Feature-Learning Networks Are Consistent Across Widths At Realistic Scales

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the effect of width on the dynamics of feature-learning neural networks across a variety of architectures and datasets. Early in training, wide neural networks trained on online data have not only identical loss curves but also agree in their point-wise test predictions throughout training. For simple tasks such as CIFAR-5m this holds throughout training for networks of realistic widths. We also show that structural properties of the models, including internal representations, pre-activation distributions, edge of stability phenomena, and large learning rate effects are consistent across large widths. This motivates the hypothesis that phenomena seen in realistic models can be captured by infinite-width, feature-learning limits. For harder tasks (such as ImageNet and language modeling), and later training times, finite-width deviations grow systematically. Two distinct effects cause these deviations across widths. First, the network output has initialization-dependent variance scaling inversely with width, which can be removed by ensembling networks. We observe, however, that ensembles of narrower networks perform worse than a single wide network. We call this the *bias* of narrower width. We conclude with a spectral perspective on the origin of this finite-width bias.

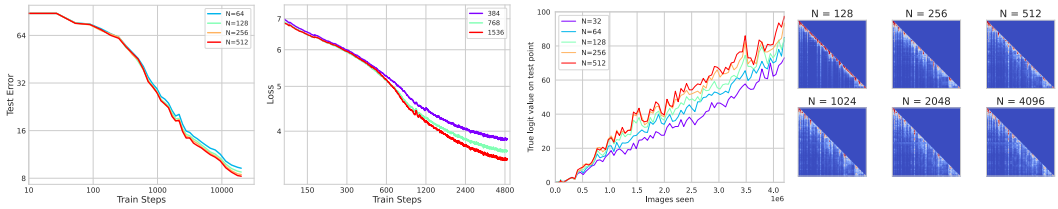
1 Introduction

Studies of large-scale language and vision models have shown that models with a larger number of parameters achieve better performance [1, 2]. Motivated by the success of large-scale models, several theories of deep learning have been developed, including large-width limits. One infinite width limit considered in [3, 4] gives rise to a initialization-independent and constant neural tangent kernel (NTK). However, modern large-scale networks adapt their features to structure in the data even at very large widths. In practice, they are not well-described by NTK theory [5, 6].

Recently, several works have identified an alternative parameterization of neural networks that preserves feature-learning even at infinite width [7–11]. In particular, the maximal update parameterization (μP) of [10] gives an infinite-width limit of a given finite-width network in standard parameterization (SP) with similar feature learning capability. These limits are attractive in that they allow for feature learning while also rendering several network properties (output logits, feature kernels, ...) deterministic rather than dependent on the precise initialization of the network. In addition, [12] found that wider networks perform better with all other architectural details held fixed. Finally, in this limit, neurons take on a simple interpretation as $i.i.d.$ draws from a width-independent distribution throughout training, enabling theoretical analysis of feature learning [11]. The existence of infinite-width feature-learning limits motivates us to ask:

Question: ***Can realistic-width neural networks be accurately described by their infinite-width feature-learning limits?***

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) CIFAR-5m (b) C4 (c) ImageNet (d) Wikitext-103

Figure 1: Consistency of large width behavior across tasks, architectures, observables. a) Loss curves for Resnets on Cifar-5M in μP are nearly to identical at large widths (see also Figure 2). b) For GPT-2 on the C4 dataset [13] the loss curves agree at early times and deviate at late times, but wider networks agree for longer (see also Figure 2 and appendices for Wikitext-103) c) The values that ResNets put on the correct logit for ImageNet appear to converge as the width grows (see also Figure 3). d) The attention matrices for transformers on Wikitext-103 become nearly identical as width increases (for quantitative metrics see Figure 4.)

We attempt to answer this question by training networks of varying widths on vision and language tasks for realistic datasets and architectures. We put all of our networks in μ-parameterization, adopting the package [14] introduced in [12]. We give an affirmative answer to the above question in the online setting. Concretely, we focus on on the online setting, where data is not repeated during SGD, and track the following quantities across widths:

- the losses throughout training;
- the predictions of the networks on individual points throughout training;
- the learned representations, summarized by the feature kernels; preactivation distributions; and, for transformers, attention matrices;
- and dynamical phenomena such as the edge of stability governing the top Hessian eigenvalues, as well as large learning rate and small batch size effects on the loss.

On each of these metrics, we show that sufficiently wide neural networks converge to consistent behavior across widths. In Figure 1, we show loss curves, logit predictions, and attention matrices approach consistent behavior as width is increased across several architectures and datasets. We further observe that the widths which achieve this consistent behavior are within the range of those used in practice. We use large-width consistency as a proxy for achieving the limiting infinite-width behavior.

We say that a network property is consistent if beyond some width, its values all lie within some small interval with high probability. We measure consistency by showing that a quantity's deviations between successive widths decrease as the widths are increased, and that its value for narrower networks systematically approaches its value for the largest trained network.

Our results show the following:

- For simple vision tasks such as CIFAR-5m [15], ResNets with practical widths achieve near consistent loss curves across widths.
- Beyond the loss curves, the individual predictions of the networks agree pointwise. That is, the logits agree on test points throughout the training process. We further show that internal representations as measured by distributions of neuron preactivations and feature kernels in various layers are consistent across widths.
- For harder tasks such as ImageNet and language modeling, loss curves are consistent across widths early in training. As training progresses, loss curves for narrow networks deviate smoothly from the loss curves of wider networks. The effective width required to reach infinite-width behavior thus increases with training time. Conversely, as network size grows we approximate the infinite width network for a larger number of training steps.
- Finite-width neural networks have variance in the learned function due to initialization seed. This variance depends inversely on the width. We study ensembles of networks over different initializations to remove this noise. Further, by analyzing ensembles of networks, we can do a

bias-variance decomposition of the effects of finite width. We find that finite-width bias plays an important role. Equivalently, ensembling narrow networks does not yield infinite-width behavior.

- In the setting of offline learning, at late times one can over-fit the training set. We observe that this leads to larger gaps in network behavior across widths, and can break the trend that wider networks perform better.
- We develop a spectral perspective on the origin of the finite-width bias by analyzing it in a simple setting of a lazy network learning a simple task. We then apply this perspective to a CNN trained on CIFAR-5m.

The consistency across large widths strongly suggests that the dynamics and predictions of realistic-scale networks can be effectively captured by their infinite-width feature learning limits. For realistic tasks, as the width is increased, a larger interval of training can be characterized by this infinite-width limit.

Our results have implications for interpretability, as the agreement of internal representations suggest that many other phenomena, such as transfer learning with linear probes or fine-tuning, in-context learning [16, 17], the emergence of outliers [18], and the emergence of induction heads [19] may be understood from the perspective of infinite-width feature learning networks.

1.1 Related Works

Empirically, the scaling of relevant quantities with width in the standard or neural-tangent parameterizations was thoroughly studied in [20]. In the latter parameterization, sufficiently wide networks give a kernel method with the infinite-width NTK. Several papers have shown that in practice the NTK limit insufficiently characterizes realistic deep neural networks [5, 21, 6]. Attempts to capture feature learning and predictor variance from perturbative series around infinite-width dynamics show that finite-width variance and kernel adaptation scale as $1/N$ [22–24] for width N. A $1/N$ scaling of generalization error with width was empirically verified on many tasks [25, 26]. The effect of width on generalization in the feature-learning regime was empirically studied in [27] in the relatively limited setting of multi-layer perceptrons (MLPs) on polynomial tasks. There, the variance of the finite-width NTK at the end of training adversely affected generalization.

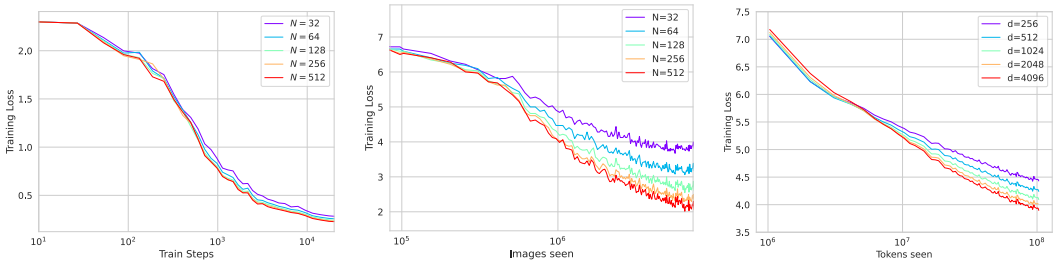
The authors of [28] identified that altering the output scale α of any network could increase or decrease feature learning in a neural network. Large values of α correspond to the "lazy limit" where the network's features don't evolve. A follow up study noticed that rescaling the output by $\alpha = \alpha_0/\sqrt{N}$ for width N networks gave consistent behavior of feature learning and losses in small scale experiments [8]. Several works have studied this regime of training in the two-layer limit, known as "mean field" parameterization, where features are still learned even at infinite width [29, 7, 30, 31]. Extensions of this model to deeper networks were studied in [32–35, 10, 11]. A theory of finite-width corrections to networks in this parameterization was studied in [36]. A very general set of parameterization principles, termed μP, was introduced to give a well defined feature learning limit for a wide range of architectures including RNNs, CNNs, MLPs and transformers [10]. [12] demonstrated that this parameterization nearly fixes optimal hyperparameters across network widths, allowing for hyperparameter transfer from small to large widths. This work also empirically noted that wider networks always outperformed narrower networks in this parameterization.

Our paper focuses on networks in μP and attempts to study the consistency of many relevant network properties across widths. We perform a fine-grained analyses of more realistic models throughout the dynamics of training. To the best of our knowledge, this is the first such paper to study the consistency of network outputs, internal representations, and dynamics across widths.

2 Consistency of large-width behavior in online learning

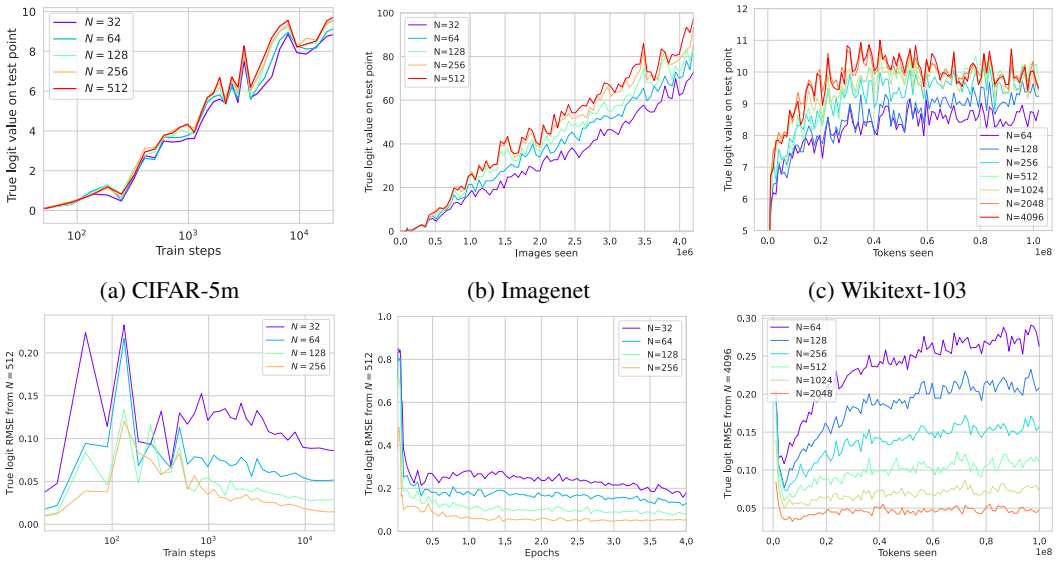
We focus on studying the effect of width in the setting of neural networks learning a task in the online setting. Online learning is representative of many modern settings of deep learning, and as will be shown in Section 3, obviates consideration of memorization and over-fitting in offline learning that can lead to large differences in networks across widths.

In what follows, the variable N will denote the width of a given network. For vision tasks, this will correspond to the number of channels in each layer. For transformers, in the notation of [37],



(a) CIFAR-5m (b) Imagenet (c) Wikitext-103

Figure 2: In the online learning setting, train loss improves as width grows. For sufficiently wide networks, the training lost is consistent across widths. For Cifar-5m this consistency is observed over all of training. For harder tasks like Imagenet and Wikitext-103, networks of different widths agree up until a width-dependent time-step where narrower networks begin performing worse.



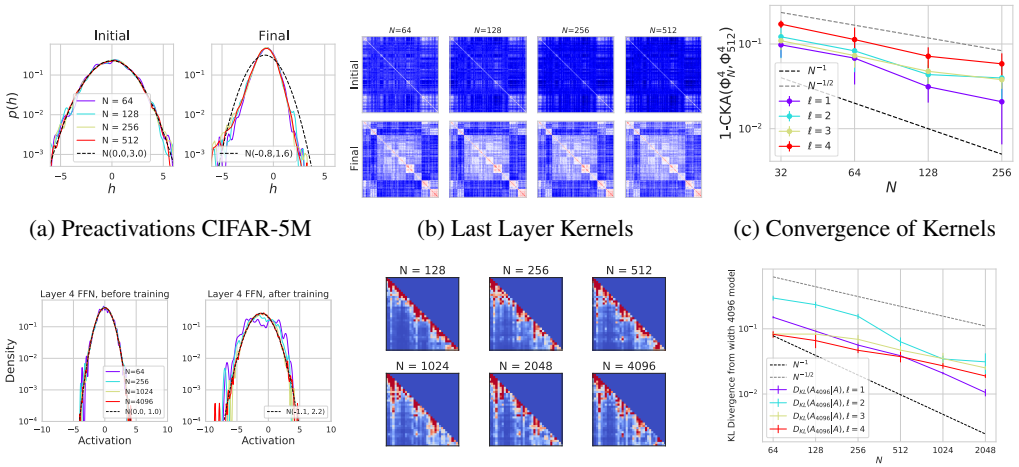
(a) CIFAR-5m (b) Imagenet (c) Wikitext-103

(d) CIFAR-5m (e) Imagenet (f) Wikitext-103

Figure 3: The output logits on a fixed test point diplays stable behavior at large enough widths. a) Value of network on correct class logit over time as width is varied for CIFAR-5m. b) Same plot for Imagenet for a fixed image in the test set c) Same plot for Wikitext-103 for a fixed masked token. Across the board the widest networks behave similarly. Next, we use the widest network as a proxy for the infinite-width limit, and compare the logit predictions of narrower networks against that. d) For CIFAR-5m, the relative root-mean-squared error over the test set of the distance to the value that the widest network puts on the correct logit. e) The same for Imagenet. f) The same for Wikitext-103. We see a striking regularity of networks converging to the widest one as the width grows. In Appendix B, we also compare networks of successive widths and show the the difference shrinks.

$N = d_{model} = hd_k = hd_v$ and $d_{ffn} = 4N$. Here, h is the number of heads, which we will keep fixed. d_{model} is the embedding dimension of the tokens as well as the dimension of the residual stream. d_k is the dimension over which the dot products in the attention are calculated and d_v is the dimension of the values in the attention layers. d_{ffn} is the hidden width of the feedforward networks (FFN).

Convergence of loss curves We begin by showing (Fig. 2) that the loss curves for sufficiently wide networks on a given task achieve consistent behavior across widths. Throughout the paper we measure train loss in terms of crossentropy. For all tasks, at early times large widths agree, but for more complicated tasks such as ImageNet or Wikitext-103, learning curves of narrower network deviate from those of wider ones.



(a) Preactivations CIFAR-5M (b) Last Layer Kernels (c) Convergence of Kernels

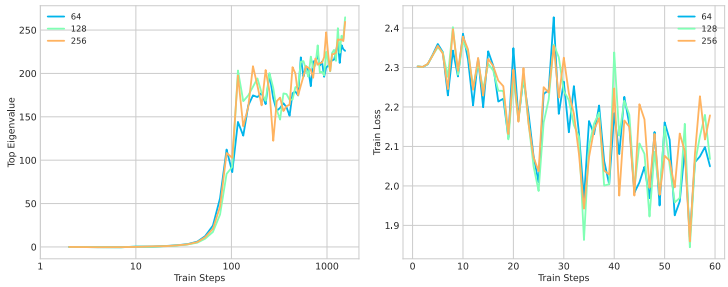
(d) FFN Preactivations Wikitext (e) Attention Matrices (f) Convergence of Attention

Figure 4: Learned features are consistent across a large range of widths in realistic tasks. (a) The distribution (over neurons) of preactivation values h in the final block of the ResNet18 trained on CIFAR-5M. At initialization, the densities are all well approximated by the Gaussian with matching mean and variance (dashed black). After feature learning, the density has shifted and become non-Gaussian (poor match with dashed black), yet is still strikingly consistent across widths. (b) Feature kernels are also consistent across widths. (c) The centered kernel alignment CKA [41, 42] of kernels increases towards 1.0 as $N \to \infty$. The $1/\sqrt{N}$ and $1/N$ trends are plotted for reference. d) The preactivation histogram for a transformer on Wikitext-103. At initialization the Gaussian of best fit is the standard normal. After training the histograms are still quite Gaussian, with different moments. e) A variant of Figure 1 d) at a smaller sequence length. Attention matrices are consistent at large widths. f) Both FFN kernels and attention matrices converge as width grows. The $1/N$ and $1/\sqrt{N}$ trends are plotted for reference.

The width beyond which networks emulate infinite-width behavior depends on the complexity of the task. For more difficult tasks, larger widths are required for the loss curves to converge. For simple tasks such as CIFAR-5m we find that widths as narrow as 128 are essentially consistent with infinite width-behavior for an entire pass through the 5 million image dataset. For ImageNet, widths near 512 are close to consistent for four passes through the dataset with heavy data augmentation. These widths are well within the range of those practically for images [38, 39]. For transformers going through a single full pass of Wikitext-103, widths on the order of 4000 are required. Early transformer models certainly had hidden widths of order 4k [40], and more recent models such as GPT-3 have widths going up to 12288 [16], so this is also within the regime of realistic width.

Pointwise convergence of predictions Beyond the convergence of the training loss curves, we observe that the logits of a network on a fixed test point become consistent as width grows. This test point can be an image in the test set or a masked token in the validation set. In plots a), b), and c) of Figure 3, we show that for a specific held-out test point, the value of the network on the correct logit becomes consistent as the width grows. In d), e), and f) we plot the root mean squared distance to the widest networks logits over the test set. We further study the difference between successive widths in Figure 11.

Convergence of representations In addition to loss and prediction dynamics, we also examine whether learned representations in these models are consistent across widths. Mean field theories of neural network dynamics predict that sufficiently wide networks should have identical kernels (and attention matrices for transformers) and that all neurons in a layer behave as independent draws from an initialization-independent single-site distribution [7, 10, 11, 43, 44]. To test whether realistic finite-width feature learning networks are accurately captured by this limit, in Figure 4, we analyze the feature kernels and preactivation distributions before and after training as well as the attention matrices in transformer models trained on Wikitext-103. We see qualitative consistency in the plots of kernels and attention matrices in b) and c) which can be made quantitatively precise by plotting the distance to the widest networks and showing systematic convergence in c) and f).



(a) Edge of Stability (b) Small Batch size effects

Figure 5: Convergence of dynamical phenomena across width for CIFAR-5m

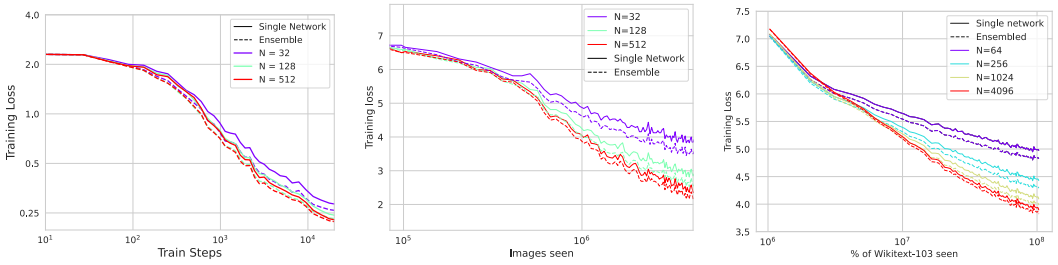
Convergence of dynamical phenomena In Figure 5a, we show that the sharpness, defined as the top eigenvalue of the loss Hessian, grows steadily to a final value that it then fluctuates around. This is a small-batch analogue of the the edge-of-stability phenomenon identified in [45]. We also show in Figure 5b that on CIFAR-5m task, at early times, the individual variations due to batch noise and large learning rate effects can be consistently captured across widths for μP networks. In Appendix D, we further demonstrate sharp agreement of large learning rate and small batch size phenomena for MLPs learning a simple task. There, we show that while μP leads to strikingly consistent loss curves, SP does not.

3 Deviations from large-width behavior

The consistency observed in Section 2 may break later during training in either the online or offline settings. In the online setting, deviations owing to narrow width compound over time and lead to two sources of error relative to the infinite width limit which we describe in 3.1. In the offline setting, where data is recycled several times, networks over-fit the training data, which can lead to larger gaps between widths and can break the trend that wider networks perform better.

Finite-width effects introduce an initialization dependence to the network, leading to additional variance in the learned function and hindering generalization [25–27]. This initialization-dependent variance can be mitigated by averaging the output logits of a sufficiently large ensemble of networks [46]. Using the bias-variance decomposition terminology, we refer to the discrepancy in performance between an ensembled network and the expected performance of a single network the *variance*, and the gap between an ensembled network and the behavior of infinite-width network as the bias of narrower width. By definition, the expected difference in loss between a single finite-width network and an infinite-width network is the sum of the bias and the variance. Below, we investigate the behavior of bias and variance in networks across various vision and language tasks.

3.1 Online training

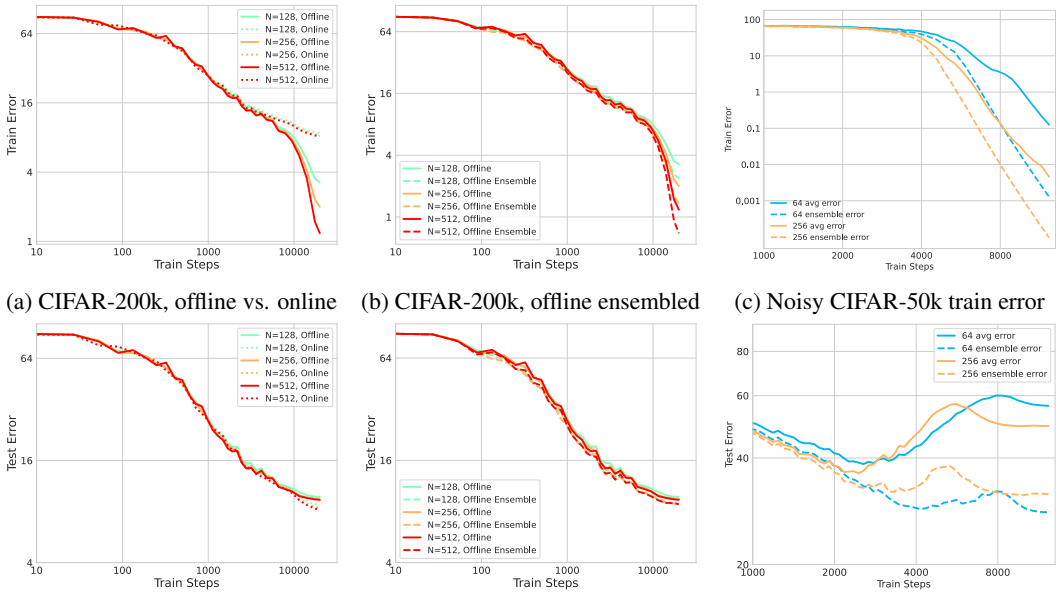


(a) CIFAR-5m (b) Imagenet (c) Wikitext-103

Figure 6: Loss curves and their ensembles in the online setting. Ensembling reduces the training loss, but a large ensemble of narrow networks do not achieve the performance of a single wider network.

Figure 6 shows that at large widths, both single networks and ensembles of networks achieve comparable error. In this regime, all the networks are consistent and increasing the width has a very marginal effect, as does ensembling. At narrower widths, variance is nontrivial (i.e. ensembling helps) but bias is much larger than variance. Single wide networks outperform ensembles of narrower networks. By comparing a) with b) and c) of Figure 6, we see that harder tasks induce larger bias gaps. Prior theoretical work [26, 27] has focused mostly on studying the variance term. In Section 4 we study the bias from a theoretical perspective.

3.2 Offline Training



(a) CIFAR-200k, offline vs. online (b) CIFAR-200k, offline ensembled (c) Noisy CIFAR-50k train error

(d) CIFAR-200k, offline vs. online. (e) CIFAR-200k, offline ensembled (f) Noisy CIFAR-50k test error

Figure 7: Top Row: Effects of offline training on train metrics. In (a) and (b) we do multi-epoch training on CIFAR-200k. We see that both bias and variance for train error are magnified by offline training and do not tend to 0 for the largest widths we could try. (c) we do multi-epoch training on noisy CIFAR-50k and again observe large bias and variance terms at large widths. Bottom Row: Effects of offline training on test metrics. In (d) and (e) we do multi-epoch training on CIFAR-200k. We see that both bias and variance for test error are near 0 at large widths. (f) We train on noisy CIFAR-50k and observe that "wider is better" is violated for ensembled networks.

In offline learning, which refers to multi-epoch training, we encounter several unexpected phenomena that challenge the width consistency observed in the previous section, even at large widths. To compare offline learning with online learning, we utilize CIFAR-200k, a 200k sized random subset of CIFAR-5m. Previous studies have demonstrated that label noise contributes to an increase in overfitting [47]. In order to investigate how width consistency changes with overfitting and double descent, we conduct experiments on a noisy label version of CIFAR-50k (50k sample from CIFAR-5m), where 50% of the labels are noisy. Additional ImageNet experiments are presented in Appendix F. As offline training achieves near-zero error, we need to compare very small quantities. To accomplish this, we will plot and compare all quantities on a logarithmic scale. The following phenomena are observed:

- Single network performance on the training set does not converge with width, even at high widths (Figure 7 (a)). In other words, the combined bias and variance does not reach zero, even with substantial widths. This is in contrast to the online runs.
- Ensembling (Figure 7 (b)) reveals that both bias and variance terms individually fail to reach zero, even at high widths.
- Regarding test performance, both bias and variance tend to zero as width increases, demonstrating an instance of benign overfitting (Figure 7 (d) and (e)).

- When working with the noisy label version of CIFAR-50k, we observe clear overfitting and stepwise double descent [47] as training progresses (Figure 7 (f)). Notably, we observe significant deviations in width for single network performance, indicating that the benign overfitting observed in Figure 7 (d) and (e) is dataset-dependent. Furthermore, variance is found to be much larger than in the non-noisy experiments.
- Surprisingly, we discover (Figure 7 (f)) that some ensembled narrower width networks outperform ensembled wider networks. This presents a counterexample to the "wider is better" phenomenon [12] for ensembled networks. We hypothesize that such counterexamples can only exist in the context of offline training.

4 Spectral perspective on the width-dependent bias

In this last section, we develop a toy model in which the effect of finite-width bias can be clearly seen. We analyze it first in the simple setting of an MLP fitting a polynomial in the lazy limit. Here, all the dynamics are well-captured by the finite-width empirical neural tangent kernel (eNTK). By studying the spectral properties of this kernel across widths, we see that finite-widths lead to an eNTK with worse finite-width bias, even after ensembling over initializations.

Concretely, we see that although the eigenvalue spectrum of the ensembled eNTK is not substantially affected by finite width, the decomposition of the task into eNTK eigenvectors changes, with narrower widths putting more of the task into smaller eigenmodes that take longer to be learned. We then apply this analysis to the after-kernel of the trained ResNets on CIFAR 5m, and find similar behavior. Prior literature has demonstrated that many of the properties of the final learned function are captured by the after-kernel [48–50].

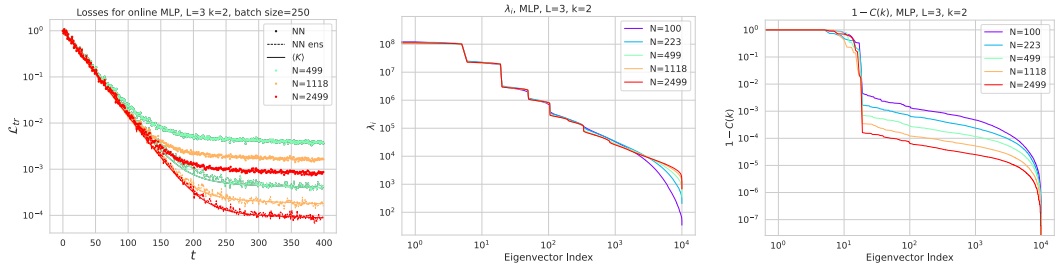
We consider a model of online learning where a large batch of data from the population distribution $p(\boldsymbol{x})$ is sampled at each step. This leads to approximate gradient flow dynamics $\frac{d}{dt}\boldsymbol{\theta} = -\frac{1}{2}\nabla_\theta \mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x},\boldsymbol{\theta}) - y(\boldsymbol{x}))^2$ (Appendix E). To analyze this equation, we choose a fixed orthonormal basis $\{\psi_k(\boldsymbol{x})\}$ for the space $L^2(\mathbb{R}^D, p(\boldsymbol{x})d\boldsymbol{x})$ of square-integrable functions on input space. The function $f(\boldsymbol{x})$, residual error $\Delta(\boldsymbol{x}) = y(\boldsymbol{x}) - f(\boldsymbol{x})$, and the kernel $K(\boldsymbol{x},\boldsymbol{x}',t)$ can be expressed in this basis as $f(\boldsymbol{x},t) = \sum_k f_k\,\psi_k(\boldsymbol{x})$, $\Delta(\boldsymbol{x},t) = \sum_k \Delta_k \psi_k(\boldsymbol{x})$, and $K(\boldsymbol{x},\boldsymbol{x}',t) = \sum_{k\ell} K_{k\ell}(t)\psi_k(\boldsymbol{x})\psi_\ell(\boldsymbol{x}')$, respectively. Their training evolution is given by:

$$\frac{d}{dt}f(\boldsymbol{x},t) = \mathbb{E}_{\boldsymbol{x}'\sim p(\boldsymbol{x})} K(\boldsymbol{x},\boldsymbol{x}',t)\Delta(\boldsymbol{x}',t) = -\sum_{k\ell} K_{k\ell}(t)\Delta_\ell(t)\psi_k(\boldsymbol{x}). \tag{1}$$

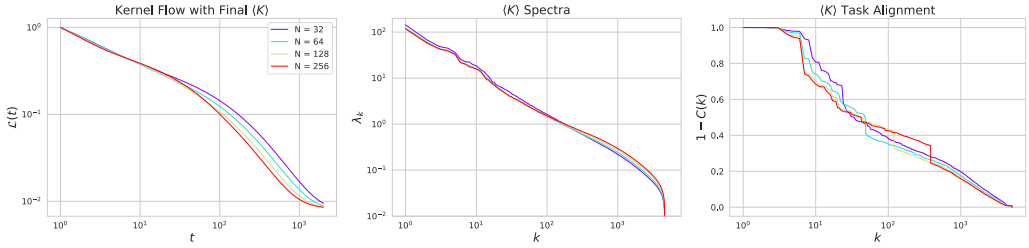
The statistics of the dynamical NTK matrix $K_{kl}(t)$ summarizes the statistics of the error dynamics $\Delta(\boldsymbol{x},t)$ at any level of feature learning. At infinite width, $K_{k\ell}(t)$ is deterministic, while at finite width, it receives a $\mathcal{O}(N^{-1})$ mean displacement and a $\mathcal{O}\left(N^{-1/2}\right)$ fluctuation around its mean [22, 23, 36]. We consider approximating the dynamics of the ensembled predictor by $\frac{d}{dt}\left\langle f_k(t)\right\rangle_{\theta_0} \approx \sum_\ell \left\langle K_{k\ell}(t)\right\rangle \left\langle \Delta_\ell(t)\right\rangle$. Here, $\langle\cdot\rangle$ denotes averages over initializations. This expression neglects the contribution from $\mathrm{Cov}(K_{k\ell},\Delta_\ell)$. We show that this approximation is accurate in depth-3 MLPs trained on Gegenbauer polynomial regression tasks in Figure 8 a). For more details see Appendix A.

In the lazy limit, the kernel is static and we choose ψ_k to diagonalize $\langle K_{k\ell}\rangle = \delta_{k\ell}\lambda_k$. This yields the loss dynamics $\mathcal{L}(t) = \sum_k \left\langle y(\boldsymbol{x})\psi_k(\boldsymbol{x})\right\rangle^2 e^{-2\lambda_k t}$. We can therefore quantify alignment of eigenfunctions to task with the cumulative power distribution $C(k) = \sum_{\ell<k}\left\langle y(\boldsymbol{x})\psi_\ell(\boldsymbol{x})\right\rangle_{\boldsymbol{x}}^2 / \left\langle y(\boldsymbol{x})^2\right\rangle_{\boldsymbol{x}}$ [51]. If $C(k)$ rises rapidly with k then the loss falls faster [51]. In this limit, there are two ingredients that could make the bias dynamics across widths distinct. First, the eigenvalues λ_k which set the timescales could be width-dependent. Second, the eigenfunctions $\psi_k(\boldsymbol{x})$ that diagonalize $\langle K\rangle$ can change with width. In Figures 8 b) and c) we show that the dominant effect is the latter. Finite width corrections do not substantially effect the spectrum but spread out target function power into slower modes in narrower networks.

To test whether these findings continue to hold in more realistic experiments, we computed the final NTKs (after kernels) of the ResNet-18 models trained on CIFAR-5M (specifically the models from Figures 3, 4). We ensemble average to get kernel $\langle K_{c,c'}(\boldsymbol{x},\boldsymbol{x}')\rangle$ for output channels c, c' and input images $\boldsymbol{x},\boldsymbol{x}'$. We then compute the kernel gradient flow corresponding to MSE training on the true target function for CIFAR-5M $\frac{d}{dt}\Delta_c(\boldsymbol{x}) = -\sum_{c'}\mathbb{E}_{\boldsymbol{x}'}\left\langle K_{c,c'}(\boldsymbol{x},\boldsymbol{x}')\right\rangle \Delta_{c'}(\boldsymbol{x}')$ from initial condition



(a) MLP Online training losses (b) MLP Kernel Spectra (c) MLP Task Power



(d) CIFAR-5M After Kernel (e) After Kernel Spectra (f) After Kernel Task Power

Figure 8: Spectral properties of the NTK can account for bias gaps across widths. (a) Depth 3 MLPs in the lazy limit ($\gamma_0^{-1} = 200$) learning a quadratic polynomial from a uniform distribution on the sphere in $D = 5$ dimensions online. Wider networks perform better (dots). Even after ensembling (dashed), wider is better, and the ensembled curves match those of the averaged eNTK (solid). (b) The spectra of the averaged eNTK across widths do not show substantial variability. (c) However, at narrower width, more of the power of the task falls into higher spectral modes, consequently leading to a slowdown in training. These results hold across dimensions, batch sizes, task complexity, and architectures. Strong feature learning can reduce this effect. See Appendix E (d) We computed the ensemble averaged *after kernels* from the CIFAR-5M ResNet-18 models and computed the theoretical kernel flow on the task. Wider models have a slightly better mean kernel for this task. (e) The eigenvalues of the final NTKs are very consistent across widths. (f) The eigenfunction-target alignment of the final kernels noticeably differ across widths, evidenced by the cumulative power distribution $C(k)$ which accounts for the gap in theoretical loss curves under kernel flow.

given by the one-hot target labels $\Delta_c(\boldsymbol{x})|_{t=0} = y_c(\boldsymbol{x})$. The convergence rate of this dynamical system is again set by the eigenvalues and eigenfunction-task alignment. In Figure 8 (d), we find that the after kernels for wider networks give slightly more rapid convergence. Figures 8 (e) and (f) show that, similar to the MLP experiment, the spectra are very consistent across widths, but the eigenfunction task alignments, measured with $C(k)$ are not. Overall, these experiments suggest that an important aspect of the bias of finite width models compared to their infinite width analogs is the deformation of their eigenfunctions.

5 Conclusion

We have demonstrated a striking consistency across widths for many quantities of interest to deep learning practitioners. Our fine-grained studies go beyond simply comparing test losses and have demonstrated that learned network functions, internal representations, and dynamical large learning rate phenomena agree for sufficiently large widths on a variety of tasks across vision and language. At later training times, or after many repetitions of the dataset, we observe systematic deviations brought on by finite width, and have chracterized them in terms of the bias and variance of the network over initializations. These studies motivate the applicability of infinite-width feature-learning models in reasoning about large scale models trained on real-world data.

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[4] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[5] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[6] Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Limitations of the ntk for understanding generalization in deep learning. *arXiv preprint arXiv:2206.10012*, 2022.

[7] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[8] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[9] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31, 2018.

[10] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[11] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *arXiv preprint arXiv:2205.09653*, 2022.

[12] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[14] Greg Yang and Edward Hu. Maximal Update Parametrization (μP) and Hyperparameter Transfer (μTransfer). `https://github.com/microsoft/mup`, 2022.

[15] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations*, 2021.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[17] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[18] Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. Positional artefacts propagate through masked language model embeddings. *arXiv preprint arXiv:2011.04393*, 2020.

[19] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[20] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *Advances in Neural Information Processing Systems*, 33:15156–15172, 2020.

[21] Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What can linearized neural networks actually say about generalization? *Advances in Neural Information Processing Systems*, 34:8998–9010, 2021.

[22] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

[23] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2020.

[24] Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*. Cambridge University Press Cambridge, MA, USA, 2022.

[25] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[26] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[27] Alexander Atanasov, Blake Bordelon, Sabarish Sainathan, and Cengiz Pehlevan. The onset of variance-limited behavior for networks in the lazy and rich regimes. *arXiv preprint arXiv:2212.12147*, 2022.

[28] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

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

[30] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

[31] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[32] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

[33] Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

[34] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

[35] Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

[36] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *arXiv preprint arXiv:2304.03408*, 2023.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[40] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[41] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

[42] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[43] Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, 2023.

[44] Adam X Yang, Maxime Robeyns, Edward Milsom, Nandi Schoots, and Laurence Aitchison. A theory of representation learning in deep neural networks gives a deep generalisation of kernel methods. *arXiv preprint arXiv:2108.13097*, 2021.

[45] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[46] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[47] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[48] Philip M Long. Properties of the after kernel. *arXiv preprint arXiv:2105.10585*, 2021.

[49] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

[50] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

[51] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

[52] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

[53] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.