A Dynamical Model of Neural Scaling Laws

Blake Bordelon [1 2] **Alexander Atanasov** [3 2] **Cengiz Pehlevan** [1 2]

Abstract

On a variety of tasks, the performance of neural networks predictably improves with training time, dataset size and model size across many orders of magnitude. This phenomenon is known as a neural scaling law. Of fundamental importance is the compute-optimal scaling law, which reports the performance as a function of units of compute when choosing model sizes optimally. We analyze a random feature model trained with gradient descent as a solvable model of network training and generalization. This reproduces many observations about neural scaling laws. First, our model makes a prediction about why the scaling of performance with training time and with model size have different power law exponents. Consequently, the theory predicts an asymmetric compute-optimal scaling rule where the number of training steps are increased faster than model parameters, consistent with recent empirical observations. Second, it has been observed that early in training, networks converge to their infinite-width dynamics at a rate $1/\textit{width}$ but at late time exhibit a rate $\textit{width}^{-c}$, where c depends on the structure of the architecture and task. We show that our model exhibits this behavior. Lastly, our theory shows how the gap between training and test loss can gradually build up over time due to repeated reuse of data.

1. Introduction

Large scale language and vision models have been shown to achieve better performance as the number of parameters and number of training steps are increased. Moreover, the scaling of various loss metrics (such as cross entropy or MSE test loss) has been empirically observed to exhibit remarkably regular, often power law behavior across several orders of magnitude (Hestness et al., 2017; Kaplan et al., 2020). These findings are termed "neural scaling laws".

Neural scaling laws play a central role in modern deep learning practice, and have substantial implications for the optimal trade-off between model size and training time (Hoffmann et al., 2022), as well as architecture selection (Alabdulmohsin et al., 2023). Understanding the origin of such scaling laws, as well as their exponents, has the potential to offer insight into better architectures, the design of better datasets (Sorscher et al., 2022), and the failure modes and limitations of deep learning systems. Yet, many questions about neural scaling laws remain open.

In this paper, we introduce and analyze a solvable model which captures many important aspects of neural scaling laws. In particular, we are interested in understanding the following empirically observed phenomena:

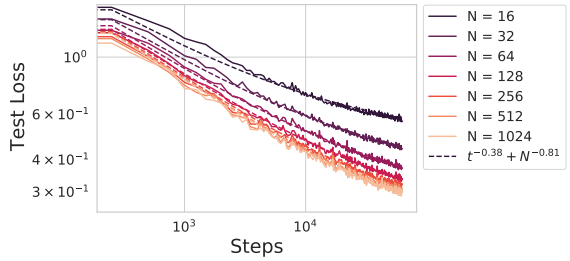
Test Loss Scales as a Power-law in Training Time and Model Size and Compute. In many domains of deep learning, the test loss of a model with N trainable parameters trained for t iterations has been found to scale as $\mathcal{L}(t, N) \approx \mathcal{L}_0 + a_t t^{-r_t} + a_N N^{-r_N}$ (Kaplan et al., 2020; Hoffmann et al., 2022). These scaling law exponents r_t, r_N generally depend on the dataset and architecture. We demonstrate scaling laws on simple vision and language tasks in Figure 1. The compute is proportional to the number of steps of gradient descent times the model size $C \propto Nt$. Setting N and t optimally gives that test loss scales as a power law in C. This is the *compute optimal scaling law*.

Compute-Optimal Training Time and Model Size Scaling Exponents Are Different. A discrepancy in exponents r_t and r_N is usually observed to some degree depdending on the data distribution and architecture Hoffmann et al. (2022); Bachmann et al. (2024). The gap between exponents would lead to asymmetric compute-optimal scaling of parameters. For compute budget C, model size should scale $N \propto C^{c_1}$ and training time $t \propto C^{c_2}$ with $c_2 > c_1$. This difference in exponents led to a change in the scaling rule for large language models, generating large performance gains.

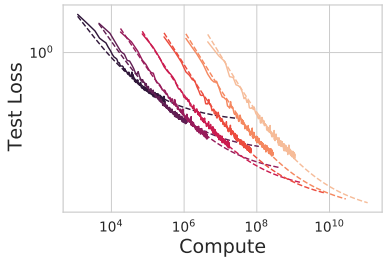
Larger Models Train Faster. Provided feature learning is held constant across model scales (i.e. adopting mean-field or μP scaling), wider networks tend to train faster (Yang et al., 2021) (Figure 1). If training proceeds in an online/one-pass setting where datapoints are not repeated,

[1]SEAS, Harvard University [2]Kempner Institute, Harvard University [3]Department of Physics, Harvard University. Correspondence to: Cengiz Pehlevan <cpehlevan@seas.harvard.edu>.

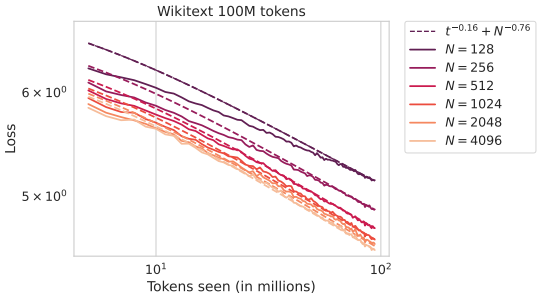
Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



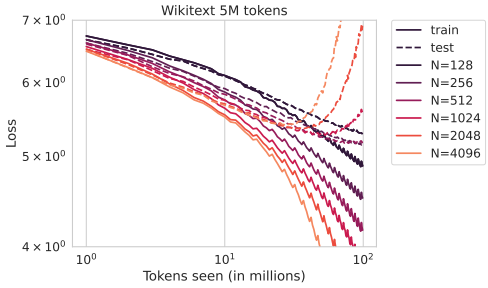
(a) One-Pass CIFAR-5M Test Dynamics



(b) CIFAR-5M Compute Scaling



(c) Wikitext-100M Train and Test



(d) Wikitext-5M Train and Test

Figure 1. Train and test losses (cross-entropy) as a function of training time t and width N. For models trained online, we do not make a distinction between training and test error because each new batch is drawn fresh and would have the same loss in expectation as an independent test set. (a) The test loss of a residual CNN on CIFAR-5M is well described by a fit of the form $\mathcal{L} \sim t^{-\alpha} + N^{-\beta}$ in the online training regime. (b) The compute optimal strategy requires scaling up both model size and training time simultaneously. (c) Transformer training on wikitext with 100M tokens before data-repetition. Model performance is monotonic in width N. (d) Wikitext with 5M subsampled tokens. Larger width N is not always better as wider models can overfit.

then the wider models will also obtain lower test loss at an equal number of iterations. This observation has been found to hold both in overparameterized and underparameterized regimes (Bordelon & Pehlevan, 2023; Vyas et al., 2023).

Models Accumulate Finite-Dataset and Finite-Width Corrections. Early training can be well described by the learning curves for stochastic gradient descent without reuse of samples (termed the online/ideal limiting dynamics), however over time the effect of reusing data accumulates and leads to worse test performance (Nakkiran et al., 2021b; Mignacco et al., 2020; Ghosh et al., 2022; Muennighoff et al., 2023). Similarly the gaps in model performance across various model sizes also grow with training time (Yang et al., 2021; Vyas et al., 2023). Figure 1 (d) shows overfitting and reversal of "wider is better" phenomenon due to data reuse.

Scaling Exponents are Task-Dependent at Late Training Time, but not at Early Time. Prior works (Dyer & Gur-Ari, 2020; Atanasov et al., 2023; Roberts et al., 2022; Bordelon & Pehlevan, 2023) predict early-time finite-width loss corrections that go as $1/\textit{width}$ near the infinite width limit in either lazy or feature-learning regimes. Bahri et al. (2021) et al provide experiments demonstrating the $1/\textit{width}$ convergence. However, finite-width models trained for a long time exhibit non-trivial exponents with respect to model width (Kaplan et al., 2020; Vyas et al., 2023). See Figure 1 for examples of nontrivial scalings at late time on CIFAR-5M and Wikitext.

Ensembling is Not the Same as Going Wider. Near the limit of infinite width, finite models can be thought of as noisy approximations of the infinite-width model with noise that can be eliminated through ensembling (Dyer & Gur-Ari, 2020; Geiger et al., 2020; Atanasov et al., 2023). However recent experiments (Vyas et al., 2023) indicate that ensembling is not enough to match performance of larger models.

These phenomena are not unique to deep networks, but can be observed in linear models, or linearized neural networks operating in the lazy/kernel regime. Though this regime does not capture feature learning, it has benefit of analytical tractability. In this paper, we focus on such linearized models to attempt to gain insight into the dynamics of training.

To attempt to explain these phenomena, we develop a mathematically tractable model of neural scaling laws which allows one to simultaneously vary time, model size, and dataset size. Our contributions are as follows:

1. We analyze the learning dynamics of a structured and randomly projected linear model trained with gradient flow, discrete time SGD, and momentum. In an asymptotic limit of the model, we obtain a dynamical mean field theory (DMFT) description of the learning curve in terms of correlation functions, which measure the cross-time correlation of training and test errors, and response functions which measure sensitivity of the dynamics to small perturbations.
2. We solve for the response functions exactly in Fourier domain. This solution reveals faster training for larger models. The low frequency range of these functions allow us to extract the long time limit of the loss.
3. We show that the model and data corrections to the dynamics accumulate over time. At early time, each of these corrections has a universal scaling, consistent with prior works (Bahri et al., 2021).
4. For power-law structured features we show that the model exhibits power law scaling of test loss with time, model size and dataset size. While the data and model exponents are the same, the time and model exponents are different in general. We show that this gives rise to an asymmetric compute optimal scaling strategy where training time increases faster than model size.
5. Our theory explains why ensembling is not compute optimal as it gives less benefit to performance than increase in model size.
6. We observe in Section 5.1 that feature learning networks can obtain better power law scalings, leading to a better compute optimal frontier. We empirically study this phenomenon in Appendix L.

1.1. Related Works

The learning curves for linear models with structured (non-isotropic) covariates, including infinite-width kernel regression, have been computed using tools from statistical physics and random matrix theory (Bordelon et al., 2020; Spigler et al., 2020; Canatar et al., 2021; Simon et al., 2021; Bahri et al., 2021; Hastie et al., 2022). Mei & Montanari (2022) analyzed a linear model with random projections of isotropic covariates. There, they study the limiting effects of width and dataset size, and observe model-wise and sample-wise double descent. In Adlam & Pennington (2020a) a related model is used to study the finite-width neural tangent kernel (NTK) (Jacot et al., 2018) of a given network. Further, d'Ascoli et al. (2020) and Adlam & Pennington (2020b) extend this analysis to understand the different sources of variance in the predictions of random feature models and the effect of ensembling and bagging on the test loss. Other works have extended this to models where an additional untrained projection is applied to the structured covariates (Loureiro et al., 2021; 2022; Zavatone-Veth et al., 2022; Atanasov et al., 2023; Maloney et al., 2022; Zavatone-Veth & Pehlevan, 2023; Ruben & Pehlevan, 2023; Simon et al., 2023). Within this literature, which considered fully trained models, the works of (Bordelon et al., 2020; Spigler et al., 2020) derived power-law decay rates for power-law features which were termed resolution limited by Bahri et al. (2021) and recovered by Maloney et al. (2022).

However, we also study the dependence on training time. The $t \to \infty$ limit of our DMFT equations recovers the final losses computed in these prior works. While these prior works find that the scaling exponents for model-size and dataset-size are the same, we find that the test loss scales with a different exponent with training time, leading to a different (model and task dependent) compute optimal scaling strategy.

DMFT methods have been used to analyze the test loss dynamics for general linear and spiked tensor models trained with high-dimensional random data (Mannelli et al., 2019; Mignacco et al., 2020; Mignacco & Urbani, 2022) and deep networks dynamics with random initialization (Bordelon & Pehlevan, 2022b; Bordelon et al., 2023). High dimensional limits of SGD have been analyzed with Volterra integral equations in the offline case (Paquette et al., 2021) or with recursive matrix equations in the online case (Varre et al., 2021; Bordelon & Pehlevan, 2022a). Random matrix approaches have also been used to study test loss dynamics in linear regression with isotropic covariates by (Advani et al., 2020) and for random feature models in (Bodin & Macris, 2021). In this work, we consider averaging over both the disorder in the sampled dataset and the random projection of the features simultaneously using DMFT.

Other models and hypotheses for scaling laws instead rely on a discrete collection of subtasks or skills which are learned as compute grows (Caballero et al., 2022; Arora & Goyal, 2023; Michaud et al., 2023). Our theory instead focuses on spectral components of a data distribution.

2. Setup of the Model

We consider a "teacher-student" setting, where data sampled from a generative teacher model is used to train a student random feature model. The teacher and student models mismatch in a particular way that will be described below. This mismatch is the key ingredient that leads to most of the phenomena that we will discuss.

Teacher Model. Take $\boldsymbol{x} \in \mathbb{R}^D$ to be drawn from a distribution $p(\boldsymbol{x})$ with a target function $y(\boldsymbol{x})$ expressible in terms of base features $\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^M$ up to noise:

$$y(\boldsymbol{x}) = \frac{1}{\sqrt{M}} \boldsymbol{w}^\star \cdot \boldsymbol{\psi}(\boldsymbol{x}) + \sigma \epsilon(\boldsymbol{x}). \tag{1}$$