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

University of Oxford
Mathematical Institute

Oxford Mathematics

Mathematical Institute

With the data $\{(x_\mu, y_\mu)\}_{\mu=1}^{m}$ where $y_\mu \in \mathbb{R}$, and network $H(x;\theta_0)$ with $H(x_\mu;\theta_0) = \hat{y}_\mu$; if we linearize the network $H(x;\theta)$ in θ about the current θ_0 and

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

we can exactly express the solution as that of a linear system. Let $x = (x_1 \cdots x_m)^T$, and treating everything in vector notation, the linear approximation of the network is $H(x;\theta) = H(x;\theta_0) + \nabla_\theta H(x;\theta)|_{\theta_0}(\theta - \theta_0) + \mathcal{O}(\|\theta - \theta_0\|^2)$. The linearization matrix J_0 can then be written as $J_0 = \nabla_\theta H(x;\theta)|_{\theta_0} \in \mathbb{R}^{m \times p}$ where its i^{th} row $\nabla_\theta H(x;\theta)|_{\theta_0}$ has entries $\left(\nabla_{\theta_1} H(x^{(i)};\theta) \cdots \nabla_{\theta_p} H(x^{(i)};\theta)\right)$.

The loss function for the linearized approximation to the network $H(x;\theta)$ about θ_0 is then given by

$$\tilde{\mathcal{L}}(\theta) = \|\hat{y} - y + J_0(\theta - \theta_0)\|_2^2.$$

Once the number of network parameters p exceeds the amount of data pairs m, the loss $\tilde{\mathcal{L}}(\theta)$ can be exactly set to zero, provided J_0 is full rank. In the $p > m$ regime there are many solutions to this underdetermined system. A natural solution is the Moore-Penrose pseudo-inverse where we start with $y - \hat{y} = J_0(\theta - \theta_0)$, multiply from the left by J_0^T and use the pseudo-inverse of the matrix $J_0^T J_0 \in \mathbb{R}^{p\times p}$ which is not full rank as $p > m$; we denote this $\theta_1 = \theta_0 + J_0^+(y - \hat{y})$. In optimization this is called the Gauss-Newton method.

If J_0 is full rank and $p > m$, there is a θ with $y - \hat{y} = J_0(\theta - \theta_0)$, but there are many such θ as J_0 has a null-space (kernel) of dimension $p - m$ and anything in this null-space can be added to θ without modifying the solution. This indicates that the linearized loss is "flat" in $p - m$ dimensions and as p grows the optimization landscape appears to be increasingly flat. Selecting the pseudo-inverse solution $\theta = \theta_0 + J_0^{+}(y - \hat{y})$ has the benefit that it minimizes $\|\theta - \theta_0\|_2$ amongst all solutions; this is a form of implicit regularization where in effect we have added a penalty on $\|\theta - \theta_0\|_2$ to the loss function, though it isn't added explicitly, rather it appears through the choice of θ amongst its many solutions.

Let J_0 be full rank and $\sigma_{min}(J_0 J_0^T)$, smallest nonzero singular value of J_0, be independent of m and p (true for most $\phi(\cdot)$ but not proven here), then using the bound

$$\|\theta - \theta_0\|_2 = \|J_0^+(y - \hat{y})\|_2 \leq \sigma_{min}^{-1/2}(J_0 J_0^T)\|y - \hat{y}\|_2$$

we can start to make some observations of the role of the dimensions. Let us consider the entries in y and $\hat{y}$ to be independent of the dimensions, then $\|y - \hat{y}\|_2 = \mathcal{O}(m^{1/2})$ due to $y, \hat{y} \in \mathbb{R}^m$ and $\|\theta - \theta_0\|_2 \leq \mathcal{O}(m^{1/2}\sigma_{min}^{-1/2}(J_0 J_0^T))$
Note that the right hand side of the above bound is independent of p, while $\theta - \theta_0 \in \mathbb{R}^p$ so it must be that if their entries are of similar magnitude then $(\theta - \theta_0)(i) = \mathcal{O}(p^{-1/2})$ which tells us that in some sense for large p there is a lack of training.

Consider the difference between the network $H(x;\theta)$ and its linear approximation $\hat{y} + J_0(\theta - \theta_0)$ in terms of the Lipshitz constant of the gradient of $H(x;\theta)$,

$$\|\nabla_\theta H(x;\theta)_{\theta_1} - \nabla_\theta H(x;\theta)_{\theta_2}\|_2 \leq L_{\nabla_\theta H}\|\theta_1 - \theta_2\|_2.$$

The difference from linear is given by

$$|H(x;\theta) - (\hat{y} + J_0(\theta - \theta_0))| \leq \mathcal{O}(L_{\nabla_\theta H})\|\theta - \theta_0\|_2^2$$

which from the prior slide is of order $\mathcal{O}(L_{\nabla_\theta H} \cdot m\sigma_{min}^{-1}(J_0 J_0^T))$. It then remains to understand how $L_{\nabla_\theta H}$ depends on the dimensions.

Linearization of the neural network

Network approaches linear in large width limit

Consider a two layer network where the first layer maps $x \in \mathbb{R}^d$ to $\mathbb{R}^p$ and the second layer maps $x \in \mathbb{R}^p$ to a scalar;

$$H(x;\theta) = p^{-1/2}\sum_{i=1}^{p} w_i^{(2)}\phi((w_i^{(1)})^T x)$$

where $w_i^{(1)} \in \mathbb{R}^d$ with entries drawn $\mathcal{N}(0, \sigma_w^2/d)$ and $w_i^{(2)}$ are drawn i.i.d. from ± 1, then the $p^{-1/2}$ scaling is needed so that $|H(x\theta)|$ is independent of p as p grows. The Lipshitz constant for the gradient of this network follows from the entries of the gradient being $\nabla_\theta H(x;\theta)_{ij} = p^{-1/2}\phi'((w_i^{(1)})^T x)x_j$ for $i \in [p]$ and $j \in [d]$. (Note we are neglecting the gradient with respect to $w_i^{(2)}$.)

Linearization of the neural network

Network approaches linear in large width limit (continued)

With $\nabla_\theta H(x;\theta)_{ij} = p^{-1/2}\phi'((w_i^{(1)})^T x)x_j$; then if $\phi'(z)$ has Lipshitz constant L_ϕ, and letting wj_i^1 denote the i^{th} entry of the weight matrix for θ_j at layer 1

$$
\begin{aligned}
\|\nabla_\theta H(x;\theta)_{\theta_1} - \nabla_\theta H(x;\theta)_{\theta_2}\|_2^2 &= p^{-1}\sum_{i=1}^{p}\sum_{j=1}^{d} x_j^2(\phi'((w1_i^{(1)})^T x) - \phi'((w2_i^{(1)})^T x))^2 \\
&\leq p^{-1}\|x\|_2^2 \sum_{i=1}^{p} L_\phi \|w1_i^{(1)} - w2_i^{(1)}\|_2^2 \\
&= p^{-1} L_\phi \|x\|_2^2 \|\theta_1 - \theta_2\|_2^2
\end{aligned}
$$

This tell us that $L_{\nabla_\theta H} = \mathcal{O}(p^{-1/2})$ and consequently

$$|H(x;\theta) - (\hat{y} + J_0(\theta - \theta_0)| \leq \mathcal{O}(p^{-1/2} m \sigma_{min}^{-1}(J_0 J_0^T) L_\phi)$$

which goes to zero with width, say if $p = m^2 \log(m)$.

The loss function for the linearized approximation to the network $H(x;\theta)$ about θ_0 is then given by

$$\tilde{\mathcal{L}}(\theta) = \|\hat{y} - y + J_0(\theta - \theta_0)\|_2^2.$$

Rather than solving for the solution θ for which the linear approximation has $\tilde{\mathcal{L}} = 0$, we could explore what occurs when updating θ through gradient descent. This approach gives us some insight into the nature of using gradient descent on the original loss function $\mathcal{L}(\theta)$ for the nonlinear $H(x,\theta)$. Letting $\Delta\theta^{(k)} = \theta^{(k)} - \theta_0$ be the k^{th} iteration of θ centred at θ_0 we have gradient descent for $\Delta\theta$ given by

$$\Delta\theta^{(k+1)} = \Delta\theta^{(k)} - \alpha\nabla_{\Delta\theta}\tilde{\mathcal{L}}(\Delta\theta)_{\Delta\theta^{(k)}}$$

Applying J_0 from the left to both sides of the gradient descent equation and letting $\tilde{y}^{(k)} = J\Delta\theta^{(k)}$ we have $\tilde{y}^{(k+1)} = \tilde{y}^{(k)} - \alpha J_0 J_0^T(\tilde{y}^{(k)} - (\hat{y} - y)$. Subtracting $(\hat{y} - y)$ from both sides and taking the norm we have

$$\begin{aligned}\tilde{\mathcal{L}}(\theta^{(k+1)}) &= \|\hat{y} - y + J_0(\Delta\theta^{(k+1)})\|_2^2 = \|(I - \alpha J_0 J_0^T)(\hat{y} - y + J_0(\Delta\theta^{(k)})\|_2^2 \\ &\leq \|I - \alpha J_0 J_0^T\|_2 \tilde{\mathcal{L}}(\theta^{(k)}).\end{aligned}$$

Let the largest and smallest singular values of $J_0 J_0^T$ be σ_{max} and σ_{min} respectively and $\kappa = \sigma_{max}/\sigma_{min}$, then, if the stepsize $\alpha < 2/(\sigma_{max} + \sigma_{min})$ we have $\tilde{\mathcal{L}}(\theta^{(k+1)}) \leq \frac{\kappa-1}{\kappa+1}\tilde{\mathcal{L}}(\theta^{(k)})$; so $\tilde{\mathcal{L}}(\theta^{(k)}) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \tilde{\mathcal{L}}(\theta^{(0)})$. Limiting small stepsize this gives the same solution as Gauss-Newton which had solution with minimal $\|\theta\|_2$. Note that the above also works if we update the loss at each iteration and use J_k where the gradient is evaluated at the $\theta^{(k)}$.

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:

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

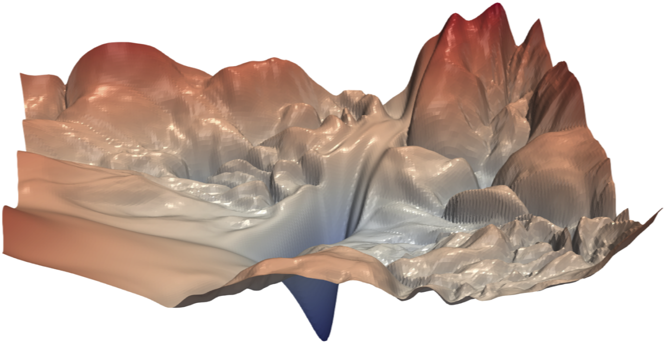
The output of the net is $H(x_\mu;\theta) = \hat{y}_\mu$ and we measure the value of the net through the average sum of squares:

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

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



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

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`

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 $h^{(\ell)} = \phi(W^{(\ell)}h^{(\ell-1)})$ with $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ and $\phi(\cdot) = \max(0,\cdot)$, 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 $H(x,\theta) = W^{(2)}\phi(W^{(1)}x)$ 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 $h^{(\ell)} = \phi(W^{(\ell)}h^{(\ell-1)}))$ with $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ and quadratic activation $\phi(z) = z^2$, 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.

UNIVERSITY OF OXFORD
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.

UNIVERSITY OF OXFORD
Mathematical Institute

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 and Wishart distributions

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

University of Oxford
Mathematical Institute

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

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

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 $\lambda_\pm := \sigma(1\pm\sqrt{\tau})^2$.

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

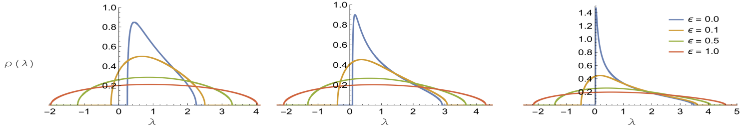


Figure 1. Spectral distributions of the Wishart + Wigner approximation of the Hessian for three different ratios of parameters to data points, ϕ. As the energy ϵ of the critical point increases, the spectrum becomes more semicircular and negative eigenvalues emerge.

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

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

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

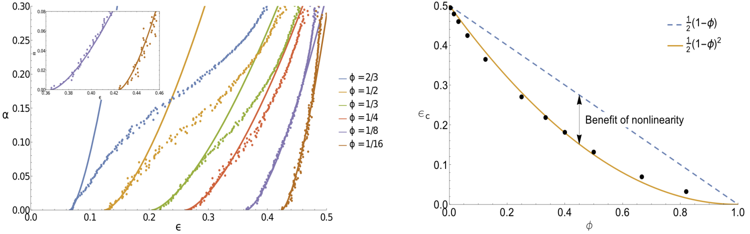
$$\epsilon_c = \frac{\phi^2(27-18\xi-\xi^2+8\xi^{3/2})}{32\tau(1-\tau)^3}, \quad \text{with} \quad \xi = 1+16\tau-8\tau^2.$$

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

Empirical values of ϵ_c and α (Pennington et al. 17')

Match of empirical and analytical calculations

UNIVERSITY OF OXFORD
Mathematical Institute



(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

Manifold of global minimizers (Yaim Cooper 21')

Dimension of global minimizers in overparameterized setting

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 m distinct data elements be defined as

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

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 $(m - rn)-$dimensional submanifold (possibly empty) of $\mathbb{R}^d$.

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

- 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.