Closed form of the Hessian spectrum for some Neural Networks

Sidak Pal Singh SSIDAK@ETHZ.CH
ETH Zürich, Switzerland

Thomas Hofmann THOMAS.HOFMANN@INF.ETHZ.CH
ETH Zürich, Switzerland

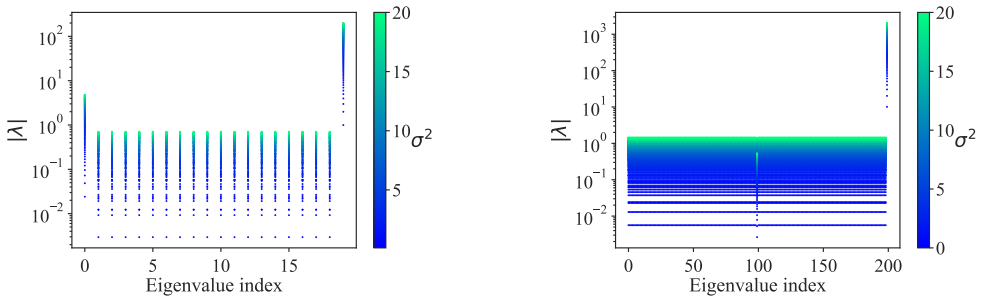
Abstract

The Hessian matrix and its spectrum hold significant theoretical and practical relevance as they capture the pairwise interaction of the parameters, and as a result have been widely used in building preconditioned optimizers, measuring generalization performance, studying the effect of learning rate and other hyperparameters, optimally pruning parameters, and more. Given its versatility and importance, several prior works have tried to characterize the Hessian spectrum through its spectral density, rank, description of the outlier and bulk in its spectrum, while often resorting to random matrix theory based approximations. However, grasping how the top eigenvalue precisely behaves has remained unclear, let alone the corresponding eigenvectors, due to a lack of its closed form for any non-trivial class of neural networks. Likewise, given the acute costs required to empirically estimate the Hessian or its various spectral measures (such as the top eigenvalue, trace, and determinant), our understanding of their behaviour continues to be somewhat muddled. In this work, we derive a closed form of all the eigenvalues and their corresponding eigenvectors for one-hidden layer, linear as well as ReLU, uni-dimensional networks with arbitrary hidden-layer width and for the loss aggregated over any number of samples. As a consequence of these theoretical results, we shed light on the previously undiscovered 'paired' nature of the spectrum outlier eigenvalues, the grouped composition of the trace, and a cell-wise decomposition of the Hessian spectrum with ReLU.

1. Setup & Results

Assume we have a one-hidden layer scalar linear network $f : \mathbb{R} \mapsto \mathbb{R}$, namely, $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{v} \rangle\, x$, with the parameters $\mathbf{w}, \mathbf{v} \in \mathbb{R}^m$. Although this is a simplified setting, it has been put to significant use by past works such as the occurrence of the catapults in loss with large learning rates [5] as well as understanding the edge-of-stability behaviour [7]. Moving further, let us consider mean-squared error (MSE) loss $\ell(\mathbf{w}, \mathbf{v}) = \frac{1}{2n} \sum_{i=1}^{n} (\langle \mathbf{w}, \mathbf{v} \rangle \cdot x_i - y_i)^2$ computed over a set of n data points $\{(x_i, y_i)\}_{i=1}^{n}$. Further, let us define the shorthands for the (uncentered) input standard deviation as $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$, the (uncentered) input-output covariance as $\overline{yx} = \frac{1}{n} \sum_{i=1}^{n} y_i x_i$, and the (uncentered) residual-input covariance as $\overline{\delta x} = \frac{1}{n} \sum_{i=1}^{n} x_i\, \delta_i$, where the residual $\delta_i = \langle \mathbf{w}, \mathbf{v} \rangle\, x_i - y_i$ denotes how far off the network is on fitting the datapoint (x_i, y_i). Then for this network, the Hessian spectrum has the following closed form:

Theorem 1 *For the setting of a one-hidden layer scalar linear network $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{v} \rangle\, x$ with $2m$ parameters and as detailed above, the Hessian $\mathbf{H}_{\mathrm{L}}$ spectrum consists of $m-1$ repeated eigenvalues*



(*a*) Width 10 (*b*) Width 100

Figure 1: Impact of data scaling on the Hessian spectrum for two networks of different width. The considered task is that of a random Gaussian regression setting.

$\lambda_{bulk} = \pm\overline{\delta x}$ and two paired outlying eigenvalues defined by the following expression:

$$\lambda_{outlier_{1,2}} = \frac{1}{2}(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2) \pm \frac{1}{2}\sqrt{(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2)^2 + 4(\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle)} \quad (1)$$

Firstly, we see that the eigenvalues scale in proportion to the input variance σ^2. A common starting element of machine learning pipelines is to normalize[1] the data, and as shown in the Figure 1, we realize how that can significantly sharpen the Hessian spectrum. Next, by noting that the trace of the Hessian in this setting is given by $\text{Tr}(\mathbf{H}_\text{L}) = \sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2$, the above expression of the outlying eigenvalues can be reformulated as,

$$\lambda_{\text{outlier}_{1,2}} = \frac{1}{2}\text{Tr}(\mathbf{H}_\text{L}) \pm \sqrt{\frac{1}{4}\text{Tr}^2(\mathbf{H}_\text{L}) - \lambda_{\text{outlier}_1}\lambda_{\text{outlier}_2}}\,. \quad (2)$$

Further, the expression $-\lambda_{\text{outlier}_1}\lambda_{\text{outlier}_2} = (\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle)$ inside the expression of the outlying eigenvalues can be rewritten as, $(\langle\mathbf{w},\mathbf{v}\rangle\sigma^2 - \overline{yx})(3\langle\mathbf{w},\mathbf{v}\rangle\sigma^2 - \overline{yx})$ which is negative if $\langle\mathbf{w},\mathbf{v}\rangle \leq \overline{yx}/\sigma^2$ and $\langle\mathbf{w},\mathbf{v}\rangle \geq \overline{yx}/3\sigma^2$ and non-negative otherwise. Note, $\overline{yx}/\sigma^2 = \mathbf{E}[yx]/\mathbf{E}[x^2] =: \theta^\star$ is precisely the closed-form solution for the scalar parameter θ in the linear regression $\mathbf{y} = \theta\,\mathbf{x}$. Effectively, the linear network under consideration is nothing but a parameterization of this scalar in the form of an inner-product between the layer weights $\mathbf{w}, \mathbf{v}$.

So, we categorize the phases of learning in three kinds: 'early phase': $\langle\mathbf{w},\mathbf{v}\rangle \leq \theta^\star/3$, 'late phase': $\theta^\star/3 \leq \langle\mathbf{w},\mathbf{v}\rangle \leq \theta^\star$ and 'divergent phase': $\theta^\star \leq \langle\mathbf{w},\mathbf{v}\rangle$. In early and divergent phases, $\lambda_{\text{outlier}_1} \geq \text{Tr}(\mathbf{H}_\text{L})$ and $\lambda_{\text{outlier}_2} \leq 0$. In the late phase, we have that $\lambda_{\text{outlier}_1} \leq \text{Tr}(\mathbf{H}_\text{L})$ and $\lambda_{\text{outlier}_2} \geq 0$. Upon reaching the solution, the Hessian is just rank one, with its spectrum given by $\lambda_{\text{outlier}_1} = \text{Tr}(\mathbf{H}_\text{L}) = \sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2$, and with the rest of the eigenvalues being 0.

An interpretation of the top Hessian eigenvalue, or sharpness. We can also express it as,

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

1. Ghorbani et al. [3] have noted how batch normalization (BN) can suppress outlier eigenvalues; but given how BN would function similar to data normalization albeit within the network, we suspect BN has a wider impact at the Hessian spectrum and not just the outlier eigenvalues.

And so, inherently the *semantics of sharpness lie in a net quantification of the discrepancy between the layer parameter norms, making the parameters co-linear, and capturing part of the target, alongside (the somewhat expected) overall parameter norm.*

The structure of eigenvectors. Moving ahead, let us shift our focus on deriving a closed form and understanding the structure of the eigenvectors, ordered as per $\boldsymbol{\theta} = (\mathbf{v}^\top, \mathbf{w}^\top)^\top$.

Theorem 2 *For the above setting, the Hessian eigenvectors corresponding to the outlying eigenvalues, determined up to scaling and sign, take the form, for $i \in \{1, 2\}$, given below:*

$$\mathbf{z}_{\textit{outlier}_i} = \begin{pmatrix} \lambda_{\textit{outlier}_i}\, \mathbf{w} + \overline{\delta x}\, \mathbf{v} \\ \overline{\delta x}\, \mathbf{w} + \lambda_{\textit{outlier}_i}\, \mathbf{v} \end{pmatrix} = \lambda_{\textit{outlier}_i} \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} + \overline{\delta x} \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} \tag{3}$$

Thus, the outlier eigenvectors live in a two-dimensional space spanned by the vectors $\left\{ \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} \right\}$ and, in fact, form its orthonormal basis. When the gradient of the loss is non-zero, we can further express these eigenvectors as, $\mathbf{z}_{\text{outlier}_i} = \frac{\lambda_{\text{outlier}_i}}{\overline{\delta x}} \nabla_{\boldsymbol{\theta}} \ell + \overline{\delta x}\, \boldsymbol{\theta}$.

In contrast, the eigenvectors corresponding to the bulk eigenvalues live in a $2m - 2$ dimensional subspace, which is essentially determined by the orthogonal complements of the vectors $\mathbf{w}+\mathbf{v}$, $\mathbf{w}-\mathbf{v}$. More explicitly, we have that:

Theorem 3 *For the above setting, the Hessian eigenvectors corresponding to the bulk eigenvalues have the form, determined up to scaling and sign, $\mathbf{z}_{\textit{bulk}} = \begin{pmatrix} \hat{\mathbf{z}}_{\textit{bulk}}^\top & \textit{sgn}(\lambda_{\textit{bulk}})\, \hat{\mathbf{z}}_{\textit{bulk}}^\top \end{pmatrix}^\top$ with*

$$\hat{\mathbf{z}}_{\textit{bulk}} = \left(\mathbf{I} - \frac{(\mathbf{w} + \textit{sgn}(\lambda_{\textit{bulk}})\, \mathbf{v})(\mathbf{w} + \textit{sgn}(\lambda_{\textit{bulk}})\, \mathbf{v})^\top}{\|\mathbf{w} + \textit{sgn}(\lambda_{\textit{bulk}})\, \mathbf{v}\|^2} \right) \mathbf{c}\,, \qquad \textit{for some vector } \mathbf{c}$$

As it turns out, the eigenvector proofs provide another way to derive the closed form of eigenvalues. Besides, the above expression of the eigenvectors also suggests that they would change smoothly over the course of training, unless there are rapid changes in the subspace spanned by the parameters.

2. Extension to the ReLU case

Assume the network is now, $f(x) = \mathbf{w}^\top (\mathbf{v} \cdot x)_+$, where, $(a)_+ = \mathbf{1}\{a > 0\}$ is the ReLU non-linearity and is applied elementwise. For this particular network, a hidden neuron j 'fires' if $v_j x > 0$, i.e., when either $v_j > 0$, $x > 0$ and $v_j < 0$, $x < 0$. Hence, this parameter space partitions in tandem with the partitions of the input space, the latter will be referred to as cells, which are namely, the $+$ cell where $x > 0$ and the $-$ cell where $x \leq 0$.

Notice, we can write $\mathbf{v} = \mathbf{v} \odot \mathbb{1}\{\mathbf{v} > 0\} + \mathbf{v} \odot \mathbb{1}\{\mathbf{v} \leq 0\} =: \mathbf{v}_+ + \mathbf{v}_-$, where $\odot$ denotes the Hadamard product and $\mathbb{1}\{\mathbf{a} > 0\}_j = \mathbb{1}\{a_j > 0\}$. Likewise, we can associate the $\mathbf{w}$ parameters to these two cells as, $\mathbf{w} = \mathbf{w}_+ + \mathbf{w}_-$, with $\mathbf{w}_+ = \mathbf{w} \odot \mathbb{1}\{\mathbf{v} > 0\}$ and $\mathbf{w}_- = \mathbf{w} \odot \mathbb{1}\{\mathbf{v} \leq 0\}$. Thus, we can express the network function in the following manner,

$$f(x) = \langle \mathbf{w}_+, \mathbf{v}_+ \rangle x\, \mathbb{1}\{x > 0\} + \langle \mathbf{w}_-, \mathbf{v}_- \rangle x\, \mathbb{1}\{x \leq 0\}$$

We specialize the previously defined shorthands for various data-dependent quantities to each of the two cells. In particular, let us define the (uncentered) standard deviation of the positive and negative datapoints as, $\sigma_+ = \sqrt{\frac{1}{n_+} \sum_{i=1}^n x_i^2\, \mathbb{1}\{x_i > 0\}}$ and $\sigma_- = \sqrt{\frac{1}{n_-} \sum_{i=1}^n x_i^2\, \mathbb{1}\{x_i < 0\}}$

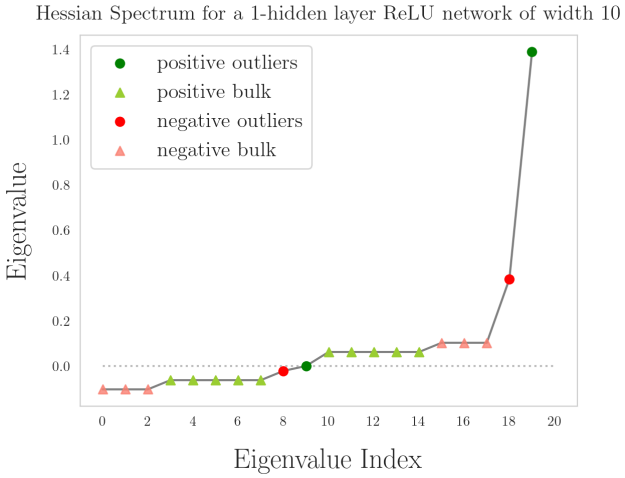


Figure 2: Hessian spectrum for ReLU networks. The eigenvalues are coloured based on their source cell, i.e., from the positive cell or the negative. Here, $q = 6$ neurons are present in the positive cell, while $m - q = 4$ in the negative cell. The dotted gray line demarcates the negative and positive eigenvalues.

respectively, and besides, the total number of points can split as $n = n_+ + n_-$. Also, denote the (uncentered) input-output covariance for the positive and negative cells respectively as follows, $\overline{yx}_+ = \frac{1}{n_+}\sum_{i=1}^{n} y_i x_i \mathbb{1}\{x_i > 0\}$, and $\overline{yx}_- = \frac{1}{n_-}\sum_{i=1}^{n} y_i x_i \mathbb{1}\{x_i \leq 0\}$. In this setting, the parameter space can be nicely partitioned in such a way that the Hessian can be decoupled.

Theorem 4 *For the setting of one-hidden layer ReLU scalar network, the Hessian is decoupled between the positive and the negative cells, and hence up to row and column permutations is,*

$$\mathbf{H}_{\mathrm{L}} = \begin{pmatrix} \frac{n_+}{n}\,{\mathbf{H}_{\mathrm{L}}}^{+} & \mathbf{0} \\ \mathbf{0} & \frac{n_-}{n}\,{\mathbf{H}_{\mathrm{L}}}^{-} \end{pmatrix} \tag{4}$$

where, the cell-wise Hessian matrices are weighed by the respective density of the cells.

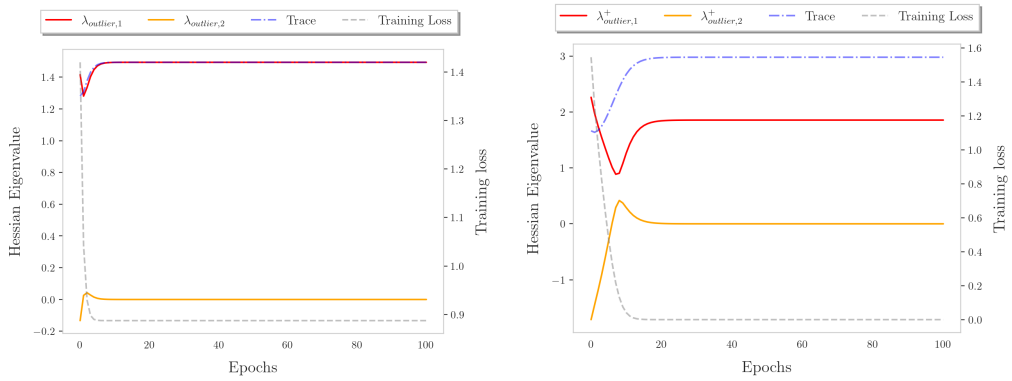
Let us assume that q coordinates of the parameter vector $\mathbf{v}$ are positive. Then we have that the spectrum in this ReLU case is a corollary of the previous Hessian decoupling result and the eigenvalues in the linear case.

Corollary 5 *The bulk Hessian spectrum consists of $q-1$ and $m-q-1$ repeated eigenvalues in signed pairs, $\lambda^{+}_{bulk} = \pm\frac{n_+}{n}\overline{x\delta}_+$, $\lambda^{-}_{bulk} = \pm\frac{n_-}{n}\overline{x\delta}_-$ and with the outlying eigenvalues being*

$$\lambda^{+}_{outlier_{1,2}} = \frac{n_+}{2n}(\sigma_+^2\|\mathbf{w}_+\|^2 + \sigma_+^2\|\mathbf{v}_+\|^2) \pm \frac{n_+}{2n}\sqrt{(\sigma_+^2\|\mathbf{w}_+\|^2 + \sigma_+^2\|\mathbf{v}_+\|^2)^2 + 4(\overline{\delta x}_+^2 + 2\sigma_+^2\overline{\delta x}_+\langle\mathbf{w}_+, \mathbf{v}_+\rangle)}$$
$$\lambda^{-}_{outlier_{1,2}} = \frac{n_-}{2n}(\sigma_-^2\|\mathbf{w}_-\|^2 + \sigma_-^2\|\mathbf{v}_-\|^2) \pm \frac{n_-}{2n}\sqrt{(\sigma_-^2\|\mathbf{w}_-\|^2 + \sigma_-^2\|\mathbf{v}_-\|^2)^2 + 4(\overline{\delta x}_-^2 + 2\sigma_-^2\overline{\delta x}_-\langle\mathbf{w}_-, \mathbf{v}_-\rangle)}$$

We see that just as in the linear case, we obtain a set of paired outlier eigenvalues, with a dependence on respective cell-wise quantities. Figure 3 highlights the paired nature of the outlying eigenvalues, for both linear and ReLU, throughout training. Similarly, the eigenvectors from the linear case also carry over to the ReLU case here.

We can see that in both cases the second outlying eigenvalue starts out negative, but which subsequently changes sign a little while into training, which would mark the exit of the early phase in our terminology, and finally converges to zero from the top (the positive side). On the other hand, when the learning rate is high, we find that the second outlying eigenvalue first drops and then increases, beyond which it wiggles around 0, and eventually becomes small and negative and



(*a*) Linear (*b*) ReLU

Figure 3: The paired nature of outlying eigenvalues throughout training. For the ReLU network, the outlying pair of eigenvalues corresponding to the positive cell are shown. Both the cases are with gradient descent on a synthetic wedge dataset, with a single hidden layer network of width 10 using a learning rate of 0.18 and momentum 0.9.

remains like that throughout. Hence, in the divergent phase, we can think of the second outlying eigenvalue approaching zero from the bottom (the negative side).

Impact of ReLU on the Hessian spectrum. First of all, we notice that the eigenvalues scale in proportion to the density (n_+/n or n_-/n) of the corresponding cell, i.e., cells which are active for a small number of samples will be less prominent in the spectrum, and vice versa. This also suggests a natural principle for the occurrence of numerous spuriously tiny eigenvalues with ReLU, as opposed to the linear case where there is a much clearer demarcation between zero and non-zero eigenvalues [6]. Lastly, while the above simple setup only has two partitions of the input space, and that too even mutually exclusive ones, we can nevertheless expect a non-trivial overlap and cross-terms to arise between the cells in the general case. But, it would form an interesting question for future work to see how well would the independent cell-wise Hessian serve as an approximation. All in all, this cell-wise decomposition of the Hessian hints at how the spectrum with a non-linearity like ReLU might be structured over and above the spectrum for a network with linear activations.

3. Extension to the bias case

In our analysis hitherto, we have assumed that the bias parameters are absent. By enabling bias parameters, we will have additional flexibility in the function class, and being a special case of $d = 2$ inputs, it will also provide us insights into how things change with increasing input dimension. So assume that our network is now, $f(\mathbf{x}) = \mathbf{w}^\top(\mathbf{v}x + \mathbf{b})$.

Theorem 6 *For the above linear network with bias, assume that* $\langle \mathbf{v}, \mathbf{b} \rangle = 0$ *and* $\|\mathbf{v}\| = \|\mathbf{b}\|$*, as well as* $\sigma^2 = 1$ *and zero-mean data* $\mathbf{E}[x] = 0, \mathbf{E}[y] = 0$ *and* $\overline{yx} = 0$*. Then the spectrum consists of the following sets of outlier eigenvalues, the first being the following pair,*

$$\lambda = \frac{1}{2}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) \pm \frac{1}{2}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 12(\overline{\delta x}^2 + \overline{\delta}^2)} \tag{5}$$

and the second is the triple $\lambda_k = t_k + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)/3$*, where* t_k *for* $k = 0, 1, 2$*:*

$$t_k = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right) - k\frac{2\pi}{3}\right)$$

with, $p = -(\overline{\delta x}^2 + \overline{\delta}^2) - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2/3$ *and* $q = -\frac{2}{27}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^3 - \frac{1}{3}\|\mathbf{v}\|^2(\overline{\delta x}^2 + \overline{\delta}^2) + \frac{2}{3}\|\mathbf{w}\|^2(\overline{\delta x}^2 + \overline{\delta}^2)$ *and the rest being the bulk eigenvalues* $\lambda_{bulk} = \pm\sqrt{\overline{\delta x}^2 + \overline{\delta}^2}$ *and* 0 *eigenvalues.*

The expression of the outlier pair strictly generalizes the case without bias, where the eigenvalues would be:

$$\lambda = \frac{1}{2}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) \pm \frac{1}{2}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 12\overline{\delta x}^2} \tag{6}$$

As fas as the eigenvalue triple is concerned, since $\cos \in [-1, 1]$, we can upper and lower bound the above solutions to the cubic as:

$$\frac{1}{3}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) - t' \leq \lambda_{0,1,2} \leq \frac{1}{3}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + t'$$

with

$$t' = 2\sqrt{\frac{-p}{3}} = 2\sqrt{\frac{1}{3}(\overline{\delta x}^2 + \overline{\delta}^2) + \frac{1}{9}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2} = \frac{2}{3}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 3(\overline{\delta x}^2 + \overline{\delta}^2)}\,.$$

The proof technique becomes highly involved and cumbersome in the above case, requiring solving the roots of a polynomial of degree 6. In general, polynomials of degree 5 or more do not have solutions in radicals, by the Abel-Ruffini theorem. However, the additional assumptions made in the theorem allow for a slight simplification of the resulting degree 6 equation, which can, in turn, be factorized into a product of a quadratic, cubic, and a linear term. Another thing worth remarking is that here we have a pair of eigenvalues centered at $1/4 \cdot \mathrm{Tr}(\mathbf{H}_\mathrm{L})$, and a triplet centered at $1/6 \cdot \mathrm{Tr}(\mathbf{H}_\mathrm{L})$, as here $\mathrm{Tr}(\mathbf{H}_\mathrm{L}) = 2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)$. Recall, for the $D = 1$ case, we had a pair of eigenvalues centered around $1/2 \cdot \mathrm{Tr}(\mathbf{H}_\mathrm{L})$. We conjecture that such a trend would also hold for the general d case, although we show it here for $D = 1, 2$ only.

4. Discussion

Summary. We provide a closed-form of the Hessian spectrum, i.e., all eigenvalues as well as eigenvectors, for both linear and ReLU networks in the scalar regression case. The obtained expressions provide insights into their intriguing nature in an exact manner, as seen via the paired nature of outlying eigenvalues and cell-wise decomposition of the Hessian for ReLU.

Related work. We would like to remark that the closest work in the literature to ours is that of [7] , who in their analysis of the Edge of Stability [1], derive the Hessian eigenvalues for an elementary 2-layer linear network with just scalar parameters $f(x) = wvx$ and a single input $n = 1$. We starkly go beyond the prior work by not only considering the general case of n inputs and vector parameters $\mathbf{w}, \mathbf{v}$, but we also cover the case of ReLU and bias as well as also derive the closed form of eigenvectors.

Conclusion. We hope that the wider research community can take advantage of our theoretical results to provide rigorous insights into the interaction of the maximum eigenvalue with the learning rate (like that showcased in the edge of stability [1]), the implicit effects of algorithms which minimize sharpness [2], as well as shed light into the flat minima hypothesis for generalization [4].

Acknowledgements

SPS would like to acknowledge the financial support from Max Planck ETH Center for Learning Systems. We also thank Weronika Ormaniec for the help in making the Figure 1 plot.

References

[1] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

[2] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[3] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density, 2019.

[4] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.

[5] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[6] Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. *Advances in Neural Information Processing Systems*, 34: 23914–23927, 2021.

[7] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

Appendix A. Eigenvalues: Linear, Unidimensional, Multiple datapoints

The loss function can be written as $\ell(\mathbf{w}, \mathbf{v}) = \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{v} \cdot x_i - y_i)^2$, where n is the number of data points under consideration. Hence the gradient with respect to the parameters comes out to be:

$$\nabla_{\mathbf{w}}\ell = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\delta_i\right)\mathbf{v} =: \overline{\delta x}\,\mathbf{v} \tag{7}$$

$$\nabla_{\mathbf{v}}\ell = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\delta_i\right)\mathbf{w} =: \overline{\delta x}\,\mathbf{w} \tag{8}$$

where, $\delta_i = \langle \mathbf{w}, \mathbf{v}\rangle\, x_i - y_i$ and we use the shorthand $\overline{x\delta}$ to designate the (uncentered) residual-input covariance. Also, let us denote the input mean as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the (uncentered) input standard deviation as $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$. Besides, let us denote the (uncentered) input-output covariance as $\overline{yx} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$. The second-order partial derivatives which will constitute the Hessian matrix turn out to be:

$$\nabla^2_{\mathbf{w},\mathbf{w}}\ell = \sigma^2\,\mathbf{v}\mathbf{v}^\top \tag{9}$$

$$\nabla^2_{\mathbf{v},\mathbf{v}}\ell = \sigma^2\,\mathbf{w}\mathbf{w}^\top \tag{10}$$

$$\nabla^2_{\mathbf{w},\mathbf{v}}\ell = \frac{\partial^2\ell}{\partial\mathbf{w}\partial\mathbf{v}} = \overline{\delta x}\,\mathbf{I}_m + \sigma^2\,\mathbf{v}\mathbf{w}^\top = \left(\nabla^2_{\mathbf{v},\mathbf{w}}\ell\right)^\top \tag{11}$$

We can also express the above in the matrix form as follows:

$$\mathbf{H}_{\mathrm{L}} = \begin{matrix}\frac{\partial}{\partial\mathbf{v}}\\ \frac{\partial}{\partial\mathbf{w}}\end{matrix}\overset{\begin{matrix}\frac{\partial}{\partial\mathbf{v}^\top} & \frac{\partial}{\partial\mathbf{w}^\top}\end{matrix}}{\begin{pmatrix}\sigma^2\,\mathbf{w}\mathbf{w}^\top & \sigma^2\,\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\mathbf{I}_m\\ \sigma^2\,\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\mathbf{I}_m & \sigma^2\,\mathbf{v}\mathbf{v}^\top\end{pmatrix}} \tag{12}$$

Solving the eigenvalues. To solve for the eigenvalues, we solve its characteristic equation, namely $|\mathbf{H}_{\mathrm{L}} - \lambda\mathbf{I}_p| = 0$, where $p = 2m$ is the number of parameters. Alternatively we have,

$$\left|\begin{pmatrix}\sigma^2\,\mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m & \sigma^2\,\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\mathbf{I}_m\\ \sigma^2\,\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\mathbf{I}_m & \sigma^2\,\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{I}_m\end{pmatrix}\right| = 0 \tag{13}$$

Via the Schur complement, we have $\left|\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\right| = |\mathbf{D}||\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|$. We can apply this to above equation, where $\mathbf{D} = \sigma^2\,\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{I}_m$, which is invertible as long as λ is non-zero and $\lambda \neq \|\mathbf{v}\|^2\sigma^2$. Hence determinant of $\mathbf{D}$ won't be zero, and the roots of the equation above (which will yield us the eigenvalues) will come from the other term, $|\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}| = 0$. Let us then calculate it:

$$\left|\sigma^2\,\mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m - (\sigma^2\,\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\,\mathbf{I}_m)(\sigma^2\,\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{I}_m)^{-1}(\sigma^2\,\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\,\mathbf{I}_m)\right| = 0 \tag{14}$$

Next, we use the Woodbury matrix identity, i.e., $(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$, which gives us that:

$$(\sigma^2\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{I}_m)^{-1} = \frac{-1}{\lambda}\mathbf{I}_m - \frac{\sigma^2}{\lambda(\lambda - \|\mathbf{v}\|^2\sigma^2)}\mathbf{v}\mathbf{v}^\top \tag{15}$$

Putting this back in the above equation and expanding the product gives us,

$$\left|\sigma^2\mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m + (\sigma^2\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\,\mathbf{I}_m)(\frac{1}{\lambda}\mathbf{I}_m + \frac{\sigma^2}{\lambda(\lambda - \|\mathbf{v}\|^2\sigma^2)}\mathbf{v}\mathbf{v}^\top)(\sigma^2\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\,\mathbf{I}_m)\right| = 0 \tag{16}$$

$$\Bigg|\sigma^2\mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m \tag{17}$$

$$+ \frac{\sigma^4\|\mathbf{v}\|^2}{\lambda}\mathbf{w}\mathbf{w}^\top + \frac{\sigma^2\overline{\delta x}}{\lambda}\mathbf{w}\mathbf{v}^\top + \frac{\sigma^6\|\mathbf{v}\|^4}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)}\mathbf{w}\mathbf{w}^\top + \frac{\sigma^4\overline{\delta x}\|\mathbf{v}\|^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)}\mathbf{w}\mathbf{v}^\top \tag{18}$$

$$+ \frac{\sigma^2\overline{\delta x}}{\lambda}\mathbf{v}\mathbf{w}^\top + \frac{\overline{\delta x}^2}{\lambda}\mathbf{I}_m + \frac{\sigma^4\overline{\delta x}\|\mathbf{v}\|^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)}\mathbf{v}\mathbf{w}^\top + \frac{\sigma^2\overline{\delta x}^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)}\mathbf{v}\mathbf{v}^\top\Bigg| = 0 \tag{19}$$

Let us analyze the coefficients for each of the matrices one by one, starting with $\mathbf{w}\mathbf{w}^\top$

$$\sigma^2 + \frac{\sigma^4\|\mathbf{v}\|^2}{\lambda} + \frac{\sigma^6\|\mathbf{v}\|^4}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)} = \frac{(\sigma^2\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2) + \sigma^4\|\mathbf{v}\|^2(\lambda - \sigma^2\|\mathbf{v}\|^2) + \sigma^6\|\mathbf{v}\|^4)}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)} \tag{20}$$

$$= \frac{\sigma^2\lambda}{\lambda - \sigma^2\|\mathbf{v}\|^2} \tag{21}$$

Next up, both $\mathbf{w}\mathbf{v}^\top$ and $\mathbf{v}\mathbf{w}^\top$ have the same coefficient:

$$\frac{\sigma^2\overline{\delta x}}{\lambda} + \frac{\sigma^4\overline{\delta x}\|\mathbf{v}\|^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)} = \frac{(\lambda - \sigma^2\|\mathbf{v}\|^2)\sigma^2\overline{\delta x} + \sigma^4\overline{\delta x}\|\mathbf{v}\|^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)} \tag{22}$$

$$= \frac{\sigma^2\overline{\delta x}}{\lambda - \sigma^2\|\mathbf{v}\|^2} \tag{23}$$

Hence, the above characteristic equation can be rewritten as,

$$\Bigg|\frac{\sigma^2\lambda}{\lambda - \sigma^2\|\mathbf{v}\|^2}\mathbf{w}\mathbf{w}^\top + \frac{\sigma^2\overline{\delta x}}{\lambda - \sigma^2\|\mathbf{v}\|^2}(\mathbf{w}\mathbf{v}^\top + \mathbf{v}\mathbf{w}^\top) + \frac{\sigma^2\overline{\delta x}^2}{\lambda(\lambda - \sigma^2\|\mathbf{v}\|^2)}\mathbf{v}\mathbf{v}^\top \tag{24}$$

$$- (\lambda - \frac{\overline{\delta x}^2}{\lambda})\mathbf{I}_m\Bigg| = 0 \tag{25}$$

Assuming $\lambda \neq 0$, multiply the equation by λ and $(\lambda - \sigma^2\|\mathbf{v}\|^2)$ (since $|\sigma^2\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{I}_m| \neq 0$) yields,

$$\left|\sigma^2\lambda^2\mathbf{w}\mathbf{w}^\top + \sigma^2\overline{\delta x}\lambda(\mathbf{w}\mathbf{v}^\top + \mathbf{v}\mathbf{w}^\top) + \sigma^2\overline{\delta x}^2\mathbf{v}\mathbf{v}^\top - (\lambda^2 - \overline{\delta x}^2)(\lambda - \sigma^2\|\mathbf{v}\|^2)\mathbf{I}_m\right| = 0 \tag{26}$$

Set $\mathbf{z} = \lambda\mathbf{w} + \overline{\delta x}\mathbf{v}$, and $\nu = (\lambda^2 - \overline{\delta x}^2)(\lambda - \sigma^2\|\mathbf{v}\|^2)$ we can express the above equation more compactly as

$$\left|\sigma^2\mathbf{z}\mathbf{z}^\top - \nu\mathbf{I}_m\right| = 0 \tag{27}$$

The determinant of the above matrix[2] is the product of its eigenvalues, which comes out as:

$$\left|\sigma^2\mathbf{z}\mathbf{z}^\top - \nu\mathbf{I}_m\right| = \nu^{m-1}(\sigma^2\|\mathbf{z}\|^2 - \nu) = 0 \tag{28}$$

This implies that they are $m-1$ repeated roots of $\nu = 0$, and once when $\sigma^2\|\mathbf{z}\|^2 - \nu = 0$.

Since $\lambda \neq \sigma^2\|\mathbf{v}\|^2$, we get $m-1$ repeated solutions of $\lambda^2 - \overline{\delta x}^2 = 0$ or $m-1$ times

$$\lambda = \pm\overline{\delta x} = \pm\frac{1}{n}\sum_{i=1}^{n}(\langle\mathbf{w},\mathbf{v}\rangle\, x_i - y_i)\, x_i\,. \tag{29}$$

The other solution corresponds to solving the following equation in λ:

$$\sigma^2\lambda^2\|\mathbf{w}\|^2 + \sigma^2\overline{\delta x}^2\|\mathbf{v}\|^2 + 2\sigma^2\overline{\delta x}\lambda\langle\mathbf{w},\mathbf{v}\rangle - \lambda^3 + \overline{\delta x}^2\|\mathbf{v}\|^2 + \overline{\delta x}^2 - \sigma^2\overline{\delta x}^2\|\mathbf{v}\|^2 = 0 \tag{30}$$

$$-\lambda^3 + (\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2)\lambda^2 + \lambda(\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle) \tag{31}$$

$$= -\lambda(\lambda^2 - (\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2)\lambda - (\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle)) = 0 \tag{32}$$

Again, $\lambda \neq 0$ by assumption above. Then solving the quadratic in λ gives us the following two roots:

$$\boxed{\lambda = \frac{1}{2}(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2) \pm \frac{1}{2}\sqrt{(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2)^2 + 4(\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle)}} \tag{33}$$

where $\overline{\delta x}^2 + 2\sigma^2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle = \overline{\delta x}(\overline{\delta x} + 2\sigma^2\langle\mathbf{w},\mathbf{v}\rangle)$, which can be better written as,

$$(\langle\mathbf{w},\mathbf{v}\rangle\sigma^2 - \overline{yx})(3\langle\mathbf{w},\mathbf{v}\rangle\sigma^2 - \overline{yx})\,.$$

Further, notice that the residual input covariance can be upper-bounded in terms of the overall loss and input variance.

$$\overline{\delta x} = \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i \leq \frac{1}{n}\sqrt{\sum_{i=1}^{n}\delta_i^2}\sqrt{\sum_{i=1}^{n}x_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2} \tag{34}$$

Thus, $\overline{\delta x} \leq \sqrt{2\ell}\,\sigma$. Further since the most negative eigenvalue is given by, $\lambda_{\text{min}} = -\overline{\delta x}$, we have that

$$\lambda_{\text{min}} \geq -\sqrt{2\ell}\,\sigma\,.$$

2. I could also make σ^2 part of the vector $\mathbf{z}$ above.

Appendix B. Eigenvalues: ReLU

Let us now analyze the case with the ReLU non-linearity. So we have as the network function, $f(x) = \mathbf{w}^\top(\mathbf{v}x)_+$, with $(z)_+ = z\mathbb{1}\{z > 0\}$. The loss function is given by, $\ell(\mathbf{w}, \mathbf{v}) = \frac{1}{2}(\mathbf{w}^\top(\mathbf{v}x)_+ - y)^2$.

Without loss of generality, assume that first q coordinates of the vector $\mathbf{v}$ have positive sign as x and collected in the vector $\mathbf{v}_+$, while the rest $m - q$ coordinates being zero. On the other hand, we can collect the negative coordinates of $\mathbf{v}$ in the vector $\mathbf{v}_-$ whose, first q components are zero and the rest $m - q$ contain the negative coordinates. Then we can write $\mathbf{v} = \mathbf{v}_+ + \mathbf{v}_-$. Since in this simple network, each parameter of $\mathbf{w}$ is coupled with a parameter in $\mathbf{v}$, based on the partition of $\mathbf{v}$ we can also split the $\mathbf{w}$ vector into two corresponding parts, $(\mathbf{w}_+)_j = \mathbf{w}_j\mathbb{1}\{\mathbf{v}_j > 0\}$ and $(\mathbf{w}_-)_j = \mathbf{w}_j\mathbb{1}\{\mathbf{v}_j < 0\}$. Alternatively, we can express the effect of non-linearity on $\mathbf{w}$ by the Hadamard product $\mathbf{w} \odot \mathbb{1}\{\mathbf{v} > 0\}$. But, to emphasize, the components of $\mathbf{w}_+$ need not be positive neither $\mathbf{w}_-$ need be negative.

And so, for $x > 0$, $f(x) = \langle \mathbf{w}_+, \mathbf{v}_+\rangle x$, while for $x \leq 0$, we have that $f(x) = \langle \mathbf{w}_-, \mathbf{v}_-\rangle x$. Or more succinctly,

$$f(x) = \langle \mathbf{w}_+, \mathbf{v}_+\rangle x\, \mathbb{1}\{x > 0\} + \langle \mathbf{w}_-, \mathbf{v}_-\rangle x\, \mathbb{1}\{x \leq 0\}$$

Hence the gradient with respect to the parameters comes out to be:

$$\nabla_{\mathbf{w}}\ell = \delta \cdot (\mathbf{v}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{v}_- x\, \mathbb{1}\{x \leq 0\}) \tag{35}$$

$$\nabla_{\mathbf{v}}\ell = \delta \cdot (\mathbf{w}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{w}_- x\, \mathbb{1}\{x \leq 0\}) \tag{36}$$

where, $\delta = \langle \mathbf{w}_+, \mathbf{v}_+\rangle x\, \mathbb{1}\{x > 0\} + \langle \mathbf{w}_-, \mathbf{v}_-\rangle x\, \mathbb{1}\{x \leq 0\} - y$. The second-order partial derivatives which will constitute the Hessian matrix turn out to be:

$$\nabla^2_{\mathbf{w},\mathbf{w}}\ell = (\mathbf{v}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{v}_- x\, \mathbb{1}\{x \leq 0\})\,(\mathbf{v}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{v}_- x\, \mathbb{1}\{x \leq 0\})^\top$$

$$\nabla^2_{\mathbf{v},\mathbf{v}}\ell = (\mathbf{w}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{w}_- x\, \mathbb{1}\{x \leq 0\})\,(\mathbf{w}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{w}_- x\, \mathbb{1}\{x \leq 0\})^\top$$

$$\nabla^2_{\mathbf{w},\mathbf{v}}\ell = \frac{\partial^2 \ell}{\partial \mathbf{w} \partial \mathbf{v}} = (\delta\, \mathbf{x}_\pm) + (\mathbf{v}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{v}_- x\, \mathbb{1}\{x \leq 0\})\,(\mathbf{w}_+ x\, \mathbb{1}\{x > 0\} + \mathbf{w}_- x\, \mathbb{1}\{x \leq 0\})^\top$$

where, $(\cdot)$ denotes a diagonal matrix formed by the corresponding vector, and $\mathbf{x}_\pm$ denotes the vector whose first q coordinates contain equal $x\mathbb{1}\{x > 0\}$ and the other $m-q$ coordinates contain $x\mathbb{1}\{x \leq 0\}$. Said differently, we have that:

$$(\delta\, \mathbf{x}_\pm) = \begin{pmatrix} \delta\, x\, \mathbb{1}\{x > 0\}\ \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \delta\, x\, \mathbb{1}\{x \leq 0\}\ \mathbf{I}_{m-q} \end{pmatrix}$$

The expressions of the above components of the Hessian can be further simplified as,

$$\nabla^2_{\mathbf{w},\mathbf{w}}\ell = \mathbf{v}_+\mathbf{v}_+^\top x^2\, \mathbb{1}\{x > 0\} + \mathbf{v}_-\mathbf{v}_-^\top x^2 \mathbb{1}\{x \leq 0\} \tag{37}$$

$$\nabla^2_{\mathbf{v},\mathbf{v}}\ell = \mathbf{w}_+\mathbf{w}_+^\top x^2\, \mathbb{1}\{x > 0\} + \mathbf{w}_-\mathbf{w}_-^\top x^2 \mathbb{1}\{x \leq 0\} \tag{38}$$

$$\nabla^2_{\mathbf{w},\mathbf{v}}\ell = (\delta\, \mathbf{x}_\pm) + \mathbf{v}_+\mathbf{w}_+^\top x^2\, \mathbb{1}\{x > 0\} + \mathbf{v}_-\mathbf{w}_-^\top x^2 \mathbb{1}\{x \leq 0\} \tag{39}$$

Averaging over Multiple datapoints. Let us average the above Hessian components over multiple datapoints,

$$\nabla^2_{\mathbf{w},\mathbf{w}} \mathrm{L} = \frac{n_+}{n}\,\sigma_+^2\;\mathbf{v}_+\mathbf{v}_+^\top + \frac{n_-}{n}\,\sigma_-^2\;\mathbf{v}_-\mathbf{v}_-^\top \tag{40}$$

$$\nabla^2_{\mathbf{v},\mathbf{v}} \mathrm{L} = \frac{n_+}{n}\,\sigma_+^2\;\mathbf{w}_+\mathbf{w}_+^\top + \frac{n_-}{n}\,\sigma_-^2\;\mathbf{w}_-\mathbf{w}_-^\top \tag{41}$$

$$\nabla^2_{\mathbf{w},\mathbf{v}} \mathrm{L} = (\overline{\delta \mathbf{x}_\pm}) + \frac{n_+}{n}\,\sigma_+^2\;\mathbf{v}_+\mathbf{w}_+^\top + \frac{n_-}{n}\,\sigma_-^2\;\mathbf{v}_-\mathbf{w}_-^\top \tag{42}$$

where, we have the the standard deviation of the positive and negative datapoints as, $\sigma_+ = \sqrt{\frac{1}{n_+}\sum_{i=1}^n x_i^2\,\mathbb{1}\{x_i > 0\}}$ and $\sigma_- = \sqrt{\frac{1}{n_-}\sum_{i=1}^n x_i^2\,\mathbb{1}\{x_i < 0\}}$ respectively, and besides, $n = n_+ + n_-$.

Besides, we let us denote the (uncentered) input-output covariance for the positive and negative cells respectively as follows,

$$\overline{yx}_+ = \frac{1}{n_+}\sum_{i=1}^n y_i x_i \mathbb{1}\{x_i > 0\}$$

$$\overline{yx}_- = \frac{1}{n_-}\sum_{i=1}^n y_i x_i \mathbb{1}\{x_i \le 0\}$$

Hessian decouples across ReLU cells. If we look at the equations above carefully, we will notice that the coordinates for $\mathbf{v}_+, \mathbf{w}_+$ are mutually exclusive from the coordinates for $\mathbf{v}_-, \mathbf{w}_-$. And hence, up to column and row permutations the Hessian can be written as a block-diagonal matrix with Hessian for the individual cells respectively in these diagonal blocks. In fact, the column space itself is a direct sum of the column space across the cells. And, so the eigenvectors will also be non-zero on mutually exclusive coordinates. As a result, we can solve the Hessian spectrum separately for the two cells.

We can also express the above in the matrix form rather compactly as follows:

$$\frac{n_+}{n}\mathbf{H}_{\mathrm{L}}{}^+ = \begin{matrix}\frac{\partial}{\partial \mathbf{v}_+} \\ \frac{\partial}{\partial \mathbf{w}_+}\end{matrix}\overset{\begin{matrix}\frac{\partial}{\partial \mathbf{v}_+^\top} & \frac{\partial}{\partial \mathbf{w}_+^\top}\end{matrix}}{\begin{pmatrix} \frac{n_+}{n}\,\sigma_+^2\;\mathbf{w}_+\mathbf{w}_+^\top & \frac{n_+}{n}\,\sigma_+^2\;\mathbf{w}_+\mathbf{v}_+^\top + \frac{n_+}{n}\,\overline{x\delta}_+\mathbf{I}_q \\ \frac{n_+}{n}\,\sigma_+^2\;\mathbf{v}_+\mathbf{w}_+^\top + \frac{n_+}{n}\,\overline{x\delta}_+\mathbf{I}_q & \frac{n_+}{n}\,\sigma_+^2\;\mathbf{v}_+\mathbf{v}_+^\top \end{pmatrix}}$$

where, $\overline{x\delta}_+ = \sigma_+^2\,\langle \mathbf{w}_+, \mathbf{v}_+ \rangle - \overline{yx}_+$. Likewise, we have that

$$\frac{n_-}{n}\mathbf{H}_{\mathrm{L}}{}^- = \begin{matrix}\frac{\partial}{\partial \mathbf{v}_-} \\ \frac{\partial}{\partial \mathbf{w}_-}\end{matrix}\overset{\begin{matrix}\frac{\partial}{\partial \mathbf{v}_-^\top} & \frac{\partial}{\partial \mathbf{w}_-^\top}\end{matrix}}{\begin{pmatrix} \frac{n_-}{n}\,\sigma_-^2\;\mathbf{w}_-\mathbf{w}_-^\top & \frac{n_-}{n}\,\sigma_-^2\;\mathbf{w}_-\mathbf{v}_-^\top + \frac{n_-}{n}\,\overline{x\delta}_-\,\mathbf{I}_{m-q} \\ \frac{n_-}{n}\,\sigma_-^2\;\mathbf{v}_-\mathbf{w}_-^\top + \frac{n_-}{n}\,\overline{x\delta}_-\,\mathbf{I}_{m-q} & \frac{n_-}{n}\,\sigma_-^2\;\mathbf{v}_-\mathbf{v}_-^\top \end{pmatrix}}$$

To emphasize, the Hessian is equivalent to:

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

Essentially, the row and column permutations correspond to a different ordering of the parameters when forming the Hessian matrix.

Solving the eigenvalues. For solving the eigenvalues, we can depend on the result for the linear case and do so separately: once for the positive cell and another time for the negative cell. The eigenvalues for the overall Hessian will just be a union of the eigenvalues obtained from solving the eigenspectrum for the cells. We will just have to multiply the obtained eigenvalues for ${\mathbf{H}_{\mathrm{L}}}^{+}$ and ${\mathbf{H}_{\mathrm{L}}}^{-}$ by the weights $\frac{n_+}{n}$ and $\frac{n_-}{n}$ respectively.

Positive Cell Eigenvalues. Here we obtain as the bulk eigenvalue, $\lambda^{+}_{\text{bulk}} = \pm\frac{n_+}{n}\overline{x\delta}_+$ repeated $q-1$ times, while as the outlier eigenvalues we have:

$$\lambda^{+}_{\text{outlier}} = \frac{n_+}{2n}(\sigma_+^2\|\mathbf{w}_+\|^2 + \sigma_+^2\|\mathbf{v}_+\|^2) \pm \frac{n_+}{2n}\sqrt{(\sigma_+^2\|\mathbf{w}_+\|^2 + \sigma_+^2\|\mathbf{v}_+\|^2)^2 + 4(\overline{x\delta}_+^2 + 2\sigma_+^2\overline{x\delta}_+\langle\mathbf{w}_+, \mathbf{v}_+\rangle)} \tag{43}$$

Negative Cell Eigenvalues. Here we obtain as the bulk eigenvalue, $\lambda^{-}_{\text{bulk}} = \pm\frac{n_-}{n}\overline{x\delta}_-$ repeated $m-q-1$ times, while as the outlier eigenvalues we have:

$$\lambda^{-}_{\text{outlier}} = \frac{n_-}{2n}(\sigma_-^2\|\mathbf{w}_-\|^2 + \sigma_-^2\|\mathbf{v}_-\|^2) \pm \frac{n_-}{2n}\sqrt{(\sigma_-^2\|\mathbf{w}_-\|^2 + \sigma_-^2\|\mathbf{v}_-\|^2)^2 + 4(\overline{x\delta}_-^2 + 2\sigma_-^2\overline{x\delta}_-\langle\mathbf{w}_-, \mathbf{v}_-\rangle)} \tag{44}$$

Appendix C. Solving for Eigenvectors

C.1. Linear, Unidimensional network with multiple datapoints

Let us recall the Hessian considered in Eq. 12 for the multiple datapoints setting in the case of linear networks.

$$\mathbf{H}_{\mathrm{L}} = \begin{matrix}\\ \frac{\partial}{\partial \mathbf{v}}\\ \frac{\partial}{\partial \mathbf{w}}\end{matrix}\overset{\begin{matrix}\frac{\partial}{\partial \mathbf{v}^\top} & \frac{\partial}{\partial \mathbf{w}^\top}\end{matrix}}{\begin{pmatrix}\sigma^2\,\mathbf{w}\mathbf{w}^\top & \sigma^2\,\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\mathbf{I}_m\\ \sigma^2\,\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\mathbf{I}_m & \sigma^2\,\mathbf{v}\mathbf{v}^\top\end{pmatrix}} \tag{45}$$

While we have solved for the eigenvalues, the form of the eigenvectors is still not apparent. In this section, we aim to work towards a closed form for the eigenvectors. For starters, let us assume an eigenvector of the Hessian matrix above, $\mathbf{z}$, is of the form $\mathbf{z} = \begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix}$. Then in order to obtain the eigenvectors we need to solve the following system of equations:

$$\begin{pmatrix}\sigma^2\,\mathbf{w}\mathbf{w}^\top & \sigma^2\,\mathbf{w}\mathbf{v}^\top + \overline{\delta x}\mathbf{I}_m\\ \sigma^2\,\mathbf{v}\mathbf{w}^\top + \overline{\delta x}\mathbf{I}_m & \sigma^2\,\mathbf{v}\mathbf{v}^\top\end{pmatrix}\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix} = \lambda\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix} \tag{46}$$

which can be expressed as,

$$\sigma^2\langle \mathbf{w}, \mathbf{z}_1\rangle \mathbf{w} + \sigma^2\langle \mathbf{v}, \mathbf{z}_2\rangle \mathbf{w} + \overline{\delta x}\, \mathbf{z}_2 = \lambda \mathbf{z}_1 \tag{47}$$

$$\sigma^2\langle \mathbf{w}, \mathbf{z}_1\rangle \mathbf{v} + \sigma^2\langle \mathbf{v}, \mathbf{z}_2\rangle \mathbf{v} + \overline{\delta x}\, \mathbf{z}_1 = \lambda \mathbf{z}_2 \tag{48}$$

Taking the inner-product with $\mathbf{v}$ on both sides of first equation and with $\mathbf{w}$ in second equation yields,

$$\sigma^2\langle \mathbf{w}, \mathbf{z}_1\rangle \langle \mathbf{w}, \mathbf{v}\rangle + \sigma^2\langle \mathbf{v}, \mathbf{z}_2\rangle \langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta x}\, \langle \mathbf{z}_2, \mathbf{v}\rangle = \lambda \langle \mathbf{z}_1, \mathbf{v}\rangle \tag{49}$$

$$\sigma^2\langle \mathbf{w}, \mathbf{z}_1\rangle \langle \mathbf{v}, \mathbf{w}\rangle + \sigma^2\langle \mathbf{v}, \mathbf{z}_2\rangle \langle \mathbf{v}, \mathbf{w}\rangle + \overline{\delta x}\, \langle \mathbf{z}_1, \mathbf{w}\rangle = \lambda \langle \mathbf{z}_2, \mathbf{w}\rangle \tag{50}$$

Subtracting the second equation from the first and rearranging gives,

$$\overline{\delta x}\, \langle \mathbf{z}_2, \mathbf{v}\rangle - \lambda \langle \mathbf{z}_1, \mathbf{v}\rangle = \overline{\delta x}\, \langle \mathbf{z}_1, \mathbf{w}\rangle - \lambda \langle \mathbf{z}_2, \mathbf{w}\rangle \tag{51}$$

Alternatively,

$$\langle \overline{\delta x}\, \mathbf{z}_2 - \lambda\, \mathbf{z}_1, \mathbf{v}\rangle = \langle \overline{\delta x}\, \mathbf{z}_1 - \lambda\, \mathbf{z}_2, \mathbf{w}\rangle \tag{52}$$

Thus we have an equation of the form $\langle \mathbf{a}, \mathbf{b}\rangle = \langle \mathbf{c}, \mathbf{d}\rangle$, whose possible solutions are:

1. $\langle \mathbf{a}, \mathbf{b}\rangle = 0$ and $\langle \mathbf{c}, \mathbf{d}\rangle = 0$. While this is a general condition, there are also some specific instantiations of this when this is possible as listed below.
   (a) $\mathbf{a} = 0$ and $\mathbf{c} = 0$.
   (b) $\mathbf{a} = 0$ and $\mathbf{d} = 0$.
   (c) $\mathbf{b} = 0$ and $\mathbf{c} = 0$.
   (d) $\mathbf{b} = 0$ and $\mathbf{d} = 0$.

   The grayed out possibilities require the parameter vectors to be zero, so we discard them as that need not be the case.

2. $\mathbf{a} = \mathbf{c}$ and $\mathbf{b} = \mathbf{d}$, and $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \neq 0$ (up to scale, since $\alpha\, \mathbf{a}$ and $\alpha^{-1}\mathbf{b}$, for $\alpha \neq 0$ is also a valid solution).

   Again this requires the parameter vectors to be equal, which may not be the case necessarily.

3. $\mathbf{a} = \mathbf{d}$ and $\mathbf{b} = \mathbf{c}$, and $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \neq 0$ (up to scale).

Outlier Eigenvectors. Considering that $\mathbf{v} \neq \mathbf{w}$ and $\mathbf{v}, \mathbf{w} \neq 0$. Then using the (3) option above, we get (for some $\alpha \neq 0$):

$$\overline{\delta x}\, \mathbf{z}_2 - \lambda \mathbf{z}_1 = \alpha \mathbf{w} \tag{53}$$

$$\overline{\delta x} \mathbf{z}_1 - \lambda\, \mathbf{z}_2 = \alpha \mathbf{v} \tag{54}$$

Then solving this pair of equations for $\mathbf{z}_1$ and $\mathbf{z}_2$ gives:

$$(\overline{\delta x}^2 - \lambda^2)\, \mathbf{z}_1 = \alpha\lambda\, \mathbf{w} + \alpha\overline{\delta x}\, \mathbf{v} \tag{55}$$

$$(\overline{\delta x}^2 - \lambda^2)\, \mathbf{z}_2 = \alpha\overline{\delta x}\, \mathbf{w} + \alpha\lambda\, \mathbf{v} \tag{56}$$

Hence, at last, the eigenvector is of the following form:

$$\mathbf{z} = \frac{\alpha}{\overline{\delta x}^2 - \lambda^2} \begin{pmatrix} \lambda \mathbf{w} + \overline{\delta x}\mathbf{v} \\ \overline{\delta x}\mathbf{w} + \lambda \mathbf{v} \end{pmatrix} \tag{57}$$

In order to check the validity of the above solution, let us multiply the Hessian with it. Besides, from there we may also see the eigenvalue corresponding to this eigenvector.

$$\begin{pmatrix} \sigma^2 \mathbf{w}\mathbf{w}^\top & \sigma^2 \mathbf{w}\mathbf{v}^\top + \overline{\delta x}\mathbf{I}_m \\ \sigma^2 \mathbf{v}\mathbf{w}^\top + \overline{\delta x}\mathbf{I}_m & \sigma^2 \mathbf{v}\mathbf{v}^\top \end{pmatrix} \begin{pmatrix} \lambda \mathbf{w} + \overline{\delta x}\mathbf{v} \\ \overline{\delta x}\mathbf{w} + \lambda \mathbf{v} \end{pmatrix} \frac{\alpha}{\overline{\delta x}^2 - \lambda^2}$$
$$= \begin{pmatrix} \lambda\sigma^2\langle \mathbf{w}, \mathbf{w}\rangle \mathbf{w} + \overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle \mathbf{w} + \overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle \mathbf{w} + \lambda\sigma^2\langle \mathbf{v}, \mathbf{v}\rangle \mathbf{w} + \overline{\delta x}^2 \mathbf{w} + \lambda\overline{\delta x}\mathbf{v} \\ \lambda\sigma^2\langle \mathbf{w}, \mathbf{w}\rangle \mathbf{v} + \overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle \mathbf{v} + \lambda\overline{\delta x}\mathbf{w} + \overline{\delta x}^2 \mathbf{v} + \overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle \mathbf{v} + \lambda\sigma^2\langle \mathbf{v}, \mathbf{v}\rangle \mathbf{v} \end{pmatrix} \frac{\alpha}{\overline{\delta x}^2 - \lambda^2}$$

The above can be simplified to:

$$\begin{pmatrix} \left(\lambda\sigma^2\|\mathbf{w}\|^2 + \lambda\sigma^2\|\mathbf{v}\|^2 + 2\overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta x}^2\right) \mathbf{w} + \lambda\overline{\delta x}\mathbf{v} \\ \lambda\overline{\delta x}\mathbf{w} + \left(\lambda\sigma^2\|\mathbf{w}\|^2 + \lambda\sigma^2\|\mathbf{v}\|^2 + 2\overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta x}^2\right) \mathbf{v} \end{pmatrix} \frac{\alpha}{\overline{\delta x}^2 - \lambda^2}$$

For this to be a valid eigenvector, the term in brackets should be equal to λ^2, i.e.,

$$\lambda\sigma^2\|\mathbf{w}\|^2 + \lambda\sigma^2\|\mathbf{v}\|^2 + 2\overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta x}^2 = \lambda^2$$

Solving this yields,

$$\lambda = \frac{1}{2}(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2) \pm \frac{1}{2}\sqrt{(\sigma^2\|\mathbf{w}\|^2 + \sigma^2\|\mathbf{v}\|^2)^2 + 4(2\overline{\delta x}\sigma^2\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta x}^2)}$$

which is precisely the solution for the eigenvalues we had obtained by solving the characteristic equation.

Besides, we need to ensure that eigenvectors are unit norm, which should give us:

Remark 1. The expression in Eqn. 57 matches the empirically obtained eigenvectors.

Remark 2. Given the decoupling of the Hessian spectrum in ReLU cells, the above eigenvector derivation should also generalize for the ReLU case.

Bulk Eigenvectors. We have the possibility (1a) remaining, and we find that as the bulk eigenvalues are $\lambda = \pm\overline{\delta x}$, this would fit neatly with constraint from Eqn. **??**. Moreover we obtain the following:

$$\lambda = \overline{\delta x}: \quad \mathbf{z}_1 = \mathbf{z}_2 \tag{58}$$
$$\lambda = -\overline{\delta x}: \quad \mathbf{z}_1 = -\mathbf{z}_2 \tag{59}$$

Further as the bulk eigenvectors have to be orthogonal to the outlier eigenvectors, we get the following constraint:

$$\lambda_{\text{outlier}}\langle \mathbf{w}, \mathbf{z}_1\rangle + \overline{\delta x}\langle \mathbf{v}, \mathbf{z}_1\rangle + \overline{\delta x}\langle \mathbf{w}, \mathbf{z}_2\rangle + \lambda_{\text{outlier}}\langle \mathbf{v}, \mathbf{z}_2\rangle = 0 \tag{60}$$

$$\langle \mathbf{w}, \lambda_{\text{outlier}}\mathbf{z}_1 + \overline{\delta x}\mathbf{z}_2\rangle + \langle \mathbf{v}, \overline{\delta x}\mathbf{z}_1 \ + \lambda_{\text{outlier}}\mathbf{z}_2\rangle = 0 \tag{61}$$

(a) Bulk eigenvalue and residual-input covariance have same signs: Let us now consider $\mathbf{z}_1 = \mathbf{z}_2$, then we get the following constraint:

$$(\lambda_{\text{outlier}} + \overline{\delta x})\langle \mathbf{w} + \mathbf{v}, \mathbf{z}_1\rangle = 0 \tag{62}$$

Hence, in this case $\mathbf{z}_1$ is of the form $\left(\mathbf{I} - \frac{(\mathbf{w}+\mathbf{v})(\mathbf{w}+\mathbf{v})^\top}{\|\mathbf{w}+\mathbf{v}\|^2}\right)\mathbf{c}$ for some vector $\mathbf{c}$, and they can simply be obtained by computing the eigenvectors of this matrix, corresponding to non-zero eigenvalues.

(b) Bulk eigenvalue and residual-input covariance have opposite signs: While plugging in $\mathbf{z}_1 = -\mathbf{z}_2$ for the other half of the bulk, we get the following constraint:

$$(\lambda_{\text{outlier}} - \overline{\delta x})\langle \mathbf{w} - \mathbf{v}, \mathbf{z}_1\rangle = 0 \tag{63}$$

Hence, in this case $\mathbf{z}_1$ is of the form $\left(\mathbf{I} - \frac{(\mathbf{w}-\mathbf{v})(\mathbf{w}-\mathbf{v})^\top}{\|\mathbf{w}-\mathbf{v}\|^2}\right)\mathbf{c}$ for some vector $\mathbf{c}$, and they can simply be obtained by computing the eigenvectors of this matrix, corresponding to non-zero eigenvalues.

Finally, the bulk eigenvectors from the cases (a) and (b) are orthogonal between themselves since,

$$\begin{pmatrix}\mathbf{z}_1 & \mathbf{z}_1\end{pmatrix}^\top \begin{pmatrix}\mathbf{z}_1 \\ -\mathbf{z}_1\end{pmatrix} = 0$$

C.2. ReLU, Unidimensional, multiple datapoints

We simply follow our strategy for the linear case, and since the Hessian for ReLU decouples into that along the positive and negative cell, we obtain the following set of eigenvectors for the outlier and bulk eigenvalues corresponding to the respective cells.

Let us briefly recall our shorthand from before, $\mathbf{w}_+ = \mathbf{w} \odot \mathbb{1}\{\mathbf{v} > 0\}$, $\mathbf{w}_- = \mathbf{w} \odot \mathbb{1}\{\mathbf{v} \leq 0\}$, $\mathbf{v}_+ = \mathbf{v} \odot \mathbb{1}\{\mathbf{v} > 0\}$ and $\mathbf{v}_- = \mathbf{v} \odot \mathbb{1}\{\mathbf{v} \leq 0\}$.

C.2.1. POSITIVE CELL EIGENVECTORS

The outlier eigenvectors are given by,

$$\mathbf{z}_+ = \frac{\alpha}{\overline{x\delta}_+^2 - \lambda_+^2} \begin{pmatrix} \lambda_{\text{outlier}}^+ \mathbf{w}_+ \, + \, \overline{x\delta}_+ \mathbf{v}_+ \\ \overline{x\delta}_+ \, \mathbf{w}_+ \, + \, \lambda_{\text{outlier}}^+ \mathbf{v}_+ \end{pmatrix} \tag{64}$$

with, $\lambda_{\text{outlier}}^+ = \frac{n_+}{2n}(\sigma_+^2 \|\mathbf{w}_+\|^2 + \sigma_+^2 \|\mathbf{v}_+\|^2) \pm \frac{n_+}{2n} \sqrt{(\sigma_+^2 \|\mathbf{w}_+\|^2 + \sigma_+^2 \|\mathbf{v}_+\|^2)^2 + 4(\overline{x\delta}_+^2 + 2\sigma_+^2 \overline{x\delta}_+ \langle \mathbf{w}_+, \mathbf{v}_+ \rangle)}$.

The bulk eigenvectors depend whether their eigenvalue is $\lambda_{\text{bulk}}^+ = \overline{x\delta}_+$ or $\lambda_{\text{bulk}}^+ = -\overline{x\delta}_+$. In the former case, we have eigenvectors of the form $\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$, where $\mathbf{z}_1 = \mathbf{z}_2$ is of the form $\left(\mathbf{I} - \frac{(\mathbf{w}_+ + \mathbf{v}_+)(\mathbf{w}_+ + \mathbf{v}_+)^\top}{\|\mathbf{w}_+ + \mathbf{v}_+\|^2} \right) \mathbf{c}$ for some vector $\mathbf{c}$, and they can simply be obtained by computing the eigenvectors of this matrix, corresponding to non-zero eigenvalues. In the latter case, we have that $\mathbf{z}_1 = -\mathbf{z}_2$ with, $\mathbf{z}_1$ of the form $\left(\mathbf{I} - \frac{(\mathbf{w}_+ - \mathbf{v}_+)(\mathbf{w}_+ - \mathbf{v}_+)^\top}{\|\mathbf{w}_+ - \mathbf{v}_+\|^2} \right) \mathbf{c}$ for some vector $\mathbf{c}$.

C.2.2. NEGATIVE CELL EIGENVECTORS

The outlier eigenvectors are given by,

$$\mathbf{z}_- = \frac{\alpha}{\overline{x\delta}_-^2 - \lambda_-^2} \begin{pmatrix} \lambda_{\text{outlier}}^- \mathbf{w}_- \, + \, \overline{x\delta}_- \mathbf{v}_- \\ \overline{x\delta}_- \, \mathbf{w}_- \, + \, \lambda_{\text{outlier}}^- \mathbf{v}_- \end{pmatrix} \tag{65}$$

with, $\lambda_{\text{outlier}}^- = \frac{n_-}{2n}(\sigma_-^2 \|\mathbf{w}_-\|^2 + \sigma_-^2 \|\mathbf{v}_-\|^2) \pm \frac{n_-}{2n} \sqrt{(\sigma_-^2 \|\mathbf{w}_-\|^2 + \sigma_-^2 \|\mathbf{v}_-\|^2)^2 + 4(\overline{x\delta}_-^2 + 2\sigma_-^2 \overline{x\delta}_- \langle \mathbf{w}_-, \mathbf{v}_- \rangle)}$.

The bulk eigenvectors depend whether their eigenvalue is $\lambda_{\text{bulk}}^- = \overline{x\delta}_-$ or $\lambda_{\text{bulk}}^- = -\overline{x\delta}_-$. In the former case, we have eigenvectors of the form $\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$, where $\mathbf{z}_1 = \mathbf{z}_2$ is of the form $\left(\mathbf{I} - \frac{(\mathbf{w}_- + \mathbf{v}_-)(\mathbf{w}_- + \mathbf{v}_-)^\top}{\|\mathbf{w}_- + \mathbf{v}_-\|^2} \right) \mathbf{c}$ for some vector $\mathbf{c}$, and they can simply be obtained by computing the eigenvectors of this matrix, corresponding to non-zero eigenvalues. In the latter case, we have that $\mathbf{z}_1 = -\mathbf{z}_2$ with, $\mathbf{z}_1$ of the form $\left(\mathbf{I} - \frac{(\mathbf{w}_- - \mathbf{v}_-)(\mathbf{w}_- - \mathbf{v}_-)^\top}{\|\mathbf{w}_- - \mathbf{v}_-\|^2} \right) \mathbf{c}$ for some vector $\mathbf{c}$.

Appendix D. Proof for the bias case

Let us start with the linear case and assume that our network is now given by,

$$f(\mathbf{x}) = \mathbf{w}^\top(\mathbf{v}x + \mathbf{b})$$

.

The expressions of the loss gradient are:

$$\nabla_{\mathbf{w}}\ell = \delta \cdot (\mathbf{v}x + \mathbf{b}) \tag{66}$$
$$\nabla_{\mathbf{v}}\ell = \delta \cdot \mathbf{w}x \tag{67}$$
$$\nabla_{\mathbf{b}}\ell = \delta \cdot \mathbf{w} \tag{68}$$

where, $\delta = f(\mathbf{x}) - y = \mathbf{w}^\top(\mathbf{v}x + \mathbf{b}) - y$. The Hessian terms are then as follows[3]:

$$\nabla^2_{\mathbf{w},\mathbf{w}}\ell = (\mathbf{v}x + \mathbf{b})(\mathbf{v}x + \mathbf{b})^\top \tag{69}$$
$$\nabla^2_{\mathbf{w},\mathbf{v}}\ell = \delta x\mathbf{I}_m + x(\mathbf{v}x + \mathbf{b})\mathbf{w}^\top \tag{70}$$
$$\nabla^2_{\mathbf{w},\mathbf{b}}\ell = \delta\mathbf{I}_m + (\mathbf{v}x + \mathbf{b})\mathbf{w}^\top \tag{71}$$
$$\nabla^2_{\mathbf{v},\mathbf{v}}\ell = x^2\mathbf{w}\mathbf{w}^\top, \quad \nabla^2_{\mathbf{v},\mathbf{b}}\ell = x\mathbf{w}\mathbf{w}^\top \tag{72}$$
$$\nabla^2_{\mathbf{b},\mathbf{b}}\ell = \mathbf{w}\mathbf{w}^\top \tag{73}$$

Next, let us aggregate the above expression over the entire dataset, with the assumption that the data is centered (which is anyways carried out in practice), i.e., $\mathbf{E}[\mathbf{x}] = 0$, yielding :

$$\nabla^2_{\mathbf{w},\mathbf{w}}\mathrm{L} = \sigma^2\mathbf{v}\mathbf{v}^\top + \mathbf{b}\mathbf{b}^\top \tag{74}$$
$$\nabla^2_{\mathbf{w},\mathbf{v}}\mathrm{L} = \overline{\delta x}\mathbf{I}_m + \sigma^2\mathbf{v}\mathbf{w}^\top \tag{75}$$
$$\nabla^2_{\mathbf{w},\mathbf{b}}\mathrm{L} = \overline{\delta}\mathbf{I}_m + \mathbf{b}\mathbf{w}^\top \tag{76}$$
$$\nabla^2_{\mathbf{v},\mathbf{v}}\mathrm{L} = \sigma^2\mathbf{w}\mathbf{w}^\top, \quad \nabla^2_{\mathbf{v},\mathbf{b}}\mathrm{L} = \mathbf{0} \tag{77}$$
$$\nabla^2_{\mathbf{b},\mathbf{b}}\mathrm{L} = \mathbf{w}\mathbf{w}^\top \tag{78}$$

In summary, the Hessian can be written as:

$$\mathbf{H}_{\mathrm{L}} = \begin{pmatrix} \sigma^2\mathbf{v}\mathbf{v}^\top + \mathbf{b}\mathbf{b}^\top & \overline{\delta x}\mathbf{I}_m + \sigma^2\mathbf{v}\mathbf{w}^\top & \overline{\delta}\mathbf{I}_m + \mathbf{b}\mathbf{w}^\top \\ \overline{\delta x}\mathbf{I}_m + \sigma^2\mathbf{w}\mathbf{v}^\top & \sigma^2\mathbf{w}\mathbf{w}^\top & \mathbf{0} \\ \overline{\delta}\mathbf{I}_m + \mathbf{w}\mathbf{b}^\top & \mathbf{0} & \mathbf{w}\mathbf{w}^\top \end{pmatrix} \tag{79}$$

For starters, assume $\sigma^2 = 1$. Then we need to solve the characteristic equation, i.e.,

$$\left| \begin{pmatrix} \mathbf{v}\mathbf{v}^\top + \mathbf{b}\mathbf{b}^\top - \lambda\mathbf{I}_m & \overline{\delta x}\mathbf{I}_m + \mathbf{v}\mathbf{w}^\top & \overline{\delta}\mathbf{I}_m + \mathbf{b}\mathbf{w}^\top \\ \overline{\delta x}\mathbf{I}_m + \mathbf{w}\mathbf{v}^\top & \mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m & \mathbf{0} \\ \overline{\delta}\mathbf{I}_m + \mathbf{w}\mathbf{b}^\top & \mathbf{0} & \mathbf{w}\mathbf{w}^\top - \lambda\mathbf{I}_m \end{pmatrix} \right| = 0 \tag{80}$$

3. The term $\nabla^2_{\mathbf{v},\mathbf{b}}\ell = x\mathbf{w}\mathbf{w}^\top$ can be generalized to the case of having a 2-dimensional input with coordinates x_1, x_2 with $x_2 \neq 1$ and then we would get $\nabla^2_{\mathbf{v}_1,\mathbf{v}_2}\ell = x_1x_2\mathbf{w}\mathbf{w}^\top$.

Instead of computing the determinant using the Schur complement, let us first rewrite the matrix above in a simpler form:[4]

$$\mathbf{H}_{\mathrm{L}} - \lambda \mathbf{I}_{3m} = \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \\ \mathbf{0} \end{pmatrix}^{\top} + \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{w} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{w} \end{pmatrix}^{\top} + \begin{pmatrix} -\lambda & \overline{\delta x} & \overline{\delta} \\ \overline{\delta x} & -\lambda & \mathbf{0} \\ \overline{\delta} & \mathbf{0} & -\lambda \end{pmatrix} \otimes \mathbf{I}_m \tag{81}$$

$$= \begin{pmatrix} \mathbf{v} & \mathbf{b} \\ \mathbf{w} & \mathbf{0} \\ \mathbf{0} & \mathbf{w} \end{pmatrix} \begin{pmatrix} \mathbf{v} & \mathbf{b} \\ \mathbf{w} & \mathbf{0} \\ \mathbf{0} & \mathbf{w} \end{pmatrix}^{\top} + \begin{pmatrix} -\lambda & \overline{\delta x} & \overline{\delta} \\ \overline{\delta x} & -\lambda & \mathbf{0} \\ \overline{\delta} & \mathbf{0} & -\lambda \end{pmatrix} \otimes \mathbf{I}_m \tag{82}$$

Then using the fact that $|\mathbf{A}_{m\times m} + \mathbf{U}_{m\times n}\mathbf{V}_{n\times m}^{\top}| = |\mathbf{A}| \cdot |\mathbf{I}_n + \mathbf{V}_{n\times m}^{\top}\mathbf{A}^{-1}\mathbf{U}_{m\times n}|$. Let's apply this to the above matrix, with $\mathbf{A} = \begin{pmatrix} -\lambda & \overline{\delta x} & \overline{\delta} \\ \overline{\delta x} & -\lambda & \mathbf{0} \\ \overline{\delta} & \mathbf{0} & -\lambda \end{pmatrix} \otimes \mathbf{I}_m$ and $\mathbf{U} = \mathbf{V} = \begin{pmatrix} \mathbf{v} & \mathbf{b} \\ \mathbf{w} & \mathbf{0} \\ \mathbf{0} & \mathbf{w} \end{pmatrix}$. First, by using the fact $(\mathbf{C} \otimes \mathbf{D})^{-1} = \mathbf{C}^{-1} \otimes \mathbf{D}^{-1}$, we have that

$$\mathbf{A}^{-1} = -\frac{1}{\lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)} \begin{pmatrix} \lambda^2 & \lambda\overline{\delta x} & \lambda\overline{\delta} \\ \lambda\overline{\delta x} & \lambda^2 - \overline{\delta}^2 & \overline{\delta}\ \overline{\delta x} \\ \lambda\overline{\delta} & \overline{\delta}\ \overline{\delta x} & \lambda^2 - \overline{\delta x}^2 \end{pmatrix} \otimes \mathbf{I}_m \tag{83}$$

Next, $\mathbf{A}^{-1}\mathbf{V}$ comes out to:

$$-\frac{1}{\lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)} \begin{pmatrix} \lambda^2\mathbf{v} + \lambda\overline{\delta x}\mathbf{w} & \lambda^2\mathbf{b} + \lambda\overline{\delta}\mathbf{w} \\ \lambda\overline{\delta x}\mathbf{v} + (\lambda^2 - \overline{\delta}^2)\mathbf{w} & \lambda\overline{\delta x}\mathbf{b} + \overline{\delta}\,\overline{\delta x}\mathbf{w} \\ \lambda\overline{\delta}\mathbf{v} + \overline{\delta}\,\overline{\delta x}\mathbf{w} & \lambda\overline{\delta}\mathbf{b} + (\lambda^2 - \overline{\delta x}^2)\mathbf{w} \end{pmatrix} \tag{84}$$

Further, the quadratic form $\mathbf{V}^{\top}\mathbf{A}^{-1}\mathbf{V}$ is given by $-\frac{1}{\lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)}\mathbf{B}$, with $\mathbf{B}$ being:

$$\mathbf{B} = \begin{pmatrix} \left(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2\right)\lambda^2 + 2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle\,\lambda - \overline{\delta}^2\|\mathbf{w}\|^2 & \langle\mathbf{v},\mathbf{b}\rangle\lambda^2 + \overline{\delta}\langle\mathbf{v},\mathbf{w}\rangle\lambda + \overline{\delta x}\langle\mathbf{b},\mathbf{w}\rangle\lambda + \overline{\delta}\ \overline{\delta x}\|\mathbf{w}\|^2 \\ \langle\mathbf{v},\mathbf{b}\rangle\lambda^2 + \overline{\delta}\langle\mathbf{v},\mathbf{w}\rangle\lambda + \overline{\delta x}\langle\mathbf{b},\mathbf{w}\rangle\lambda + \overline{\delta}\ \overline{\delta x}\|\mathbf{w}\|^2 & \left(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2\right)\lambda^2 + 2\overline{\delta}\langle\mathbf{w},\mathbf{b}\rangle\,\lambda - \overline{\delta x}^2\|\mathbf{w}\|^2 \end{pmatrix} \tag{85}$$

Then, $|\mathbf{I}_2 - \frac{1}{\lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)}\mathbf{B}| = \frac{(-1)^2}{\lambda^2\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)^2}|\mathbf{B} - \lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)\mathbf{I}_2|$, and the determinant of the remaining 2×2 matrix is given by:

$$\begin{aligned} &\left(\left(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2\right)\lambda^2 + 2\overline{\delta x}\langle\mathbf{w},\mathbf{v}\rangle\,\lambda - \overline{\delta}^2\|\mathbf{w}\|^2 - \lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)\right) \times \\ &\left(\left(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2\right)\lambda^2 + 2\overline{\delta}\langle\mathbf{w},\mathbf{b}\rangle\,\lambda - \overline{\delta x}^2\|\mathbf{w}\|^2 - \lambda\left(\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\right)\right) \\ &- \left(\langle\mathbf{v},\mathbf{b}\rangle\lambda^2 + \overline{\delta}\langle\mathbf{v},\mathbf{w}\rangle\lambda + \overline{\delta x}\langle\mathbf{b},\mathbf{w}\rangle\lambda + \overline{\delta}\ \overline{\delta x}\|\mathbf{w}\|^2\right)^2 \end{aligned}$$

In the case when $\sigma^2 = 1$, we have that $\overline{\delta x} = \langle\mathbf{w},\mathbf{v}\rangle - \overline{yx}$ and $\overline{\delta} = \langle\mathbf{w},\mathbf{b}\rangle - \overline{y}$.

4. If this does not work, I could start with assuming that the $\mathbf{v}$ and $\mathbf{b}$ are orthogonal.

Expanding the above equation and setting it to zero yields the following:

$$\lambda^6 + (\overline{\delta x}^2 + \overline{\delta}^2)^2\,\lambda^2 - 2(\overline{\delta x}^2 + \overline{\delta}^2)\,\lambda^4 - (\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2)\lambda^5 \tag{86}$$

$$- 2(\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle)\lambda^4 + (\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2)\,\lambda^3 \tag{87}$$

$$+ 2(\overline{\delta x}^2 + \overline{\delta}^2)(\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle)\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2\lambda \tag{88}$$

$$+ (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2)\lambda^4 + 4\overline{\delta x}\,\overline{\delta}\langle \mathbf{w}, \mathbf{v}\rangle\langle \mathbf{w}, \mathbf{b}\rangle\lambda^2 \cancel{+\overline{\delta x}^2\,\overline{\delta}^2\|\mathbf{w}\|^4} \tag{89}$$

$$+ 2\overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\lambda^3 + 2\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\lambda^3 \tag{90}$$

$$- 2\overline{\delta x}^3\langle \mathbf{w}, \mathbf{v}\rangle\|\mathbf{w}\|^2\lambda - 2\overline{\delta}^3\langle \mathbf{w}, \mathbf{b}\rangle\|\mathbf{w}\|^2\lambda \tag{91}$$

$$- (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2\lambda^2 - (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2\lambda^2 \tag{92}$$

$$- \big(\langle \mathbf{v}, \mathbf{b}\rangle^2\lambda^4 + \langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle^2\lambda^2 \cancel{+\overline{\delta}^2\,\overline{\delta x}^2\|\mathbf{w}\|^4} \tag{93}$$

$$+ 2\langle \mathbf{v}, \mathbf{b}\rangle\,\langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle\lambda^3 + 2\langle \mathbf{v}, \mathbf{b}\rangle\overline{\delta}\,\overline{\delta x}\|\mathbf{w}\|^2\lambda^2 \tag{94}$$

$$+ 2\langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle\overline{\delta}\,\overline{\delta x}\|\mathbf{w}\|^2\lambda\big) \tag{95}$$

$$= 0 \tag{96}$$

The constant terms get cancelled, and rewriting the above expression by collecting the coefficients of various degree terms separately gives us:

$$\begin{aligned}
&\lambda^6 - (\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2)\lambda^5 \\
&+ \Big[-2(\overline{\delta x}^2 + \overline{\delta}^2) - 2(\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle) + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2) - \langle \mathbf{v}, \mathbf{b}\rangle^2\Big]\,\lambda^4 \\
&+ \Big[(\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2) + 2\overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + 2\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\Big]\,\lambda^3 \\
&- \big[2\langle \mathbf{v}, \mathbf{b}\rangle\,\langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle\big]\,\lambda^3 \\
&+ \Big[(\overline{\delta x}^2 + \overline{\delta}^2)^2 + 2(\overline{\delta x}^2 + \overline{\delta}^2)(\overline{\delta x}\langle \mathbf{w}, \mathbf{v}\rangle + \overline{\delta}\langle \mathbf{w}, \mathbf{b}\rangle) + 4\overline{\delta x}\,\overline{\delta}\langle \mathbf{w}, \mathbf{v}\rangle\langle \mathbf{w}, \mathbf{b}\rangle\Big]\,\lambda^2 \\
&- \Big[(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2 + (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2 + \langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle^2 + 2\langle \mathbf{v}, \mathbf{b}\rangle\overline{\delta}\,\overline{\delta x}\|\mathbf{w}\|^2\Big]\,\lambda^2 \\
&- \Big[(\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2 + 2\overline{\delta x}^3\langle \mathbf{w}, \mathbf{v}\rangle\|\mathbf{w}\|^2 + 2\overline{\delta}^3\langle \mathbf{w}, \mathbf{b}\rangle\|\mathbf{w}\|^2 + 2\langle \overline{\delta}\mathbf{v} + \overline{\delta x}\mathbf{b}, \mathbf{w}\rangle\overline{\delta}\,\overline{\delta x}\|\mathbf{w}\|^2\Big]\,\lambda^1 \\
&= 0
\end{aligned}$$

Let us try to first solve the above equation in a simpler setting. To do so, we assume: $\overline{yx} = 0$ and $\overline{y} = 0$. This means that: $\overline{\delta x} = \langle \mathbf{w}, \mathbf{v}\rangle$ and $\overline{\delta} = \langle \mathbf{w}, \mathbf{b}\rangle$. We get the following simplified form:

$$
\begin{aligned}
&\lambda^6 - (\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2)\lambda^5 \\
&+ \left[-2(\overline{\delta x}^2 + \overline{\delta}^2) - 2(\overline{\delta x}^2 + \overline{\delta}^2) + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2) - \langle \mathbf{v}, \mathbf{b}\rangle^2\right] \lambda^4 \\
&+ \left[(\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2) + 2\overline{\delta}^2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + 2\overline{\delta x}^2(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\right] \lambda^3 \\
&- \left[4\langle \mathbf{v}, \mathbf{b}\rangle\, \overline{\delta x}\, \overline{\delta}\right] \lambda^3 \\
&+ \left[(\overline{\delta x}^2 + \overline{\delta}^2)^2 + 2(\overline{\delta x}^2 + \overline{\delta}^2)^2 + \cancel{4\overline{\delta x}^2\, \overline{\delta}^2}\right] \lambda^2 \\
&- \left[(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2 + (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2 + \cancel{4\overline{\delta x}^2\, \overline{\delta}^2} + 2\langle \mathbf{v}, \mathbf{b}\rangle \overline{\delta}\, \overline{\delta x}\|\mathbf{w}\|^2\right] \lambda^2 \\
&- \left[(\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2 + 2\overline{\delta x}^4\|\mathbf{w}\|^2 + 2\overline{\delta}^4\|\mathbf{w}\|^2 + 4\overline{\delta}^2\, \overline{\delta x}^2\|\mathbf{w}\|^2\right] \lambda^1 \\
&= 0
\end{aligned}
$$

which can be further simplified as,

$$
\begin{aligned}
&\lambda^6 - (\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2)\lambda^5 \\
&+ \left[-4(\overline{\delta x}^2 + \overline{\delta}^2) + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2) - \langle \mathbf{v}, \mathbf{b}\rangle^2\right] \lambda^4 \\
&+ \left[(\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2) + 2\overline{\delta}^2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + 2\overline{\delta x}^2(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2) - 4\langle \mathbf{v}, \mathbf{b}\rangle\, \overline{\delta x}\, \overline{\delta}\right] \lambda^3 \\
&+ \left[3(\overline{\delta x}^2 + \overline{\delta}^2)^2 - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2 - (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2 - 2\langle \mathbf{v}, \mathbf{b}\rangle\overline{\delta}\, \overline{\delta x}\|\mathbf{w}\|^2\right] \lambda^2 \\
&- 3(\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2\lambda^1 \\
&= 0
\end{aligned}
$$

We effectively have a quintic equation above and quintic equations, in general, do not have roots in radicals due to the Abel-Ruffini theorem. Although this specific equation might still be solvable in radicals. Let's assume that the underlying quintic is of the form:

$$(\lambda^2 + A\lambda + B)(\lambda^3 + C\lambda^2 + D\lambda + E)$$

which we can be expressed as:

$$\lambda^5 + (A + C)\lambda^4 + (B + AC + D)\lambda^3 + (E + AD + BC)\lambda^2 + (AE + BD)\lambda + BE$$

Matching this with the above equation with λ factored out, we get the following series of equations:

$$
\begin{aligned}
A + C &= -(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2) \\
B + AC + D &= -4(\overline{\delta x}^2 + \overline{\delta}^2) + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2) - \langle \mathbf{v}, \mathbf{b}\rangle^2 \\
E + AD + BC &= (\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2) + 2\overline{\delta}^2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + 2\overline{\delta x}^2(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2) - 4\langle \mathbf{v}, \mathbf{b}\rangle\, \overline{\delta x}\, \overline{\delta} \\
AE + BD &= 3(\overline{\delta x}^2 + \overline{\delta}^2)^2 - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2 - (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2 - 2\langle \mathbf{v}, \mathbf{b}\rangle\overline{\delta}\, \overline{\delta x}\|\mathbf{w}\|^2 \\
BE &= -3(\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2
\end{aligned}
$$

And to remind again the above quintic is:

$$
\begin{aligned}
&\lambda^5 - (\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{w}\|^2)\lambda^4 \\
&+ \left[-4(\overline{\delta x}^2 + \overline{\delta}^2) + (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\|\mathbf{b}\| + \|\mathbf{w}\|^2) - \langle \mathbf{v}, \mathbf{b}\rangle^2\right] \lambda^3 \\
&+ \left[(\overline{\delta x}^2 + \overline{\delta}^2)(\|\mathbf{v}\|^2 + \|\mathbf{b}\|^2 + 3\|\mathbf{w}\|^2) + 2\overline{\delta}^2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + 2\overline{\delta x}^2(\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2) - 4\langle \mathbf{v}, \mathbf{b}\rangle\, \overline{\delta x}\, \overline{\delta}\right] \lambda^2 \\
&+ \left[3(\overline{\delta x}^2 + \overline{\delta}^2)^2 - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta x}^2 - (\|\mathbf{b}\|^2 + \|\mathbf{w}\|^2)\|\mathbf{w}\|^2\overline{\delta}^2 - 2\langle \mathbf{v}, \mathbf{b}\rangle\overline{\delta}\,\overline{\delta x}\|\mathbf{w}\|^2\right] \lambda^1 \\
&- 3(\overline{\delta x}^2 + \overline{\delta}^2)^2\|\mathbf{w}\|^2 \\
&= 0
\end{aligned}
$$

Based on the above equations, let us employ two additional assumptions, i.e., $\langle \mathbf{v}, \mathbf{b}\rangle = 0$ and $\|\mathbf{v}\| = \|\mathbf{b}\|$. We have as the choice of the coefficients::

$$
\begin{aligned}
A &= -(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) && (97)\\
B &= -3(\overline{\delta x}^2 + \overline{\delta}^2) && (98)\\
C &= -(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) && (99)\\
D &= -(\overline{\delta x}^2 + \overline{\delta}^2) && (100)\\
E &= (\overline{\delta x}^2 + \overline{\delta}^2)\|\mathbf{w}\|^2 && (101)
\end{aligned}
$$

We get as a factored quadratic:

$$\lambda^2 - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\lambda - 3(\overline{\delta x}^2 + \overline{\delta}^2) = 0$$

$$\lambda = \frac{1}{2}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) \pm \frac{1}{2}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 12(\overline{\delta x}^2 + \overline{\delta}^2)} \tag{102}$$

This strictly generalizes the case without bias, where we had the solution:

$$\lambda = \frac{1}{2}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) \pm \frac{1}{2}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 12\overline{\delta x}^2} \tag{103}$$

Remember, under our assumptions $\overline{\delta x} = \langle \mathbf{w}, \mathbf{v}\rangle$ and $\overline{\delta} = \langle \mathbf{w}, \mathbf{b}\rangle$. Hence this forms for an interesting generalization in the case with multi-dimensional input, with term in the square root being the sum of squares of the inner-product of the second layer weight $\mathbf{w}$ and the columns of the first layer matrix.

Solving the cubic. Other than the solutions from the above quadratic, we should also check for solutions in:

$$\lambda^3 - (\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)\lambda^2 - (\overline{\delta x}^2 + \overline{\delta}^2)\lambda + (\overline{\delta x}^2 + \overline{\delta}^2)\|\mathbf{w}\|^2 = 0$$

We can put this into depressed form, i.e., $t^3 + pt + q = 0$, by making the substitution:

$$\lambda = t + \frac{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)}{3}\,.$$

Here, p and q come out to be:

$$p = -(\overline{\delta x}^2 + \overline{\delta}^2) - \frac{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2}{3} \tag{104}$$

$$q = \frac{1}{27}\left[-2(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^3 - 9(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)(\overline{\delta x}^2 + \overline{\delta}^2) + 27(\overline{\delta x}^2 + \overline{\delta}^2)\|\mathbf{w}\|^2\right] \tag{105}$$

The coefficient q can be further simplified as:

$$q = -\frac{2}{27}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^3 - \frac{1}{3}\|\mathbf{v}\|^2(\overline{\delta x}^2 + \overline{\delta}^2) + \frac{2}{3}\|\mathbf{w}\|^2(\overline{\delta x}^2 + \overline{\delta}^2) \tag{106}$$

Using Viète's trigonometric expression of the roots in three-real roots case, we have that the solutions to t are given by t_k for $k = 0, 1, 2$:

$$t_k = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right) - k\frac{2\pi}{3}\right) \tag{107}$$

Since $cos \in [-1, 1]$, we can upper and lower bound the above solutions to the cubic as:

$$\frac{1}{3}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) - t' \leq \lambda_{0,1,2} \leq \frac{1}{3}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2) + t'$$

with $t' = 2\sqrt{\frac{1}{3}(\overline{\delta x}^2 + \overline{\delta}^2) + \frac{1}{9}(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2} = \frac{2}{3}\sqrt{(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2)^2 + 3(\overline{\delta x}^2 + \overline{\delta}^2)}$.

Notice, that $\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2$ is $1/2 \cdot \text{Tr}(\mathbf{H}_\text{L})$. So, the two solutions to the quadratic on the page before are centered at $1/4 \cdot \text{Tr}(\mathbf{H}_\text{L})$, while the three solutions to the cubic get centered around $1/6 \cdot \text{Tr}(\mathbf{H}_\text{L})$.