



Mathematical  
Institute

# Random DNN training landscape: Hessian eigen- values

THEORIES OF DEEP LEARNING: C6.5, VIDEO 8

*Prof. Jared Tanner*  
*Mathematical Institute*  
*University of Oxford*

Oxford  
Mathematics



# Loss function for a simple fully connected two layer NN

Sum of squares loss

Consider a data set  $X \in \mathbb{R}^{n \times m}$  of  $m$  data entries in  $\mathbb{R}^n$ , associated target outputs (such as labels)  $Y \in \mathbb{R}^{n_2 \times m}$  (for simplicity we let  $n_2 = n$ ), and (very) simple two layer net:

$$h_1 = \phi(W^{(1)}x_0) \quad \text{note, no bias, and } \phi(\cdot) = \max(0, \cdot)$$

$$h_2 = W^{(2)}h_1 \quad \text{note, no bias or nonlinear activation.}$$

The output of the net is  $H(x_\mu; \theta) = \hat{y}_\mu$  and we measure the value of the net through the average sum of squares:

$$\mathcal{L} = (2m)^{-1} \sum_{\mu=1}^m \sum_{i=1}^n (\hat{y}_{i,\mu} - y_{i,\mu})^2$$

and define a weighted loss accuracy as  $\epsilon = n^{-1}\mathcal{L}$ .

## Hessian for two layer net (without activation)

Omitting diagonal nonlinear activation matrices.

Let  $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$  be the error in the  $i^{\text{th}}$  entry of the output for data entry indexed by  $\mu$ , and  $\theta = \{W^{(1)}, W^{(2)}\} \in \mathbb{R}^{2n^2}$  be the net parameters, then the hessian of the loss function has entries

$$H_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial \theta_\alpha \partial \theta_\beta} =: H_0 + H_1$$

with positive semi-definite and error dependent components:

$$[H_0]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^m \sum_{i=1}^n \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\alpha} \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\beta} = m^{-1} [JJ^T]_{\alpha,\beta}$$

$$[H_1]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^m \sum_{i=1}^n e_{i,\mu} \frac{\partial^2 \hat{y}_{i,\mu}}{\partial \theta_\alpha \partial \theta_\beta}.$$

There are  $mn$  data entries and  $2n^2$  NN parameters, with  $\tau = 2n/m$  the relative over ( $\tau > 1$ ) or under ( $\tau < 1$ ) parameterisation.

# Loss function landscape through Hessian eigenvalues

Local shape of loss landscape



Functions, say  $\mathcal{L}$ , which have Hessians that are:

- ▶ positive definite (all positive eigenvalues) are convex and have a single global minima and unique minimiser,
- ▶ positive semi-definite have single global minima but non-unique minimiser due to the null-space
- ▶ indefinite (positive and negative eigenvalues) are non-convex and may be a complicated landscape with multiple local minimisers.

For the simple two layer network we considered the network has Hessian  $H = H_0 + H_1$  with  $H_0$  positive semidefinite and of size independent of the error, while  $H_1$  is indefinite with magnitude depending on the size of  $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ .

One can interpret properties of the landscape through the Hessian by considering simplified models:

- ▶ The weights are i.i.d. random normal variable,
- ▶ The data are i.i.d. random variables,
- ▶ The residuals  $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$  are normal random variables, say  $\mathcal{N}(0, 2\epsilon)$  with  $\epsilon = n^{-1}\mathcal{L}$  (which also allows the gradient to vanish as  $m, n \rightarrow \infty$  while  $m/n$  remains fixed; the focus is on fixed points where the gradient is zero),
- ▶ The matrices  $H_0$  and  $H_1$  are *freely independent* which allows us to compute the spectra of  $H_0 + H_1$  from their individual spectra.

<http://proceedings.mlr.press/v70/pennington17a.html>

# Wigner and Wishart distributions

Deterministic eigenvalue distributions of random matrices: the large  $n, p$  limit.

Wigner matrices, entries drawn  $\mathcal{N}(0, \sigma^2)$ , have eigenvalues drawn from the semi-circle law:

$$\rho_{sc}(\lambda) = \begin{cases} \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - \lambda^2} & \text{if } |\lambda| \leq 2\sigma \\ 0 & \text{otherwise} \end{cases}$$

Wishart matrices,  $X = JJ^T$  product of  $J \in \mathbb{R}^{n \times p}$  drawn  $\mathcal{N}(0, \sigma^2/p)$  have eigenvalues drawn from the Marchenko-Pastur distribution:

$$\rho_{MP}(\lambda) = \begin{cases} \rho(\lambda) & \text{if } \tau = n/p < 1 \\ (1 - \tau^{-1})\delta(\lambda) + \rho(\lambda) & \text{otherwise} \end{cases}$$

where  $\rho(\lambda) := (2\pi\lambda\sigma\tau)^{-1} \sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)}$  for  $\lambda \in [\lambda_-, \lambda_+]$  and  $\lambda_{\pm} := \sigma(1 \pm \sqrt{\tau})^2$ .

# Stieltjes and $\mathcal{R}$ Transforms of probability distributions

Method to compute the spectrum under addition.

The probability distribution of the sum of two (freely independent) random matrix distributions can be calculated using the transforms:

## Stieltjes and $\mathcal{R}$ Transforms

For  $z \in \mathbb{C}/\mathbb{R}$  the Stieltjes Transform,  $G_\rho(z)$ , of a probability distribution and its inverse are given by

$$G_\rho(z) = \int_{\mathbb{R}} \frac{\rho(t)}{z-t} dt \quad \text{and} \quad \rho(\lambda) = -\pi^{-1} \lim_{\epsilon \rightarrow 0_+} \text{Imag}(G_\rho(\lambda + i\epsilon)).$$

The Stieltjes and  $\mathcal{R}$  Transform of  $\rho$  are related by the solutions of  $\mathcal{R}_\rho(G_\rho(z)) + 1/G_\rho(z) = z$  and has the property that if  $\rho_1$  and  $\rho_2$  are freely independent then  $\mathcal{R}_{\rho_1+\rho_2} = \mathcal{R}_{\rho_1} + \mathcal{R}_{\rho_2}$ .

<https://terrytao.wordpress.com/tag/stieltjes-transform-method/>

## Recall the Hessian for two layer net (without activation)

Stieltjes and  $\mathcal{R}$  Transform for joint spectra

Let  $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$  be the error in the  $i^{\text{th}}$  entry of the output for data entry indexed by  $\mu$ , and  $\theta = \{W^{(1)}, W^{(2)}\} \in \mathbb{R}^{2n^2}$  be the net parameters, then the hessian of the loss function has entries

$$H_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial \theta_\alpha \partial \theta_\beta} =: H_0 + H_1$$

with positive semi-definite and error dependent components:

$$[H_0]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^m \sum_{i=1}^n \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\alpha} \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\beta} = m^{-1} [JJ^T]_{\alpha,\beta}$$

$$[H_1]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^m \sum_{i=1}^n e_{i,\mu} \frac{\partial^2 \hat{y}_{i,\mu}}{\partial \theta_\alpha \partial \theta_\beta}.$$

Where we assumed that  $H_0$  and  $H_1$  can be modelled as being drawn from Wishart and Wigner distributions respectively.

Using the Pennington model ( $\tau = \phi = 2n/m$  and  $\epsilon = n^{-1}\mathcal{L}$ ) we have  $\rho_{H_0}(\lambda) = \rho_{MP}(\lambda; 1, \tau)$  and  $\rho_{H_1}(\lambda) = \rho_{SC}(\lambda; \sqrt{2\epsilon})$ .

Their  $\mathcal{R}$  transforms are respectively

$$\mathcal{R}_{H_0} = \frac{1}{1 - z\tau} \quad \text{and} \quad \mathcal{R}_{H_1} = 2\epsilon z,$$

from which follows the probability distribution,  $\rho_H(\lambda; \epsilon, \tau)$ :

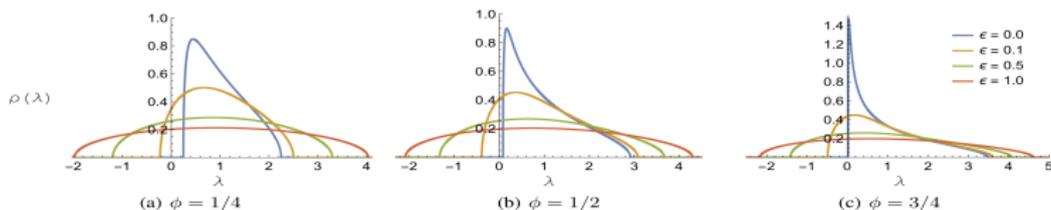


Figure 1. Spectral distributions of the Wishart + Wigner approximation of the Hessian for three different ratios of parameters to data points,  $\phi$ . As the energy  $\epsilon$  of the critical point increases, the spectrum becomes more semicircular and negative eigenvalues emerge.

<http://proceedings.mlr.press/v70/pennington17a.html>

# Fraction of negative eigenvalues (Pennington et al. 17')

Breakpoint dependence on  $\epsilon_c$  and oversampling  $\tau$

Consider the fraction of negative eigenvalues of  $\rho_H(\lambda)$ :

$$\alpha(\epsilon, \tau) := \int_{-\infty}^0 \rho_H(\lambda; \epsilon, \tau) d\lambda.$$

Fraction of negative eigenvalues (without ReLU)

For  $\rho_H(\lambda)$  modelling the Hessian of the two layer net, when  $\alpha$  is small it is well approximated by

$$\alpha(\epsilon, \tau) \approx \alpha_0(\tau) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2}$$

where

$$\epsilon_c = \frac{1}{16} (1 - 20\tau - 8\tau^2 + (1 + 8\tau)^{3/2}).$$

<http://proceedings.mlr.press/v70/pennington17a.html>

# The two layer ReLU net (Pennington et al. 17')

Now including a ReLU nonlinear activation

The introduction of the ReLU nonlinear activation changes the Hessian, roughly setting to zero half of the entries and generating a block off-diagonal structure in  $H_1$  with  $\mathcal{R}_{H_1}(z) = \frac{\epsilon \tau z}{2 - \epsilon \tau^2 z^2}$ .

Continuing to model  $H_0$  as Wishart (less clear an assumption):

Fraction of negative eigenvalues (with ReLU)

For  $\rho_H(\lambda)$  modelling the Hessian of the two layer net, when  $\alpha$  is small it is well approximated by

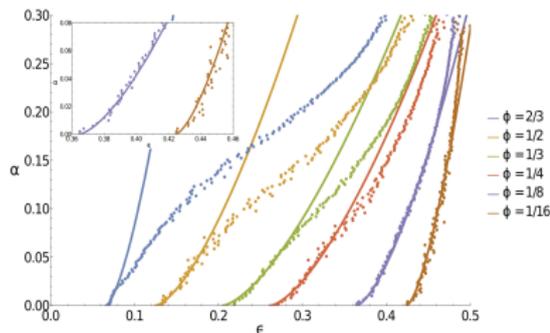
$$\alpha(\epsilon, \tau) \approx \tilde{\alpha}_0(\tau) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2} \quad \text{where}$$

$$\epsilon_c = \frac{\sigma^2(27 - 18\xi - \xi^2 + 8\xi^{3/2})}{32\tau(1 - \tau)^3}, \quad \text{with} \quad \xi = 1 + 16\tau - 8\tau^2.$$

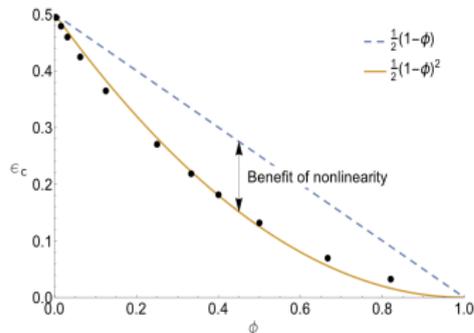
<http://proceedings.mlr.press/v70/pennington17a.html>

# Empirical values of $\epsilon_c$ and $\alpha$ (Pennington et al. 17')

Match of empirical and analytical calculations



(a) Index of critical points versus energy



(b) Energy of minimizers versus parameters/data points

Figure 6. Empirical observations of the distribution of critical points in single-hidden-layer tanh networks with varying ratios of parameters to data points,  $\phi$ . (a) Each point represents the mean energy of critical points with index  $\alpha$ , averaged over  $\sim 200$  training runs. Solid lines are best fit curves for small  $\alpha \approx \alpha_0 |\epsilon - \epsilon_c|^{3/2}$ . The good agreement (emphasized in the inset, which shows the behavior for small  $\alpha$ ) provides support for our theoretical prediction of the  $3/2$  scaling. (b) The best fit value of  $\epsilon_c$  from (a) versus  $\phi$ . A surprisingly good fit is obtained with  $\epsilon_c = \frac{1}{2}(1 - \phi)^2$ . Linear networks obey  $\epsilon_c = \frac{1}{2}(1 - \phi)$ . The difference between the curves shows the benefit obtained from using a nonlinear activation function.

<http://proceedings.mlr.press/v70/pennington17a.html>