Mathematical
Institute

Observations of the loss landscape: impact of parameterization and architecture

Theories of Deep Learning: C6.5, Video 10

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

UNIVERSITY OF OXFORD
Mathematical Institute

Consider a fully connected L layer deep net given by

$$h^{(\ell)} = W^{(\ell)} z^{(\ell)} + b^{(\ell)}, \qquad z^{(\ell+1)} = \phi(h^{\ell}), \qquad \ell = 0, \ldots, L-1,$$

for $\ell = 1, \ldots, L$ with nonlinear activation $\phi(\cdot)$ and $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$. The trainable parameters for the DNN, $\theta := \{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L}$ are learned by minimizing a high dimensional, $|\theta| \sim n^2 L$, loss function such as

$$\mathcal{L}(\theta; X, Y) = (2m)^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n_L} (H(x_\mu(i); \theta) - y_{i,\mu})^2.$$

The shape of $\mathcal{L}(\theta)$ and our knowledge about a good initial minimizer $\theta^{(0)}$ strongly influence our ability to learn the parameters θ for the DNN.

Landscape loss function: VGG9 (Li et al. 18')

One dimensional views of a loss landscape

UNIVERSITY OF OXFORD
Mathematical Institute

DNN loss $\mathcal{L}(\theta)$ between two minimizers, $\theta^s(1-\alpha)+\alpha\theta^l$ trained with small and large batches; horizontal axis α in (a) and (d).

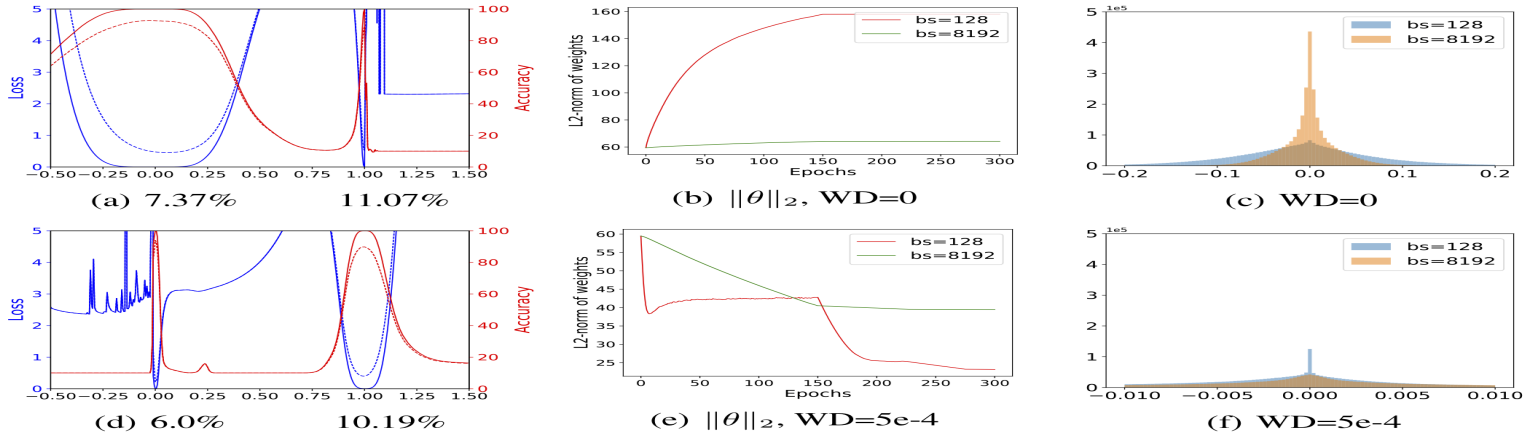


Figure 2: (a) and (d) are the 1D linear interpolation of VGG-9 solutions obtained by small-batch and large-batch training methods. The blue lines are loss values and the red lines are accuracies. The solid lines are training curves and the dashed lines are for testing. Small batch is at abscissa 0, and large batch is at abscissa 1. The corresponding test errors are shown below. (b) and (e) shows the change of weights norm $||\theta||_2$ during training. When weight decay is disabled, the weight norm grows steadily during training without constraints (c) and (f) are the weight histograms, which verify that small-batch methods produce more large weights with zero weight decay and more small weights with non-zero weight decay.

VGG9 is a CNN (Simonyan et al. 15')

`https://arxiv.org/pdf/1409.1556.pdf`

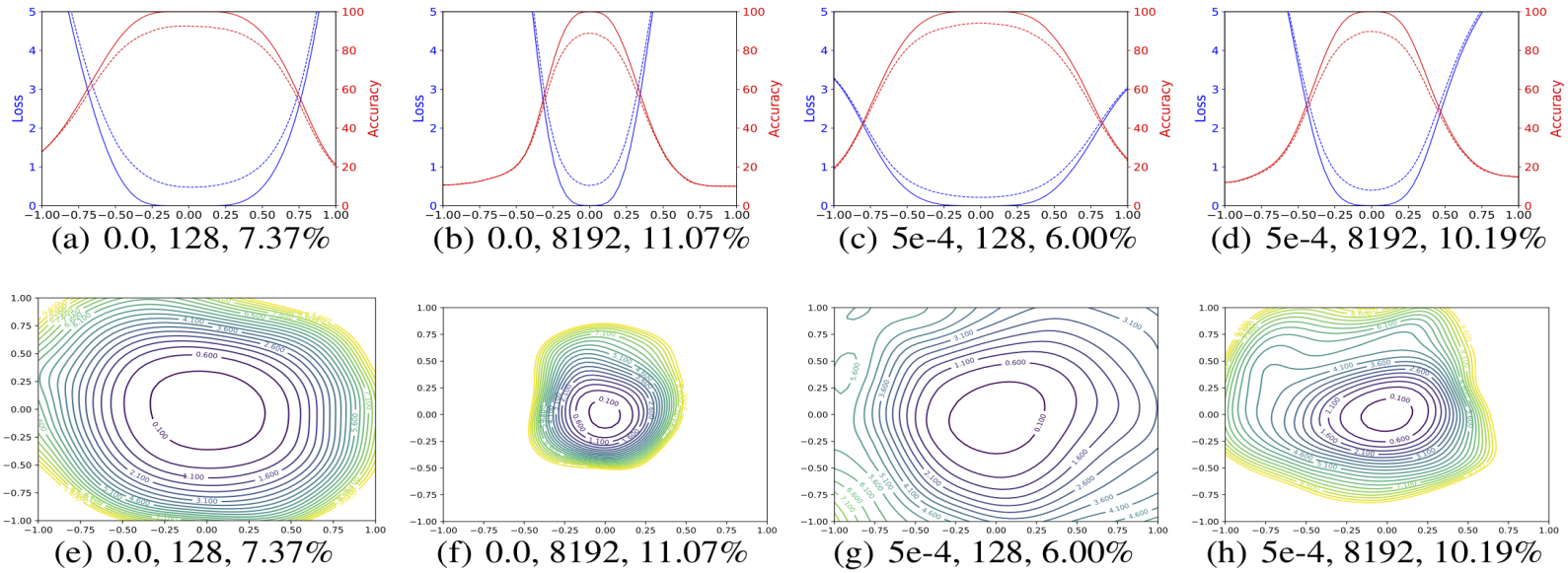
`http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf`

Landscape loss function: VGG9 (Li et al. 18')

One and two dimensional landscape near SGD minima

Mathematical Institute

Impact of training rate weight decay and batch size on level curves of $\mathcal{L}(\theta^* + \alpha\delta + \beta\eta)$. Larger batch size narrows the loss function. Weight decay broadens the loss function.



(a) 0.0, 128, 7.37% (b) 0.0, 8192, 11.07% (c) 5e-4, 128, 6.00% (d) 5e-4, 8192, 10.19%

(e) 0.0, 128, 7.37% (f) 0.0, 8192, 11.07% (g) 5e-4, 128, 6.00% (h) 5e-4, 8192, 10.19%

Figure 3: The 1D and 2D visualization of solutions obtained using SGD with different weight decay and batch size. The title of each subfigure contains the weight decay, batch size, and test error.

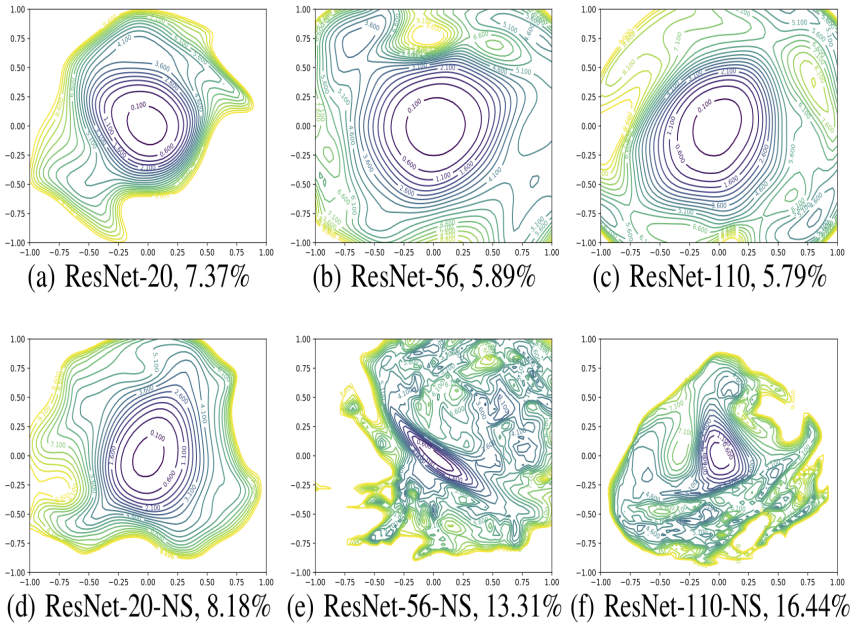
`http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf`

Loss landscape example: ResNet skip (Li et al. 18')

Architecture influences landscape: depth and skip connections

University of Oxford
Mathematical Institute

No-short is a standard fully connected DNN, ResNet (He et al. 15') has additional connections between every second layer.



(a) ResNet-20, 7.37% (b) ResNet-56, 5.89% (c) ResNet-110, 5.79%

(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%

Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.

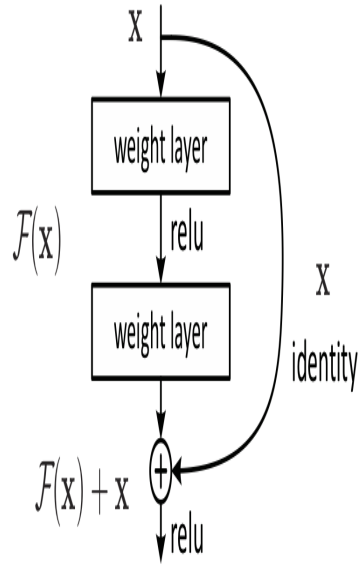


Figure 2. Residual learning: a building block.

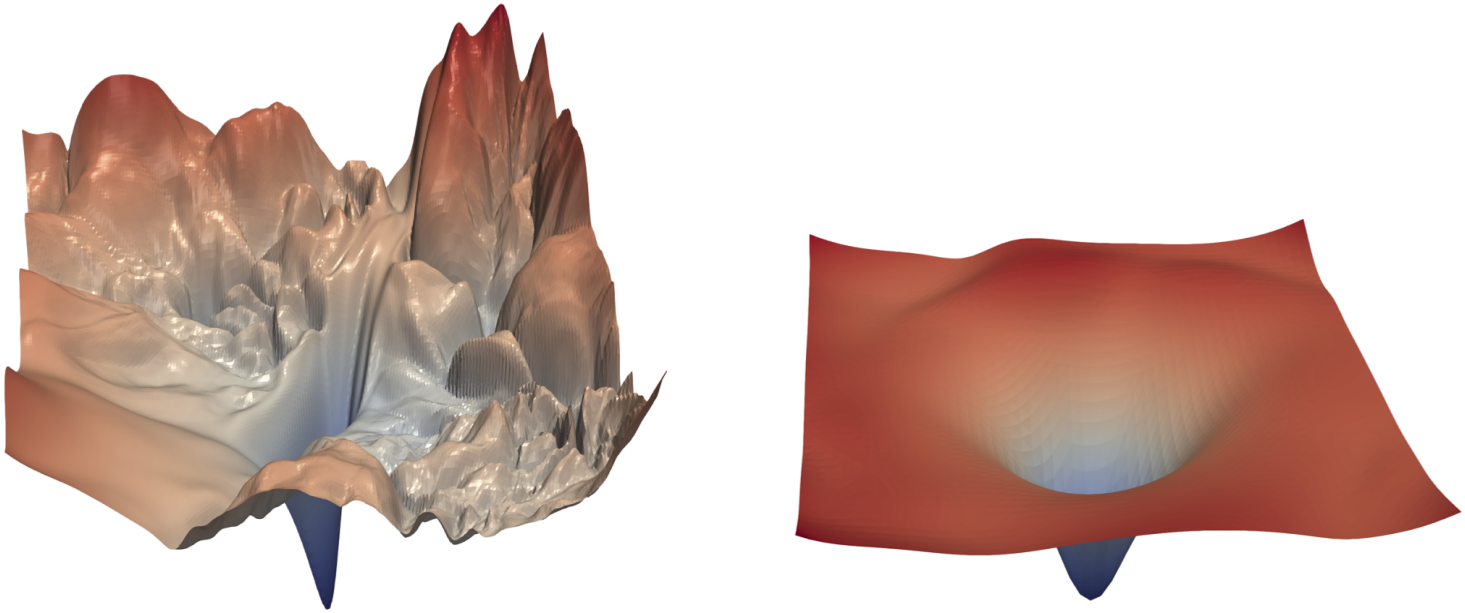
https://arxiv.org/pdf/1512.03385.pdf

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

Loss landscape example: ResNet-56 (Li et al. 18')

Loss landscapes are generally highly non-convex

University of Oxford
Mathematical Institute



(a) without skip connections

(b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

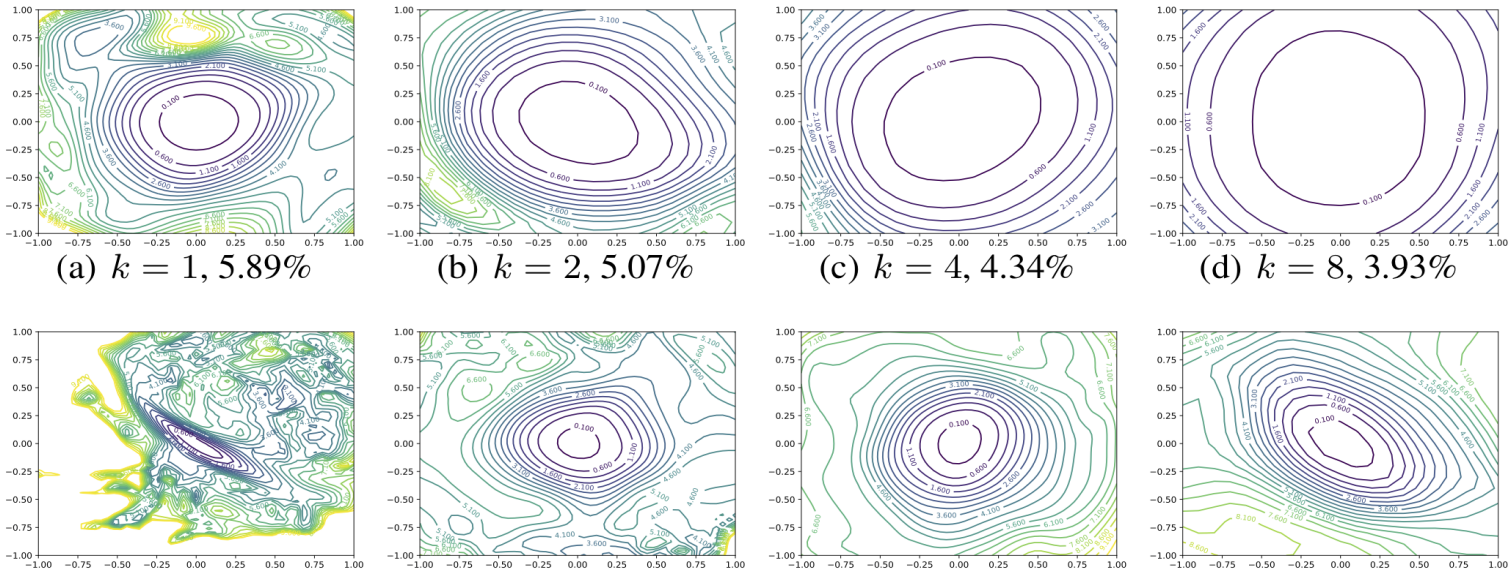
`http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf`

Loss landscape example: ResNet skip (Li et al. 18')

Architecture influences landscape: width

Mathematical Institute

Increasing width over parameterises the net and broadens minima.



(a) $k = 1$, 5.89% (b) $k = 2$, 5.07% (c) $k = 4$, 4.34% (d) $k = 8$, 3.93%

(e) $k = 1$, 13.31% (f) $k = 2$, 10.26% (g) $k = 4$, 9.69% (h) $k = 8$, 8.70%

Figure 6: Wide-ResNet-56 on CIFAR-10 both with shortcut connections (top) and without (bottom). The label $k = 2$ means twice as many filters per layer. Test error is reported below each figure.

`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`

Summar: influences on the loss landscape

Impact of: architecture, depth, width, batch size, weight decay

- Smaller mini-batches and weight decay seem to find minimisers θ^* with in wider basins, resulting in less sensitivity to $\mathcal{L}(\theta^* + \epsilon)$ for small ϵ.
- Inclusion of skip connections, ie. ResNet architectures, and/or increased over parameterisation through greater width appear to give wider less complex landscapes.
- There are theoretical results indicating that width and over parameterisation result in landscapes for which minimisers are simply connected.