998. URL https://www.jstor.org/stable/2669609.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.03530.