Closed form of the Hessian Spectrum

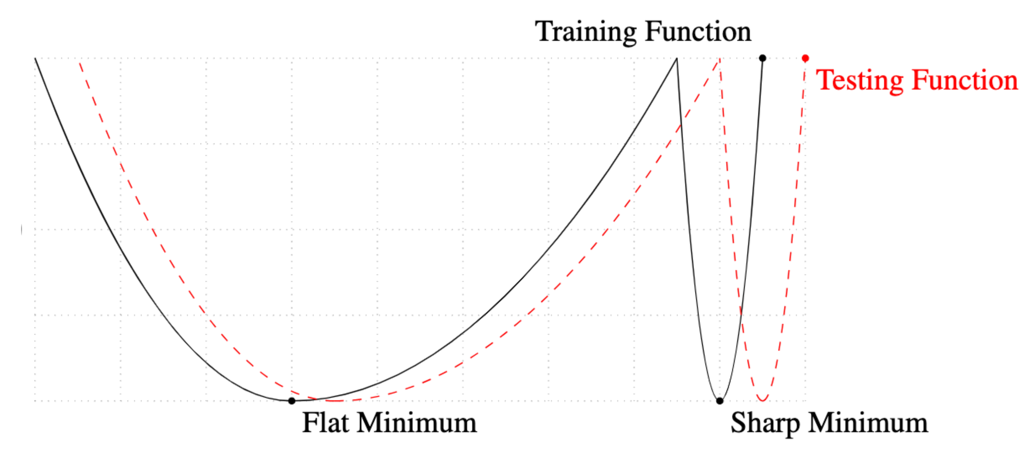
for some Neural Networks

Sidak Pal Singh & Thomas Hofmann

ETH zürich

The Hessian matrix is of fundamental significance

- Wide body of research shows that 'flatter' minima generalize better
- Algorithms like Sharpness-Aware Minimisation (SAM) work quite well
- Learning seems to happen at the Edge-of-Stability (EoS), $\eta \approx 2/\lambda_{\max}(\mathbf{H})$



But, how are the eigenvalues/eigenvectors really like?

What does 'sharpness' even mean?

Insights from a popular toy-model

Setup: 1 hidden-layer univariate network (linear/ReLU)

$$f(x) = \langle \mathbf{w} \, , \sigma(\mathbf{v}\, x) \rangle$$

$$\mathbf{w}, \mathbf{v} \in \mathbb{R}^m$$

- Valid for arbitrary number of datapoints and any layer-width; MSE loss

Key Result :
(Linear case)

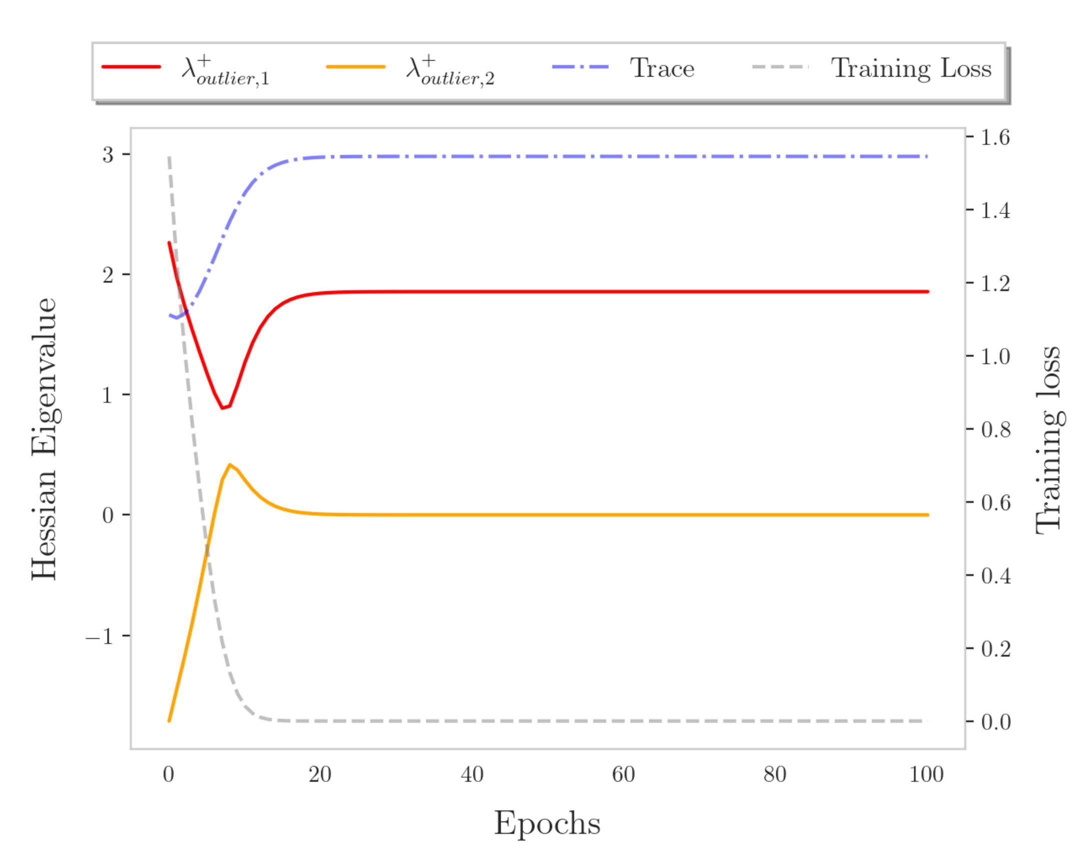
The above network with $2m$ parameters has an eigenspectrum consisting of $m-1$ repeated eigenvalues $\lambda_{\text{bulk}} = \pm\, \overline{x\delta}$ and an outlying eigenvalue pair given by

$$\lambda_{\text{outlier}} = \frac{1}{2}\left(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2\right) \pm \frac{1}{2}\sqrt{\left(\sigma^2\|\mathbf{w}\|^2 - \sigma^2\|\mathbf{v}\|^2\right)^2 + 4\sigma^4\left(\|\mathbf{w}\|^2\,\|\mathbf{v}\|^2 - \langle \mathbf{w}, \mathbf{v}\rangle^2\right) + 4\left(2\langle \mathbf{w}, \mathbf{v}\rangle - \overline{yx}\right)^2}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is the (uncentered) input variance and $\overline{\delta x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,\delta_i$, with $\delta_i = \langle \mathbf{w}, \mathbf{v}\rangle\, x_i - y_i$ is the residual-input covariance

Insights about the Eigenspectrum:

- Outlier eigenvalues exists as pairs; only one remains at convergence
- Sharpness quantifies *discrepancy b/w layer norms, co-linearity of parameters, extent of target captured,* besides overall parameter norm
- ReLU leads to cell-wise decomposition, but each cell like linear case

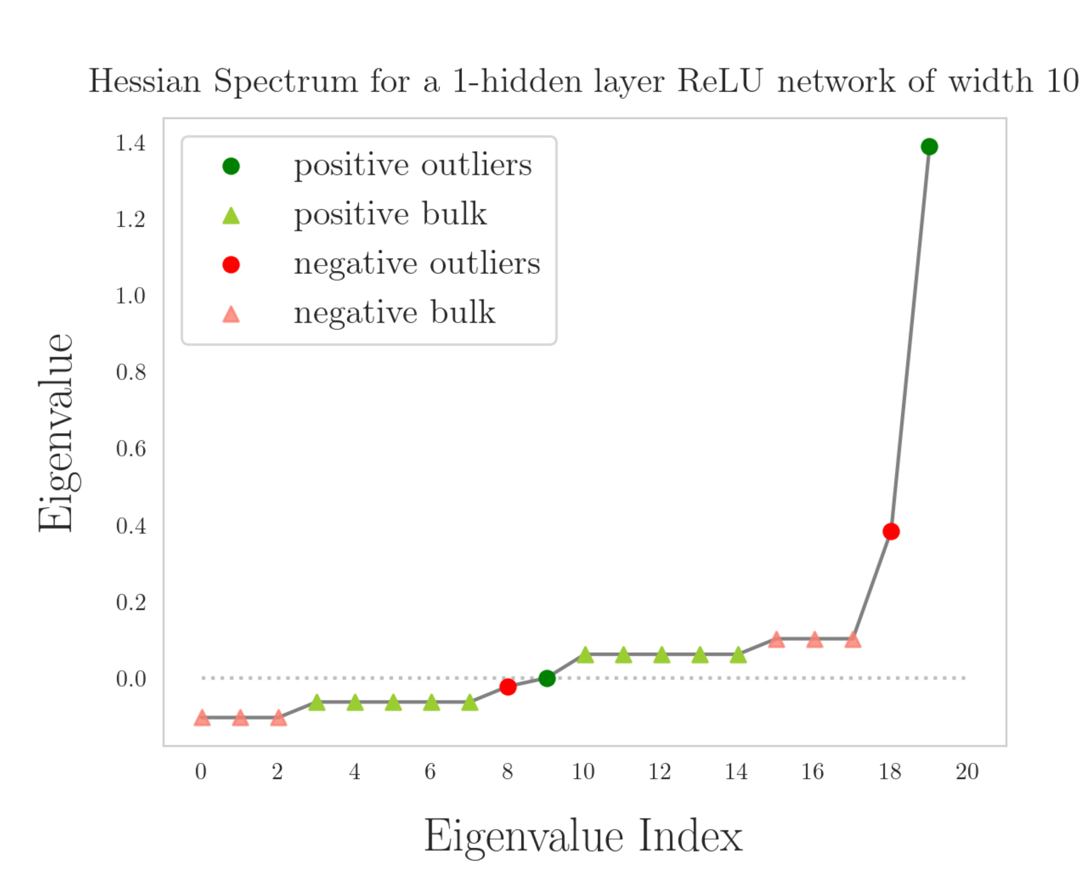


Evolution of Outlier Eigenvalue Pair

Key Result:
(ReLU case)

The Hessian undergoes a cell-wise decomposition, which here is fully decoupled:

$$\mathbf{H}_{\mathrm{L}} = \begin{pmatrix} \frac{n_+}{n}\mathbf{H}_{\mathrm{L}}^{+} & \mathbf{0} \\ \mathbf{0} & \frac{n_-}{n}\mathbf{H}_{\mathrm{L}}^{-} \end{pmatrix}$$



Bulk vs Outlier spectrum: ReLU

Eigenvector Structure:

$$\mathbf{z}_{\text{outlier}_i} = \begin{pmatrix} \lambda_{\text{outlier}_i}\,\mathbf{w} + \overline{x\delta}\,\mathbf{v} \\ \overline{x\delta}\,\mathbf{w} + \lambda_{\text{outlier}_i}\,\mathbf{v} \end{pmatrix}$$

Here, outlier eigenvectors are Linear combination of parameter and gradient vectors