Thanks for your careful reading and positive feedback! We address major comments here (and the rest in the revision).

Further discussion/interpretation, and implications for ML (R1 and R2) The 9th content page allows us to greatly expand our discussion throughout the paper, and to add a conclusion section highlighting the following implications:

(a) An increasingly large body of literature studies generalization in random features regression models derived from the CK or NTK, and associated multiple-descent phenomena. In the linear-width regime, these results rely on asymptotic approximations for the Stieltjes transforms and resolvents of these kernels. Such studies have largely been limited to single-layer networks, and our results and techniques may enable their extension to deep networks with many layers.

(b) The linear-width asymptotic regime may provide a theoretically tractable setting for studying feature learning and "non-lazy" network training, and it is arguably closer to the operating regimes of neural networks in practical applications. Our experiments suggest an interesting possible mechanism of training in this regime, and our theoretical analysis of the spectra for random weights may provide a first step towards understanding this phenomenon.

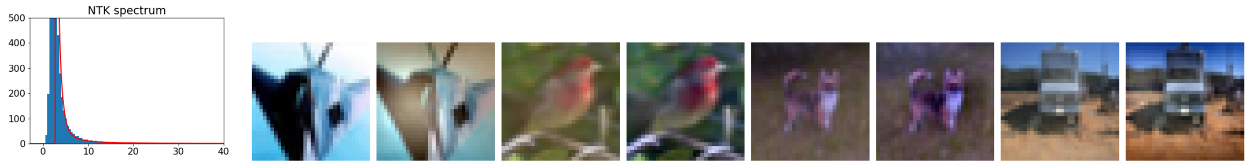
Assumption of pairwise orthogonality (R2 and R4) Thanks very much for these comments. As R2 points out, we believe this assumption is significantly more general than white noise. In the revision, we will add a discussion that the assumption encompasses many settings of independent samples with input dimension $d_0 \asymp n$, including:

(a) *Non-white* Gaussian inputs $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$, for any Σ satisfying $\operatorname{Tr}\Sigma = 1$ and $\|\Sigma\| \leq C/d_0$. Note that such data can have spectral distribution very different from $\mathbf{x}_\alpha$ with i.i.d. entries (which would be the Marcenko-Pastur law).

(b) More generally, inputs that may be expressed as $\mathbf{x}_\alpha = f(\mathbf{z}_\alpha)/\sqrt{d_0}$, where $\mathbf{z}_\alpha \in \mathbb{R}^m$ has independent entries satisfying a log-Sobolev inequality, and $f : \mathbb{R}^m \to \mathbb{R}^{d_0}$ is any Lipschitz function.

(c) Inputs $\mathbf{x}_\alpha$ drawn from certain multi-class Gaussian mixture models, satisfying the high-dimensional asymptotic assumptions of [CBG '16], [LLC '18], [LC '18]. The mixture components can differ in both mean and covariance.

Derivation of Theorem 3.7 from Lemma 3.5 (R3) We believe this derivation is the main point of theoretical novelty in our work, and is not standard. Each X_ℓ in $z_{-1}\,\mathrm{Id} + z_0 X_0^\top X_0 + \ldots + z_L X_L^\top X_L$ has a complicated dependence on $X_0, \ldots, X_{\ell-1}$, so this is not a classical RMT model. We develop the new idea of analyzing the extended matrix model $(z_{-1}\,\mathrm{Id} + z_0 X_0^\top X_0 + \ldots + z_L X_L^\top X_L)^{-1}(w_{-1}\,\mathrm{Id} + w_0 X_0^\top X_0 + \ldots + w_L X_L^\top X_L)$ in order to recursively characterize the spectrum by induction on depth. The resulting fixed-point equations are also non-standard, and led to new challenges in inductively showing uniqueness of their fixed points and providing a numerical algorithm for solving these equations.



Removal of 10 leading PCs (R2 and R4) This figure shows the NTK spectrum after mean-centering each CIFAR-10 class, rather than removing 10 PCs. The fit is OK but not perfect. Also shown are example images before (left) / after (right) removing the 10 PCs. Differences are hard to discern, and we will add a page of such images to the appendix.

Outliers and Adam optimizer (R2) We agree that the role of Adam is unclear, and we will make our code publicly available for further exploration. Training using full-batch gradient descent is slow—we tried based on R2's feedback, but had difficulty producing results with comparable generalization in a short time. In our Adam experiments, we tested various network depths, widths, and learning rates: Outlier eigenvalues emerged only in experiments that yielded good generalization, and not in those where the learned function generalized poorly. Also, these phenomena for the CK are perhaps more fundamental, and this then does not relate to the specific gradient flow derivation of the NTK.

NTK remains constant over training (R2) Our apologies for this confusion, and we will clarify in the revision: We do not claim the NTK remains approximately constant in this regime. The training dynamics described in Section 2.1 hold regardless of whether $K^{\mathrm{NTK}}(t)$ evolves or is fixed, and the eigenvalue $\lambda_\alpha(t)$ always determines the *instantaneous* decay of the training error along $\mathbf{v}_\alpha(t)$ at the instant t. "Training occurs most rapidly along the eigenvectors of the largest eigenvalues" is just an informal statement of this, with the understanding that the eigenvectors also evolve over training. We will also clarify in the intro that our theory pertains only to random weights and not to this evolution.

Miscellaneous Scaling by $1/\sqrt{d_\ell}$ is specifically important for the NTK as it affects the scaling of the derivative in deriving the NTK (R2). $b_\sigma = 0$ indeed has implications for classification and training, and we will add discussion and references to [CBG '16], [PW '17] (R2). The NTK spectrum here has two non-point-mass bulk components (R2). Convergence in Thm 3.4 holds marginally for each ℓ (R4). $d_0 \to \infty$ is necessary to ensure the approximate pairwise orthogonality, but is not otherwise used in the proof (R4). Thanks very much for the missing references! (R1, R2, R4)