Mathematical Institute

Topology of the loss landscape: global and local structures

Theories of Deep Learning: C6.5,
Lecture / Video 9

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford Mathematics

UNIVERSITY OF OXFORD
Mathematical Institute

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:

$$\begin{aligned} h_1 &= \phi(W^{(1)} x_0) \quad \text{note, no bias, and } \phi(\cdot) = \max(0, \cdot) \\ h_2 &= W^{(2)} h_1 \quad \text{note, no bias or nonlinear activation.} \end{aligned}$$

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:

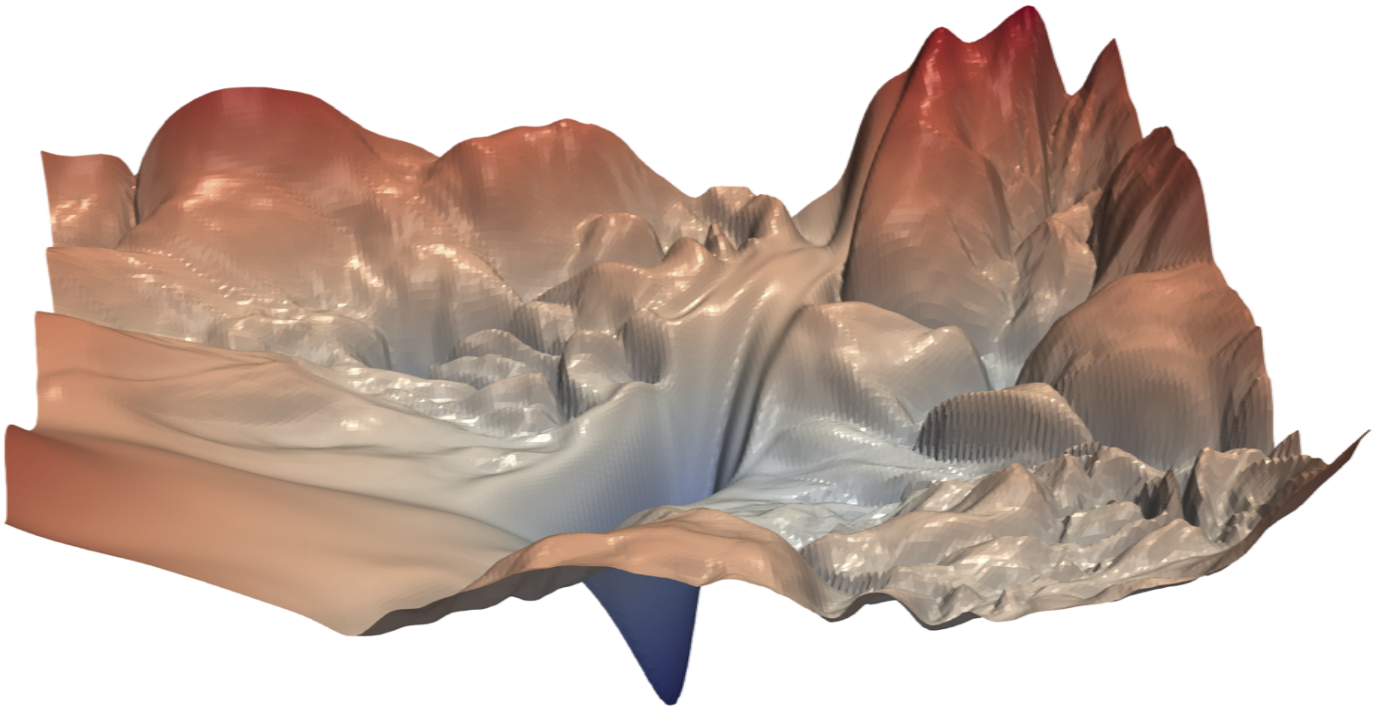
$$\mathcal{L} = (2m)^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n} (\hat{y}_{i,\mu} - y_{i,\mu})^2$$

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

Loss landscape example: 56 layers fully connected (Li et al. 18')

Loss landscapes of DNNs are typically non-convex

UNIVERSITY OF OXFORD
Mathematical Institute



http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf

Topology of loss landscape (Freeman et al. 16')

Number of connected components

Mathematical Institute

Consider our loss function: $\mathcal{L}(\theta; X, Y) = n^{-1} \sum_{\mu=1}^{n} l(\theta; x_\mu, y_\mu)$ and its associated level set

$$\Omega_{\mathcal{L}}(\lambda) = \{\theta : \ \mathcal{L}(\theta; X, Y) \leq \lambda\}$$

Of particular interest are the number of connected components, say N_λ, in $\Omega_{\mathcal{L}}(\lambda)$. If $N_\lambda = 1$ for all λ then $\mathcal{L}(\theta; X, Y)$ has no isolated local minima and any descent method can obtain a global minima.

If $N_\lambda > 1$ there may be "spurious valleys" in which the minima in the connected component does not achieve the global minima. `https://arxiv.org/pdf/1611.01540.pdf`

Topology of loss landscape (Freeman et al. 16')

There are datasets for which ReLU has a complex landscape

UNIVERSITY OF OXFORD
Mathematical Institute

Linear network: single component

Let $H(x;\theta)$ be an L layer net given by $h^{(\ell)} = W^{(\ell)} h^{(\ell-1)}$ with $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, then if $n_\ell > \min(n_0, n_L)$ for $0 < \ell < L$, the sum of squares loss function has a single connected component

ReLU network: multiple components

Let $H(x;\theta)$ be an L layer net given by $h^{(\ell)} = \sigma(W^{(\ell)} h^{(\ell-1)})$ with $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ and $\sigma(\cdot) = \max(0, \cdot)$, then for any choice of n_ℓ there is a distribution of data (X, Y) such that there are more than one single connected component.

Topology of loss landscape: (Venturi et al. 16')

Over parameterisation can generate a single connected component

UNIVERSITY OF OXFORD
Mathematical Institute

ReLu activation network: nearly connected

Consider a 2 layer ReLu network $H(x,\theta) = W^{(2)}\sigma(W^{(1)}x)$ with $W^{(1)} \in \mathbb{R}^{m\times n}$ and $W^{(2)} \in \mathbb{R}^m$, then for any two parameters θ_1 and θ_2 with $\mathcal{L}(\theta_i) \leq \lambda$ for $i = 1, 2$, then there is a path $\gamma(t)$ between θ_1 and θ_2 such that $\mathcal{L}(\theta_{\gamma(t)}) \leq \max(\lambda, m^{-1/n})$.

quadratic activation network: single component

Let $H(x,\theta)$ be an L layer net given by $h^{(\ell)} = \sigma(W^{(\ell)}h^{(\ell-1)})$ with $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ and quadratic activation $\sigma(z) = z^2$, then once the number of parameters $n_\ell \geq 3N^{2^\ell}$ where N is the number of data entries, then the sum of squares loss function has a single connected component. For the two layer case with a single quadratic activation this simplifies to $n > 2N$.

`https://arxiv.org/pdf/1802.06384.pdf`

Hessian for two layer net (without activation)

Omitting diagonal nonlinear activation matrices.

Mathematical Institute

Let $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ be the error in the i^{th} entry of the output for data entry indexed by μ, 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.

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 $m, n \to \infty$ 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 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.

University of Oxford
Mathematical Institute

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

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

The Stieltjes and $\mathcal{R}$ Transform of ρ are related by the solutions of $\mathcal{R}_\rho(G_\rho(z)) + 1/G_\rho(z) = z$ and has the property that if ρ_1 and ρ_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/`

Let $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ be the error in the i^{th} entry of the output for data entry indexed by μ, 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:

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

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

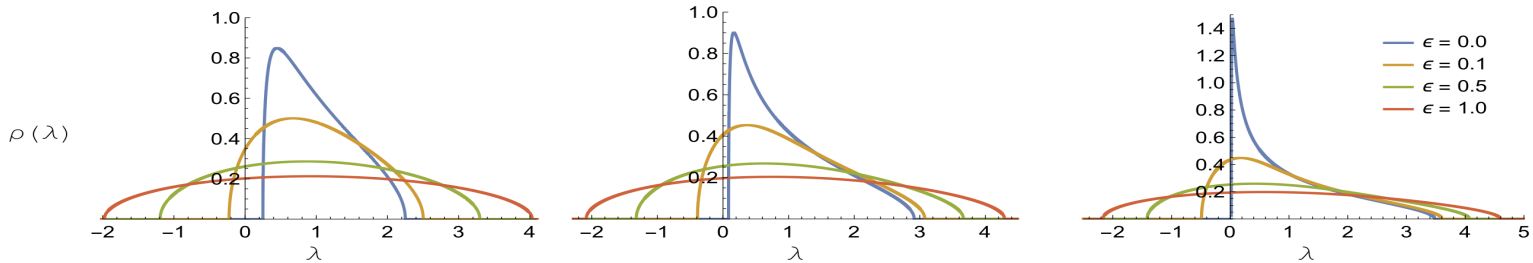
Where we assumed that H_0 and H_1 can be modelled as being drawn from Wishart and Wigner distributions respectively.

Landscape via random matrix theory (Pennington et al. 17')

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, $\rho_H(\lambda; \epsilon, \tau)$:



(a) $\phi = 1/4$ (b) $\phi = 1/2$ (c) $\phi = 3/4$

Figure 1. Spectral distributions of the Wishart + Wigner approximation of the Hessian for three different ratios of parameters to data points, ϕ. As the energy ϵ 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 ϵ_c and oversampling τ

Mathematical Institute

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 α 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

UNIVERSITY OF OXFORD
Mathematical Institute

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 $\mathcal{R}_{H1}(z) = \frac{\epsilon\tau z}{2-\epsilon\tau^2 z^2}$.

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 α is small it is well approximated by

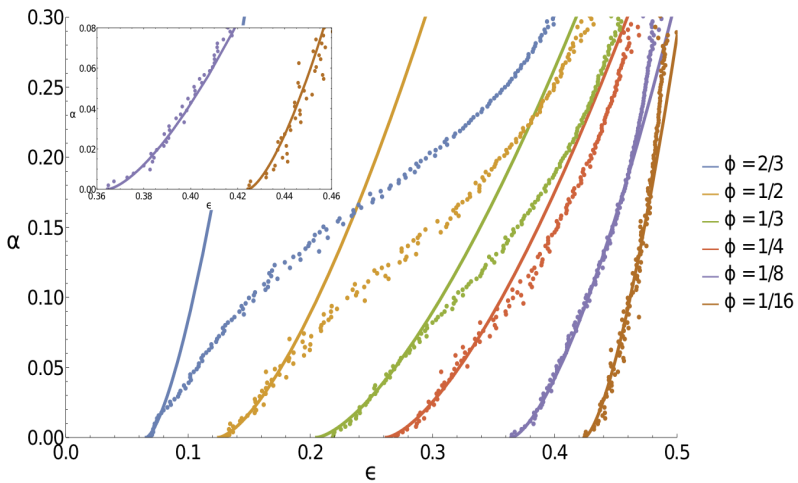
$$\alpha(\epsilon,\tau) \approx \tilde{\alpha}_0(\tau) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2} \quad \text{where}$$

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

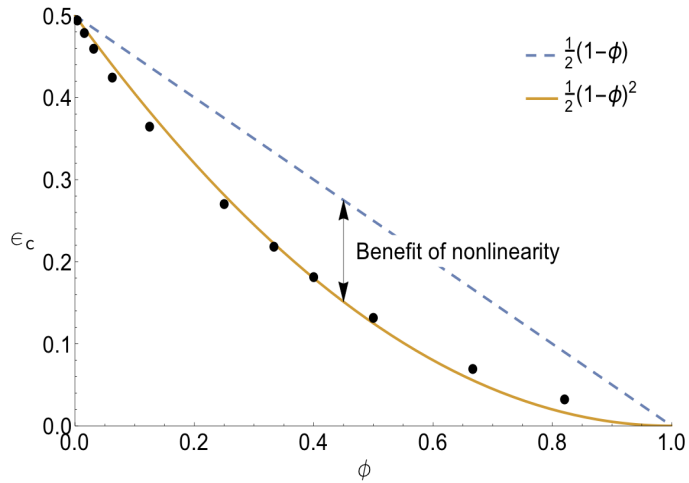
`http://proceedings.mlr.press/v70/pennington17a.html`

Empirical values of ϵ_c and α (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, ϕ. (a) Each point represents the mean energy of critical points with index α, averaged over ~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 α) provides support for our theoretical prediction of the $3/2$ scaling. (b) The best fit value of ϵ_c from (a) versus ϕ. 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

University of Oxford
Mathematical Institute

Dimension of global minimizer submanifold

Let $H(x;\theta)$ be a DNN from $\mathbb{R}^n$ to $\mathbb{R}^r$ with smooth nonlinear activation $\phi(\cdot)$, let the loss function over d distinct data elements be defined as

$$\mathcal{L} = (2m)^{-1} \sum_{\mu=1}^{d} \|H(x_\mu;\theta) - y_\mu)\|_2^2,$$

and let $\Omega^*_{\mathcal{L}}(0) = \{\theta : \ \mathcal{L}(\theta; X, Y) = 0\}$ be the set of weight and bias trainable parameters for which the DNN exactly fits the d data elements. Then, subject to possibly arbitrarily small perturbation, the set $\Omega^*_{\mathcal{L}}(0)$ is a smooth $(d - rn)-$dimensional submanifold (possibly empty) of $\mathbb{R}^d$.

https://epubs.siam.org/doi/pdf/10.1137/19M1308943

DNN loss landscape summary

Structure of the loss landscape dimensionality dependence

Mathematical Institute

- Loss landscapes for DNNs can be non-convex and hence difficult to optimise.
- The number of components of a loss landscape level curve can be analysed, and in some settings has a single component greatly aiding its optimisation.
- Increasing width of a DNN can improve the loss landscape.
- The local shape of random nets can be analysed, showing that when near a minima the Hessian has only non-negative eigenvalues.
- When the amount of data exceeds the product of the input and output dimensions, DNNs with smooth non-linear activations which exactly fit the data, have smooth manifold of a known dimension.