1 Additional experiments

1.1 MLP-Mixer on CIFAR10

We trained the MLP-Mixer model from `https://arxiv.org/abs/2105.01601`, size `S/16`, on CIFAR10 to parallel the experiments from Figure 7 in the main text on a third architecture. We trained using SGD with momentum, MSE loss, batch size 128, and a cosine learning rate schedule with 1 epoch of linear warmup. The base learning rate was varied from 0.00625 to 1.6. We find similar trends to ResNet:

- $\mathcal{K}$ **stays in range** $[0.3, 1.0]$ **over a wide range of learning rates.** We vary learning rates by a factor of 128 and the typical $\mathcal{K}$ value only varies by a factor of 3 (Figure 1, top left). This suggests there is some effect stabilizing its growth.
- $\lambda_{\max}$ **remains far from the edge of stability.** Even for the largest learning rates, $\eta\lambda_{max} \approx 1$, far from the critical value of 2 (Figure 1, top right).
- $\mathcal{K}$ **close to** 1 **impedes training.** Larger learning rates spend more time with $\mathcal{K}$ close to 1, which leads to slower improvements in loss and error rate (Figure 1, bottom row).

1.2 ResNet50 and ViT on Imagenet - cross entropy loss

We conducted experiments to test the strengths and limitations of the analysis extending $\mathcal{K}$ to non-MSE loss (Section 2). We trained ResNet50 and ViT on Imagenet. The ViT implementation was the `S/16` size from `https://arxiv.org/abs/2010.11929`. Both models were trained using SGD with momentum, batch size 1024, on cross-entropy loss. We used a linear warmup for 5 epochs followed by cosine learning rate schedule for both models.

We used the analysis in Section 2 to compute an estimator of the noise kernel norm given by:

$$\hat{\mathcal{K}}_{mom} \equiv \frac{\eta}{2\alpha B}\text{tr}\left[\frac{1}{D}\mathbf{H}_{GN}\right] \tag{1}$$

where the Gauss-Newton component of the Hessian $\mathbf{H}_{GN} \equiv \mathbf{J}^{\text{T}}\mathbf{H}_{\mathbf{z}}\mathbf{J}$, where $\mathbf{H}_{\mathbf{z}}$ is the loss Hessian with respect to the logits. In order to compute the trace of $\mathbf{H}_{GN}$ efficiently over all of Imagenet, we used the Bartlett Gauss Newton estimator found in `https://arxiv.org/abs/2305.14342`. This let us estimate $\hat{\mathcal{K}}_{mom}$ with an epoch's worth of backwards passes. The results are found in Figure 2, with ResNet50 in the left column, and ViT in the right column.

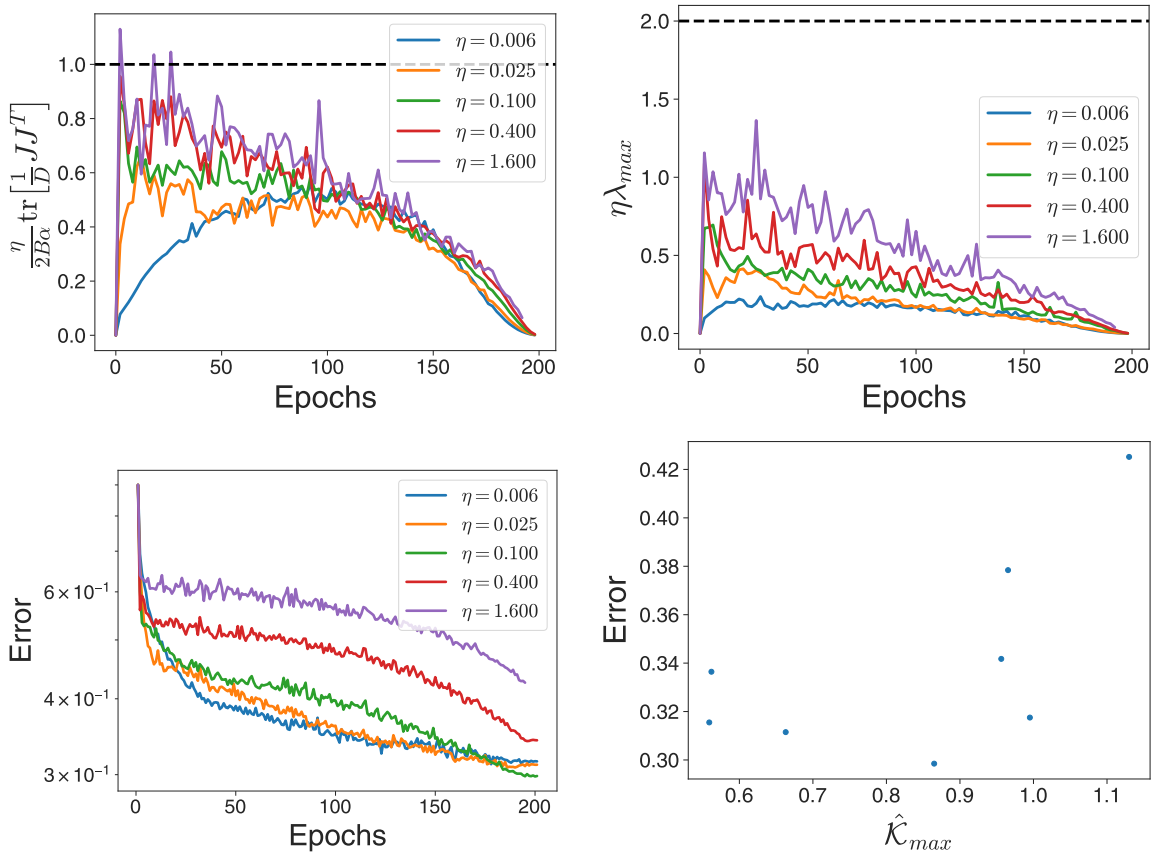

Figure 1: MLP-Mixer trained on CIFAR10. At large learning rates $\hat{\mathcal{K}}$ is near 1 at early times, and at intermediate times values cluster over a large range of learning rates (top left). Maximum eigenvalue remains below edge of stability (top right). Learning is slow when $\hat{\mathcal{K}}$ is near 1 (bottom left), and best performance is for intermediate values of $\hat{\mathcal{K}}$ (bottom right).

We found qualitative similarities with the experiments studying $\mathcal{K}$ in the MSE setting:

- $\mathcal{K}$ **remains in a small range over a wide range of learning rates.** Over a range of learning rates of factor 100, $\mathcal{K}$ only changed by a factor of ~ 5 (Figure 2, top row).
- **There is an** $O(1)$ **threshold of** $\mathcal{K}$ **corresponding to stable training.** The stability threshold was higher than $\mathcal{K} = 1$ in both examples. For ResNet50 it appears to be slightly below 2, for MLP-Mixer slightly above 2.
- $\mathcal{K}$ **is predictive of training success.** In both cases $\mathcal{K} < 0.5$ and $\mathcal{K} > 2.0$ lead to either inefficient or unstable training respectively (Figure 2, middle and bottom rows).

These experiments suggest that extending the analysis of the MSE case to cross-entropy

via the Gauss-Newton matrix is promising, but still requires work. In particular, a better estimator is needed to bring the stability threshhold to the predictable value $\mathcal{K} = 1$. We discuss some of the issues with the approximation in Section 2.2.

1.3 Averaged versions of Figure 2

We include a version of Figure 2 from the main text, averaging over 100 different instantiations of SGD noise (Figure 3). This shows that the deviations for small batch size are in fact real, and not merely artifacts of the more noisy trajectories.

2 Beyond MSE loss

Here we consider the stability of SGD under more general convex losses. We will derive a stability condition by expanding around a minimum. The upshot is that under certain assumptions, we can derive a noise kernel norm $\mathcal{K}$ for non-MSE losses, and there is a regime where we have the estimator

$$\mathcal{K} \approx \hat{\mathcal{K}}_{tr} \equiv \frac{\eta}{2B} \mathrm{tr}\left(\frac{1}{D}\mathbf{J}^{\mathrm{T}}\mathbf{H}_{\mathbf{z}}^{*}\mathbf{J}\right) \tag{2}$$

where $\mathbf{H}_{\mathbf{z}}^{*}$ is the Hessian of the loss with respect to the logits at the minimum. We note that $\mathbf{J}^{\mathrm{T}}\mathbf{H}_{\mathbf{z}}^{*}\mathbf{J}$ is the Gauss-Newton part of the Hessian.

2.1 Theory: expansion around a fixed point

Consider a linear model $\mathbf{z}_t = \mathbf{J}\boldsymbol{\theta}_t$. Here θ_t is the P-dimensional parameter vector, and $\mathbf{z}_t$ is the output. If each data point has C outputs, then we flatten them so that $\mathbf{z}_t$ has dimension CD. $\mathbf{J}$ is the (flattened) Jacobian with dimension $CD \times P$.

Consider the loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{D}\sum_{\alpha=1}^{D} \mathcal{L}_z(\mathbf{z}_\alpha(\theta_t)) \tag{3}$$

Here $\mathcal{L}_z$ is the per-example loss, convex in the inputs. The update equation for $\boldsymbol{\theta}$ under SGD with batch size B is

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\frac{\eta}{B}\mathbf{J}^{\mathrm{T}}\mathbf{P}_t\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}_t) \tag{4}$$

where $\mathbf{P}_t$ is the projection matrix with exactly B 1s on the diagonal, drawn i.i.d. at each step. The update equation for $\mathbf{z}_t$ is

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\frac{\eta}{B}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{P}_t\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}_t) \tag{5}$$

In general this is a non-linear stochastic system in $\mathbf{z}_t$, whose moments don't close at any finite order. However, we can make progress by expanding around a minimum. Let $\mathbf{z}^*$ be a minimum of the loss. We have:

$$\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}) = \mathbf{H}_{\mathbf{z}}(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) + O(||\mathbf{z} - \mathbf{z}^*||^2) \tag{6}$$

where $\mathbf{H}_{\mathbf{z}}$ is the PSD Hessian of $\mathcal{L}$ with respect to the logits $\mathbf{z}$. Therefore near $\mathbf{z}^*$ we can write:

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\frac{\eta}{B}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{P}_t\mathbf{H}_{\mathbf{z}}(\mathbf{z}^*)(\mathbf{z}_t - \mathbf{z}^*) + O(||\mathbf{z} - \mathbf{z}^*||^2) \tag{7}$$

Let $\tilde{\mathbf{z}} \equiv \mathbf{z} - \mathbf{z}^*$. Neglecting terms of $O(||\tilde{\mathbf{z}}^*||^2)$ we have:

$$\tilde{\mathbf{z}}_{t+1} - \tilde{\mathbf{z}}_t = -\frac{\eta}{B}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{P}_t\mathbf{H}^*_{\mathbf{z}}\tilde{\mathbf{z}}_t \tag{8}$$

where we denote $\mathbf{H}^*_{\mathbf{z}} \equiv \mathbf{H}_{\mathbf{z}}(\mathbf{z}^*)$.

This is similar to the dynamical equation for the MSE case, but with an additional PSD matrix factor. The second moment equations are:

$$\begin{aligned}\mathrm{E}_{\mathbf{P}}[\tilde{\mathbf{z}}_{t+1}\tilde{\mathbf{z}}_{t+1}^\top - \tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^\top|\tilde{\mathbf{z}}_t] = &-\frac{\eta}{D}(\mathbf{J}\mathbf{J}^\top\mathbf{H}^*_{\mathbf{z}}\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^\top + \tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^\top\mathbf{H}^*_{\mathbf{z}}\mathbf{J}\mathbf{J}^\top) + \frac{\eta^2}{D^2}\mathbf{J}\mathbf{J}^\top\mathbf{H}^*_{\mathbf{z}}\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^\top\mathbf{H}^*_{\mathbf{z}}\mathbf{J}\mathbf{J}^\top \\ &+ (\beta^{-1} - 1)\frac{\eta^2}{D^2}\mathbf{J}\mathbf{J}^\top\mathbf{H}^*_{\mathbf{z}}\mathrm{diag}\left[\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^\top\right]\mathbf{H}^*_{\mathbf{z}}\mathbf{J}\mathbf{J}^\top\end{aligned} \tag{9}$$

Using the PSDness of $\mathbf{H}^*_{\mathbf{z}}$, we can define the modified covariance matrix $\tilde{\mathbf{\Sigma}}_t \equiv \mathbf{H}^{1/2}_{\mathbf{z}}\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t^{\mathrm{T}}\mathbf{H}^{1/2}_{\mathbf{z}}$. The dynamics are given by:

$$\mathrm{E}_{\mathbf{P}}[\tilde{\mathbf{\Sigma}}_{t+1} - \tilde{\mathbf{\Sigma}}_t|\tilde{\mathbf{\Sigma}}_t] = -\eta(\tilde{\mathbf{\Theta}}\tilde{\mathbf{\Sigma}}_t + \tilde{\mathbf{\Sigma}}_t\tilde{\mathbf{\Theta}}) + \eta^2(\tilde{\mathbf{\Theta}}\tilde{\mathbf{\Sigma}}_t\tilde{\mathbf{\Theta}} + (\beta^{-1} - 1)\tilde{\mathbf{\Theta}}(\mathbf{H}^*_{\mathbf{z}})^{1/2}\mathrm{diag}\left[(\mathbf{H}^*_{\mathbf{z}})^{-1/2}\tilde{\mathbf{\Sigma}}_t(\mathbf{H}^*_{\mathbf{z}})^{-1/2}\right](\mathbf{H}^*_{\mathbf{z}})^{1/2}\tilde{\mathbf{\Theta}}) \tag{10}$$

where we define $\tilde{\mathbf{\Theta}} \equiv \frac{1}{D}(\mathbf{H}^*_{\mathbf{z}})^{1/2}\mathbf{J}\mathbf{J}^{\mathrm{T}}(\mathbf{H}^*_{\mathbf{z}})^{1/2}$. Note that $\tilde{\mathbf{\Theta}}$ is the Gram matrix of the Gauss-Newton part of the Hessian, up to a normalizing constant - they have the same non-zero eigenvalues.

We can once again attempt to work in a diagonal basis to reduce the complexity of the analysis. Consider the eigendecomposition $\tilde{\mathbf{\Theta}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$. If $\tilde{\mathbf{S}} \equiv \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{\Sigma}}\mathbf{V}$, we have:

$$\begin{aligned}\mathrm{E}_{\mathbf{P}}[\tilde{\mathbf{S}}_{t+1} - \tilde{\mathbf{S}}_t|\tilde{\mathbf{S}}_t] = &-\eta(\mathbf{\Lambda}\tilde{\mathbf{S}}_t + \tilde{\mathbf{S}}_t\mathbf{\Lambda}) + \eta^2(\mathbf{\Lambda}\tilde{\mathbf{S}}_t\mathbf{\Lambda} + \\ &(\beta^{-1} - 1)\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}(\mathbf{H}^*_{\mathbf{z}})^{1/2}\mathrm{diag}\left[(\mathbf{H}^*_{\mathbf{z}})^{-1/2}\mathbf{V}\tilde{\mathbf{S}}_t\mathbf{V}^{\mathrm{T}}(\mathbf{H}^*_{\mathbf{z}})^{-1/2}\right](\mathbf{H}^*_{\mathbf{z}})^{1/2}\mathbf{V}\mathbf{\Lambda})\end{aligned} \tag{11}$$

This equation defines a linear operator $\tilde{\mathbf{T}}$ whose maximum eigenvalue defines stability. We have

$$\mathrm{E}_{\mathbf{P}}[(\tilde{\mathbf{S}}_{t+1})_{\mu\nu}|\tilde{\mathbf{S}}_t] = \sum_{\beta\gamma} \tilde{\mathbf{T}}_{\mu\nu,\beta\gamma}(\tilde{\mathbf{S}}_t)_{\beta\gamma} \tag{12}$$

where $\tilde{\mathbf{T}}$ is given by

$$\begin{aligned}\tilde{\mathbf{T}}_{\mu\nu,\beta\gamma} =& \delta_{\mu\beta,\nu\gamma}(1-\eta(\lambda_\mu+\lambda_\nu)+\eta^2\lambda_\mu\lambda_\nu) \\ &+ (\beta^{-1}-1)\eta^2\lambda_\mu\lambda_\nu\left[\sum_{\alpha,\delta,\epsilon,\phi,\psi} \mathbf{V}_{\phi\mu}(\mathbf{H}^*_{\mathbf{z}})^{1/2}_{\alpha\phi}(\mathbf{H}^*_{\mathbf{z}})^{-1/2}_{\delta\alpha}\mathbf{V}_{\delta\beta}\mathbf{V}_{\epsilon\gamma}(\mathbf{H}^*_{\mathbf{z}})^{-1/2}_{\alpha\epsilon}(\mathbf{H}^*_{\mathbf{z}})^{1/2}_{\alpha\psi}\mathbf{V}_{\psi\nu}\right]\end{aligned} \tag{13}$$

where λ_μ is the μth eigenvalue from $\mathbf{\Lambda}$. If we reduce the $(DC)^2 \times (DC)^2$ system by restricting to the diagonal $\mathbf{p} = \mathrm{diag}(\tilde{\mathbf{S}})$

$$(\mathbf{p}_{t+1})_\mu = \sum_\beta \tilde{\mathbf{T}}_{\mu\mu,\beta\beta}(\mathbf{p}_t)_\beta \tag{14}$$

which becomes, in matrix notation

$$\mathbf{p}_{t+1} = \mathbf{D}\mathbf{p}_t,\ \mathbf{D} \equiv [(\mathbf{I}-\eta\mathbf{\Lambda})^2 + \eta^2(\beta^{-1}-1)\mathbf{\Lambda}^2\tilde{\mathbf{C}}] \tag{15}$$

with

$$\tilde{\mathbf{C}}_{\beta\mu} \equiv \sum_{\alpha,\delta,\phi}[\mathbf{V}_{\phi\mu}(\mathbf{H}^*_{\mathbf{z}})^{1/2}_{\alpha\phi}]^2[(\mathbf{H}^*_{\mathbf{z}})^{-1/2}_{\delta\alpha}\mathbf{V}_{\delta\beta}]^2 \tag{16}$$

Note: $\mathbf{H}^*_{\mathbf{z}}$ is block-diagonal with respect to the $C \times C$ blocks for the D datapoints. If $\mathbf{H}^*_{\mathbf{z}}$ is diagonal within each block (no logit-logit interactions), then $\tilde{\mathbf{C}} = \mathbf{C}$ from the MSE case. Otherwise, $\tilde{\mathbf{C}}$ is a slightly different positive matrix.

This means that we can derive a noise kernel norm $\mathcal{K}$ following the analysis in Section 3.2 of the main text, using $\mathbf{A} = (\mathbf{I}-\eta\mathbf{\Lambda})^2$, $\mathbf{B} = \eta^2(\beta^{-1}-1)\mathbf{\Lambda}\tilde{\mathbf{C}}\mathbf{\Lambda}$.

2.2 Relationship to previous analysis

This analysis is analogous to the MSE case, with the modified NTK $\tilde{\mathbf{\Theta}}$ taking the role of the NTK - meaning the Gauss-Newton eigenvalues are key. If $\tilde{\mathbf{C}} \approx \frac{1}{D}\mathbf{1}\mathbf{1}^{\mathrm{T}}$, then we recover the estimators from Section 3.3, replacing $\hat{\Theta}$ with $\tilde{\mathbf{\Theta}}$ - or alternatively,

$$\mathcal{K} \approx \hat{\mathcal{K}}_{tr} \equiv \frac{\eta}{2B}\mathrm{tr}\left(\tilde{\mathbf{\Theta}}\right) = \frac{\eta}{2B}\mathrm{tr}\left(\frac{1}{D}\mathbf{J}^{\mathrm{T}}\mathbf{H}^*_{\mathbf{z}}\mathbf{J}\right) \tag{17}$$

The last expression is written in terms of the Gauss-Newton matrix at the minimum.

However, there are a few ways this quantity may suffer compare to the MSE one:

- **Expansion around $\mathbf{z}^*$.** In order to derive a linear recurrence relation, we expanded around the minimum $\mathbf{z}^*$. If the dynamics is near but not at a minimum, an accurate computation would require finding $\mathbf{z}^*$, and computing $\tilde{\mathbf{\Theta}}$ there. If the dynamics is not near a minimum, then the accuracy of the stability condition is unclear.

- **Restriction of $\tilde{\mathbf{T}}$ to the diagonal.** In order to derive $\mathcal{K}$ we reduce to the dynamics of the diagonal of the covariance only. For MSE loss previous work has justified this approximation in certain high dimensional limits; for more general loss functions this is not clear.

- **Nontrival structure of $\mathbf{H}^*_{\mathbf{z}}$.** In order to use efficient high-dimensional approximators of $\mathcal{K}$, it is useful for $\mathbf{C}$ to have a low-rank structure. In the MSE case this can be a good approximation because eigenvectors are delocalized in the coordinate basis; in the more general setting, this may no longer be the case. For example, cross-entropy could introduce additional correlations across members of the same class, different inputs, or the same inputs, different classes.

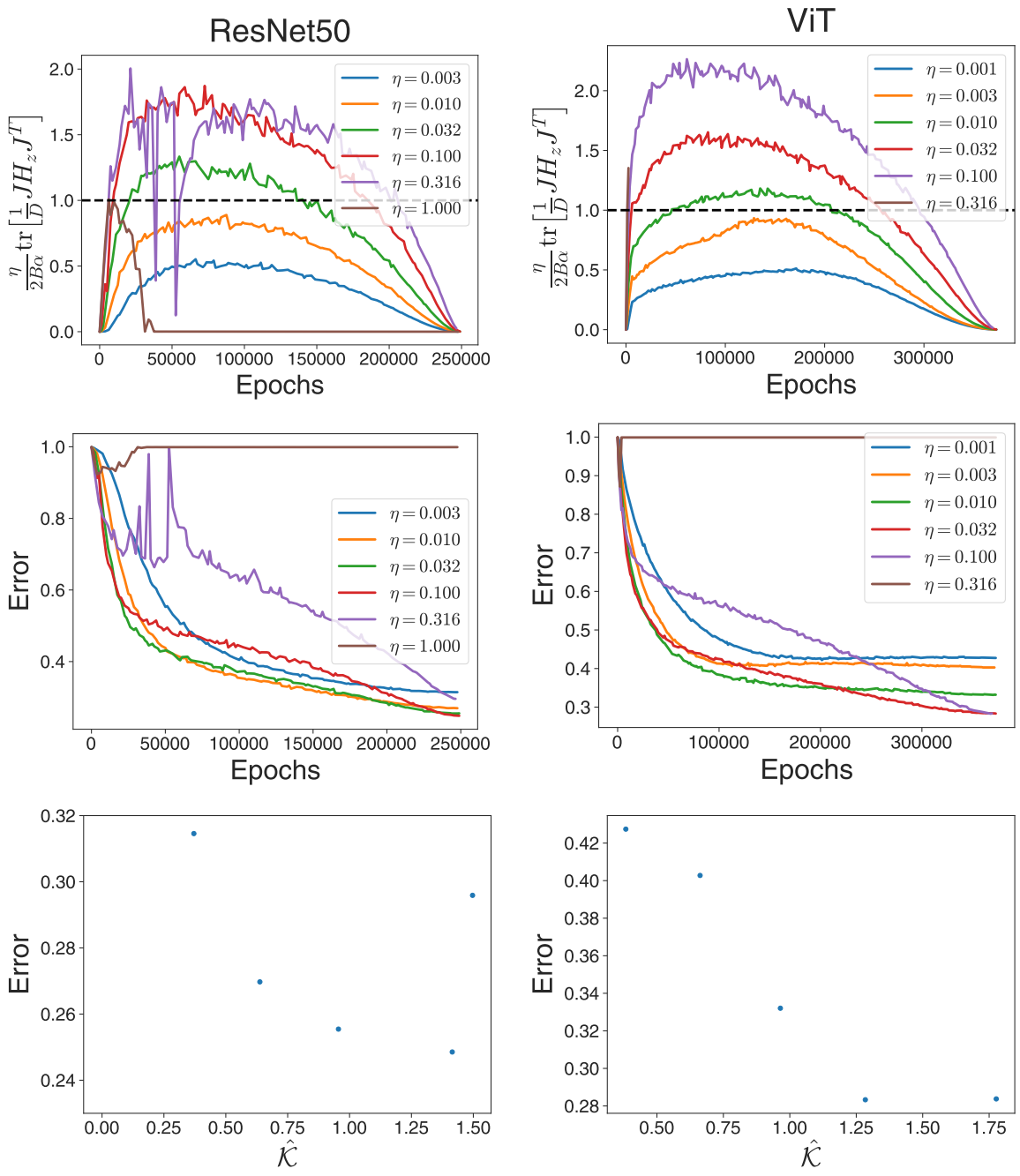

Figure 2: ResNet50 (left column) and ViT (right column) trained on Imagenet with cross-entropy loss. $\mathcal{K}$ was approximated using the Gauss-Newton trace, estimated using the Bartlett-Gauss-Newton estimator. Learning rate variation of 1000 leads to $\hat{\mathcal{K}}$ variation of a factor of ~ 5. $\hat{\mathcal{K}}$ seems to have a critical value around 2 (top and middle row). There appears to be an $O(1)$ value of $\hat{\mathcal{K}}$ predictive of low error (bottom row), but more work is needed to refine the measurement.

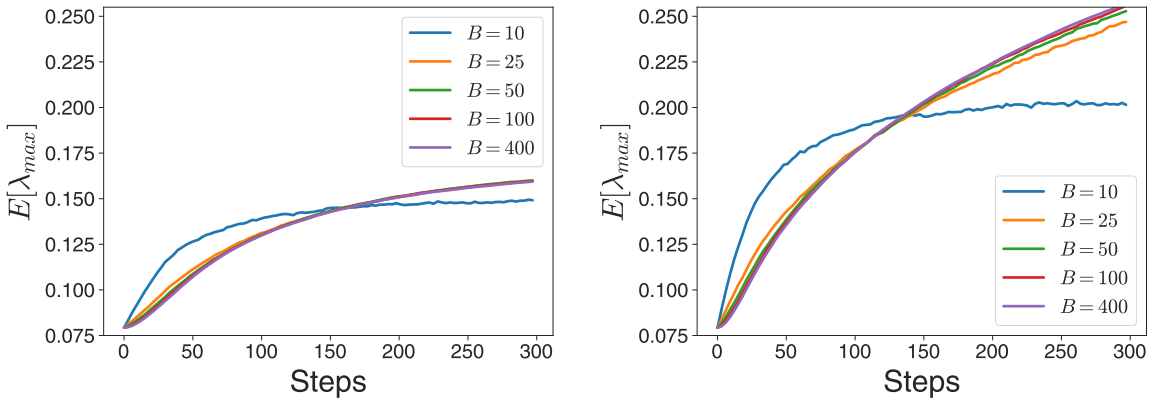

Figure 3: Maximum NTK eigenvalue from quadratic regression model, averaged over 100 random seeds. Corresponds to experiments from Figure 2 in the main text.