Essentially No Barriers in Neural Network Energy Landscape

Felix Draxler [1 2] **Kambis Veschgini** [2] **Manfred Salmhofer** [2] **Fred A. Hamprecht** [1]

Abstract

Training neural networks involves finding minima of a high-dimensional non-convex loss function. Knowledge of the structure of this energy landscape is sparse. Relaxing from linear interpolations, we construct continuous paths between minima of recent neural network architectures on CIFAR10 and CIFAR100. Surprisingly, the paths are essentially flat in both the training and test landscapes. This implies that neural networks have enough capacity for structural changes, or that these changes are small between minima. Also, each minimum has at least one vanishing Hessian eigenvalue in addition to those resulting from trivial invariance.

1. Introduction

Neural networks have achieved remarkable success in practical applications such as object recognition (He et al., 2016; Huang et al., 2017), machine translation (Bahdanau et al., 2015; Vinyals & Le, 2015), speech recognition (Hinton et al., 2012; Graves et al., 2013; Xiong et al., 2017) etc. Theoretical insights on why neural networks can be trained successfully despite their high-dimensional and non-convex loss functions are few or based on strong assumptions such as the eigenvalues of the Hessian at critical points being random (Dauphin et al., 2014), linear activations (Choromanska et al., 2014; Kawaguchi, 2016) or wide hidden layers (Soudry & Carmon, 2016; Nguyen & Hein, 2017).

In the current literature, minima of the loss function are typically depicted as points at the bottom of a valley with a certain width that reflects the generalisation of the network with parameters given by the location of the minimum (Keskar et al., 2016). This is also the picture obtained when the loss function of neural networks is visualised in low dimension (Li et al., 2017).

In this work, we argue that neural network loss minima are not isolated points in parameter space, but essentially form a connected manifold. More precisely, the part of the parameter space where the loss remains below a certain low threshold forms one single connected component.

We support the above claim by studying the energy landscape of several ResNets and DenseNets on CIFAR10 and CIFAR100: For pairs of minima, we construct continuous paths through parameter space for which the loss remains very close to the value found directly at the minima. An example for such a path is shown in Figure 1.

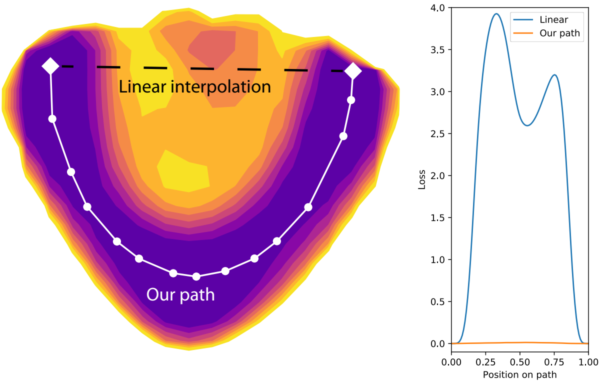

Figure 1. Left: A slice through the one million-dimensional training loss function of DenseNet-40-12 on CIFAR10 and the minimum energy path found by our method. The plane is spanned by the two minima and the mean of the nodes of the path. *Right:* Loss along the linear line segment between minima, and along our high-dimensional path. Surprisingly, the energy along this path is essentially flat.

Our main contribution is the finding of paths

1. that connect minima trained from different initialisations which are not related to each other via known loss-conserving operations such as rescaling,
2. along which the training loss remains essentially at the same value as at the minima,
3. along which the test loss remains essentially constant while the test error rate slightly increases.

[1] Heidelberg Collaboratory for Image Processing (HCI), IWR, University of Heidelberg, D-69120 Heidelberg, Germany [2] Institut für Theoretische Physik, University of Heidelberg, D-69120 Heidelberg, Germany. Correspondence to: Felix Draxler <felix.draxler@iwr.uni-heidelberg.de>.

Initial version under review at the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Submitted on February 9th, 2018. Copyright 2018 by the author(s).

The abundance of such paths suggests that modern neural networks have enough parameters such that they can achieve good predictions while a big part of the network undergoes structural changes. In closing, we offer qualitative justification of this behaviour that may offer a handle for future theoretical investigation.

2. Related Work

In discussions about why neural networks generalise despite the extremely large number of parameters, one often finds the argument that wide minima generalise better (Keskar et al., 2016). This picture is confirmed when visualising the parameter space on a random plane around a minimum (Li et al., 2017). We draw a completely different image of the loss landscape: Minima are not located in finite-width valleys, but there are paths through the parameter space along which the loss remains very close to the value at the minima.

It has previously been argued that minima of networks with ReLU activations lie in strictly flat valleys (Dinh et al., 2017): One can scale all parameters in one layer by a constant α and by α^{-1} the following layer without changing the output of the network. Here, we provide a different class of such a valley: We construct paths between independent minima that are essentially flat.

(Freeman & Bruna, 2016) showed that local minima are connected without large barriers for a CNN on MNIST and an RNN on PTB next word prediction. On CIFAR10 however, they found significant barriers between minima for the CNN considered. We extend their work in two ways: First, we consider ResNets and DenseNets that outperform plain CNNs by a large margin. Second, we apply a state of the art method for connecting minima from molecular statistical mechanics: The Automated Nudged Elastic Band (AutoNEB) algorithm (Kolsbjerg et al., 2016). It is based on the Nudged Elastic Band (NEB) algorithm (Jónsson et al., 1998). We additionally systematically replace paths with unnaturally high loss barrier. Combining the above we find paths with essentially no energy barrier.

NEB has so far been applied to a multi-layer perceptron with a single hidden layer (Ballard et al., 2016). High energy barriers between the minima of network were found when using three hidden neurons, and disappeared upon adding more neurons to the hidden layer. In a follow-up, (Ballard et al., 2017) trained a multi-layer perceptron with a single hidden layer on MNIST. They found that with l^2-regularisation, the landscape had no significant energy barriers. However, for their network they report an error rate of 14.8% which is higher than the 12% achieved even by a linear classifier (LeCun et al., 1998) and the 0.35% achieved with a standard CNN (Ciresan et al., 2011).

In this work, we apply AutoNEB to a nontrivial network for the first time, and make the surprising observation that different minima of state of the art networks on CIFAR10 and CIFAR100 are connected through essentially flat paths.

After submission of this work to the International Machine Learning Conference (ICML) 2018, (Garipov et al., 2018) independently reported that they also constructed paths between neural network minima. They study the loss landscape of several architectures on CIFAR10 and CIFAR100 and report the same surprising observation: minima are connected by paths with constantly low loss.

3. Method

In the following, we use the terms *energy* and *loss* interchangeably.

3.1. Minimum Energy Path

A neural network loss function depends on the architecture, the training set and the network parameters θ. Keeping the former two fixed, we simply write $L(\theta)$ and start with two parameter sets θ_1 and θ_2. In our case, they are minima of the loss function, i.e. they result from training the networks to convergence. The goal is to find the continuous path p^* from θ_1 to θ_2 through parameter space with the lowest maximum loss:

$$p(\theta_1, \theta_2)^* = \underset{p \text{ from } \theta_1 \text{ to } \theta_2}{\arg\min} \left\{ \max_{\theta \in p} L(\theta) \right\}.$$

For this optimisation to be tractable, the loss function must be sufficiently smooth, i.e. contain no jumps along the path. The output and loss of neural networks are continuous functions of the parameters (Montúfar et al., 2014); only the derivative is discontinuous for the case of ReLU activations. However, we cannot give any bounds on how steep the loss function may be. We address this problem by sampling all paths very densely.

Such a lowest path p^* is called the *minimum energy path* (MEP) (Jónsson et al., 1998). We refer to the parameter set with the maximum loss on a path as the “saddle point” of the path because it is a true saddle point of the loss function.

In low-dimensional spaces, it is easy to construct the exact minimum energy path between two minima, for example by using dynamic programming on a densely sampled grid.

This is not possible for present day’s neural networks with parameter spaces that have millions of dimensions. We thus must resort to methods that construct an approximation of the MEP between two points using some local heuristics. In particular, we resort to the Automated Nudged Elastic Band (AutoNEB) algorithm (Kolsbjerg et al., 2016). This method is based on the Nudged Elastic Band (NEB) algorithm (Jónsson et al., 1998).

NEB bends a straight line segment by applying gradient forces until there are no more gradients perpendicular to the path. Then, as for the MEP, the highest point of the resulting path is a critical point. While this critical point is not necessarily the saddle point we were looking for, it gives an upper bound for the energy at the saddle point.

In the following, we present the mechanical model behind and the details of NEB. We then proceed to AutoNEB.

Mechanical Model A chain of $N + 2$ pivots (parameter sets) p_i for $i = 0, \dots, N + 1$ is connected via springs of stiffness k. The initial and the final pivots are fixed to the minima to connect, i.e. $p_0 = \theta_1$ and $p_{N+1} = \theta_2$. Using gradient descent, the path that minimises the following energy function is found:

$$E(p) = \sum_{i=1}^{N} L(p_i) + \sum_{i=0}^{N} \frac{1}{2} k \left\| p_{i+1} - p_i \right\|^2 \tag{1}$$

The problem with this energy formulation lies in the choice of the spring constant: If, on the one hand, k is too small, the distances between the pivots become larger in areas with high energy. However, identifying the highest point on the path and its energy is the very goal of the algorithm, so the sampling rate should be high in the high-energy regions. If, on the other hand, k is chosen too large, it becomes energetically advantageous to shorten and hence straighten the path as the spring energy grows quadratically with the total length of the path. This cuts into corners of the loss surface and the resulting path can miss the saddle point.

Nudged Elastic Band Inspired by the above model, (Jónsson et al., 1998) presented the *Nudged Elastic Band* (NEB). For brevity, we directly present the improved version by (Henkelman & Jónsson, 2000). The force resulting from Equation (1) consists of a force derived from the loss and a force originating from the springs:

$$F_i = -\nabla_{p_i} E(p) = F_i^L + F_i^S$$

For NEB, the physical forces are modified, or *nudged*, so that the loss force only acts perpendicularly to the path and the spring force only parallelly to the path (see also Figure 2):

$$F_i^{\text{NEB}} = F_i^L\big|_{\perp} + F_i^S\big|_{\|}.$$

The direction of the path is defined by the local tangent $\hat{\tau}_i$ to the path. The two forces now read:

$$\begin{aligned} F_i^L\big|_{\perp} &= -(\nabla L(p_i) - (\nabla L(p_i) \cdot \hat{\tau}_i)\hat{\tau}_i) \\ F_i^S\big|_{\|} &= (F_i^S \cdot \hat{\tau}_i)\hat{\tau}_i \end{aligned} \tag{2}$$

where the spring force opposes unequal distances along the path:

$$F_i^S = -k(\|p_i - p_{i-1}\| - \|p_{i+1} - p_i\|) \tag{3}$$

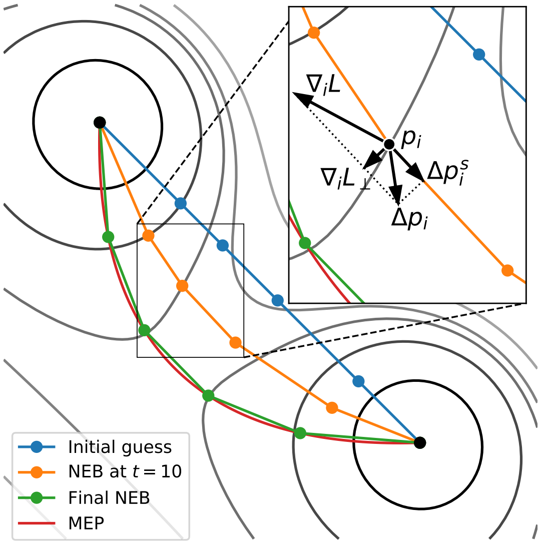

Figure 2. Two dimensional loss surface, with two minima connected by a minimum energy path (MEP) and a nudged elastic band (NEB) at iteration 0, 10 and converged. Construction of NEB update Δp_i for one pivot. The tangent points to the neighbouring pivot with higher energy. Re-distribution Δp_i^s acts parallelly and the loss force $\nabla_i L$ perpendicularly to the tangent.

In this formulation, high energy pivots no longer "slide down" from the saddle point. The spring force only re-distributes pivots on the path, but does not straighten it. Pivots can be spaced unequally by introducing target distances or unequal spring constants into Equation (3).

The local tangent is chosen to point in the direction of one of the adjacent pivots ($\mathcal{N}$ normalises to length one):

$$\hat{\tau}_i = \mathcal{N} \begin{cases} p_{i+1} - p_i & \text{if } L(p_{i+1}) > L(p_{i-1}) \\ p_i - p_{i-1} & \text{else.} \end{cases}$$

This particular choice of $\hat{\tau}$ prevents kinks in the path and ensures a good approximation near the saddle point (Henkelman & Jónsson, 2000).

The above procedure requires the following hyperparameters: The spring stiffness k and number of pivots N.

(Sheppard et al., 2008) claim that a wide range of k leads to the same result on a given loss surface. However, if chosen too large, the optimisation can become unstable. If it is too small, an excessive number of iterations are needed before the pivots become equally distributed. We did not find a value for k that worked well across different loss surfaces and number of pivots N. Instead, we re-distribute the pivots in each iteration t and set the actual spring force to zero. The loss force is still restricted to act parallelly to the path. In the literature, this is sometimes referred to as the *string method* (Sheppard et al., 2008).

Algorithm 1 shows how the initial path is iteratively updated using the above forces. As a companion, Figure 2 visualises the forces in one update step for a two dimensional example. In this formulation, we use gradient descent to update the path. Any other gradient based optimiser can be used. It typically introduces additional hyperparameters, for example a learning rate γ. The number of iterations T should be chosen large enough for the optimisation to converge.

Algorithm 1 NEB

```
**Input:** initial path $p^{(0)}$ with $N+2$ pivots,
  $p_0^{(0)} = \theta_1$ and $p_{N+1}^{(0)} = \theta_2$.
**for** t = $1, \dots, T$ **do**
    Redistribute pivots on path $p^{(t-1)}$ and store as $p$.
    **for** i = $1, \dots, N$ **do**
        Compute projected loss force $F_i = F_i^L \big|_\perp$.
        Store pivot $p_i^{(t)} = p_i + \gamma F_i$.
    **end for**
**end for**
**return** final path $p^{(T)}$
```

The evaluation time of Algorithm 1 rises linearly with the number of iterations and the number of pivots on the path. Computing the NEB forces can trivially be parallelised over the pivots.

The number of pivots N trades off between computational effort on the one hand and subsampling artefacts on the other hand. In neural networks, it is not known what sampling density is needed for traversing the parameter space. We use an adaptive procedure that inserts more pivots where needed:

AutoNEB The Automated Nudged Elastic Band (AutoNEB, Algorithm 2) wraps the above NEB algorithm (Kolsbjerg et al., 2016). Initially, it runs NEB with low N for a small number of iterations T. It is then checked if the current pivots are sufficient to accurately sample the path. If this is not the case, new pivots are added at locations where it is estimated that the path requires more accuracy.

As a criterion, new pivots are inserted where the true loss values deviate from the linear interpolation between each neighbouring pivot pair larger than a certain threshold, as visualised in Figure 3. This requires handling unequal spaces between pivots.

3.2. Local minimum energy paths

AutoNEB is not guaranteed to find the true MEP. Instead, it can get stuck in local minimum energy paths (local MEPs). This means that the saddle point energies reported by AutoNEB can only be an upper bound for the unknown minimal saddle point losses.

Algorithm 2 AutoNEB

```
**Input:** Minima to connect $\theta_1, \theta_2$.
Initialise $N$ pivots equally spaced on line segment $(\theta_1, \theta_2)$.
**for** $t' = 1, \dots, T'$ **do**
    Optimise path using NEB (Algorithm 1).
    Evaluate loss along NEB.
    Insert pivots where residuum is large.
**end for**
**return** path after final iteration.
```

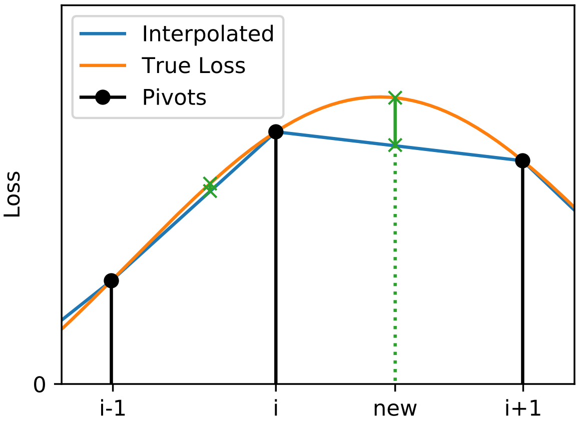

Figure 3. New items are inserted in each cycle of AutoNEB when the true energy at an interpolated position between two points rises too high compared to the interpolated energy. Between i and $i+1$, a new pivot is inserted. Between $i-1$ and i, the difference is small enough that no additional pivot is needed.

The good news is that the graph of minima and local MEPs has an ultrametric property: Suppose some local MEPs from a minimum A to B and from B to C are known. We call them p_{AB} and p_{BC}. The respective saddle point energies give an upper bound for the true saddle point energies (marked with an asterisk):

$$L^*_{AB} \le L_{AB} = \max_{\theta \in p_{AB}} L(\theta)$$
$$L^*_{BC} \le L_{BC} = \max_{\theta \in p_{BC}} L(\theta)$$

Additionally, the concatenation of the two paths yields an upper bound for the true saddle point energy between A and C (ultrametric triangle inequality):

$$L^*_{AC} \le \max\{L_{AB}, L_{BC}\}$$

This is easy to see: When concatenating the paths p_{AB} and p_{BC}, this gives a new path p_{AC} connecting A to C. The saddle point is located at the maximum loss along a

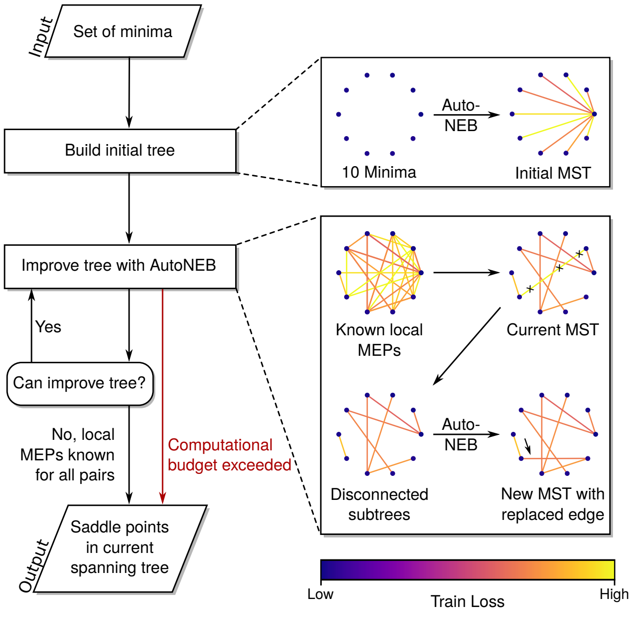

Figure 4. Overview over Algorithm 3 and examples for the first nine iterations and some later iteration. First, all minima are connected to one particular minimum. Then, AutoNEB computes new local MEPs to circumvent the worst local MEPs in the minimum spanning tree. This is repeated until local MEPs are known between all pairs of minima or the procedure is stopped early. Whenever the algorithm stops, an upper bound for each pair of minima is available via the minimum spanning tree.

path and hence the saddle point energy of p_{AC} is $L_{AC} = \max\{L_{AB}, L_{BC}\}$.

This has three consequences:

1. As soon as the minima and computed local MEPs form one connected graph, upper bounds for all saddle energies are available. We can hence very quickly get upper bounds for all pairs of minima by connecting one minimum to all others.

2. When AutoNEB finds a bad local MEP, this can be addressed by computing paths between other pairs of minima. As soon as a lower path is found by concatenating other paths, the bad local MEP can be removed. This means that the bad local paths can easily be corrected for.

3. When we evaluate the saddle point energies of a set of computed local MEPs, we can ignore paths with higher energy than the concatenation of paths with a lower maximal energy.
   These lowest local MEPs form a minimum spanning tree in the available graph (Gower & Ross, 1969). A Minimum Spanning Tree (MST) can be found efficiently, e.g. using Kruskal's algorithm.

Algorithm 3 Energy Landscape Exploration

```
Input: set of minima $\theta_i$.
Connect $\theta_1$ to all $\theta_i, i \neq 1$, yielding a spanning tree.
repeat
  Remove edge $p_o$ with highest loss from spanning tree.
  From each resulting tree, try to select one minimum,
    so that no local MEP is known for the pair.
  if search failed then
    Re-insert $p_o$ and ignore it when searching for the
      highest edge in the future.
  else
    Compute new path $p_n$ using AutoNEB.
    if $L_{p_n} < L_{p_o}$ then
      Add $p_n$ to the tree, making tree “lighter”.
    else
      Re-insert $p_o$ to the tree (no better path was found).
    end if
  end if
until one local MEP is known for each pair of minima
  or computational budget is exceeded.
return saddle points in minimum spanning tree.
```

Guided by the above, we resort to the heuristic (surely not new, though we have no reference) spelled out in Algorithm 3 and visualised in Figure 4 to determine which pair of minima to connect next.

The procedure suggests new tuples of minima until local MEPs are known between all pairs of minima. Since running AutoNEB is computationally expensive (effort on the order of training the corresponding network), we stop the iteration when the minimum spanning tree contains only similar saddle point energies.

4. Experiments

We connect minima of different ResNets (He et al., 2016) and DenseNets (Huang et al., 2017) on the image classification tasks CIFAR10 and CIFAR100. We train several instances of the network from distinct random initialisations following the instructions in the original literature. Then we connect pairs of minima using AutoNEB.

We report the average cross-entropy loss and misclassification rates over the full training and test data for the minima found. For the final evaluation, we reduce the saddle points to the minimum spanning tree with the corresponding training loss as weight.

4.1. ResNet

We train ResNets on both CIFAR10 and CIFAR100 (ResNet-20, -32, -44 and -56) following the training procedure in (He et al., 2016). For each tuple of architecture and dataset, we

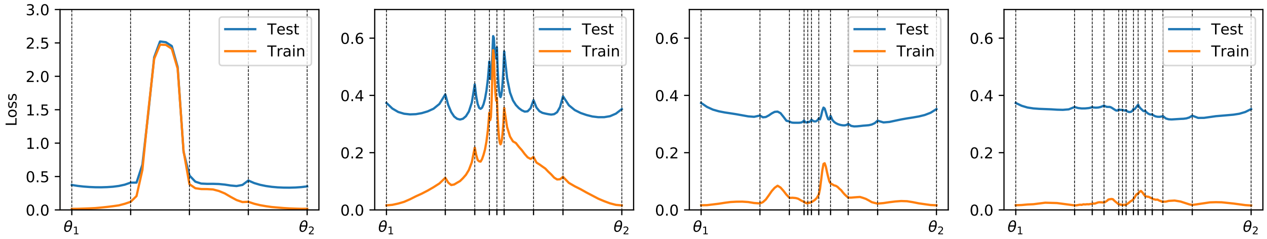

Figure 5. Typical snapshots of the loss along path at the end of AutoNEB cycles, here for a ResNet-20 on CIFAR10: (1) After the first cycle, typically one or two corners are cut. New pivots are inserted at high loss values, here between the second and the third pivot. (2) After four cycles of high learning rate, the highest loss on the path is reduced by a factor of five. Between pivots we find low energy regions that we attribute to the high learning rate 0.01. (3) The first round with low learning rate 0.001 reduces the energy by another factor of two. (4) After 14 cycles, no major energy bumps exist between the pivots, the procedure is converged.

train ten networks. The pairs to connect are ordered by Algorithm 3.

AutoNEB (Algorithm 2) is run for a total of 14 cycles of NEB per minimum pair. The loss is evaluated for each pivot on a random batch of 512 training samples for ResNet-20 and ResNet-32 and 256 training samples for ResNet-44 and ResNet-56 (four respectively two times the batch size for training).

After each cycle, new pivots are inserted at positions where the loss exceeds the energy estimated by linear interpolation between pivots by at least 20% compared to the total energy difference along the path. Comparing to the total loss difference prioritises big errors which is beneficial as each additional pivot implies one more loss evaluations per iteration. The energy is evaluated on 9 points between each pair of neighbouring pivots.

As in the original paper, SGD with momentum 0.9 and l^2-regularisation with $\lambda = 0.0001$ is used.

We use the following learning schedule, inspired by the original training procedure, see also Figure 5:

1. Four cycles of 1000 steps (about 100 or 50 epochs for batch size of 512 respectively 256) each with learning rate 0.1.
2. Two cycles with 2000 steps and learning rate 0.1. The number of steps was increased as it did not prove necessary inserting new pivots after 1000 steps.
3. This was followed by four cycles of 1000 steps with learning rate 0.01. In this phase, the energy drops significantly.
4. No big improvement was seen in the last four cycles of 1000 steps each with a learning rate of 0.001.

4.2. DenseNet

We train a DenseNet-40-12 and a DenseNet-100-12-BC on both CIFAR10 and CIFAR100 following the training procedure in (Huang et al., 2017). The AutoNEB cycles were configured exactly as for the ResNets except for the batch sizes which was set to 256. As before, we train ten minima and apply our connection procedure in Algorithm 3 until no major improvements are made.

4.3. Results

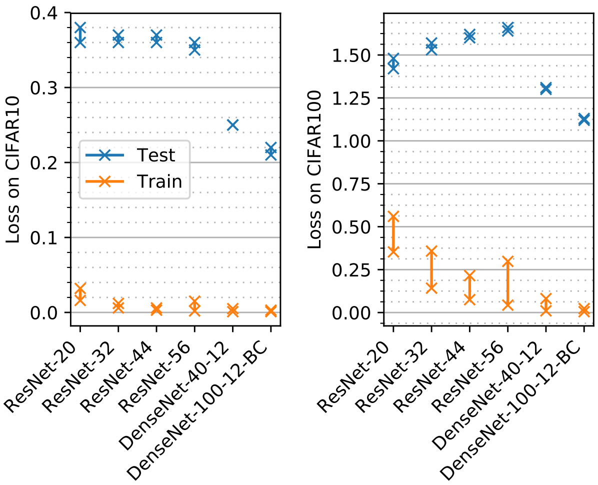

Figure 6. Comparison of minimum and saddle point loss on training and test set. CIFAR10 is on the left and CIFAR100 on the right. Each pair of points represents the average loss at the minima (lower point) and the corresponding mean saddle point loss (higher point), connected by a straight line. The lower row (orange) was evaluated on the training set and the higher losses (blue) on the test set. On the training set, the saddle point loss is small compared to the test set. On the test set, the points of the minima and the saddle points are very close.

The saddle point losses for both training and test sets found by AutoNEB are shown in Figure 6. For reference, the corresponding numbers can be found in Table 1.

The training energies can be compared to a few other characteristic loss values of a neural network, ordered from high to low:

Table 1. Quantitative results. "Min." denotes the average value at the minima. For the saddle point values ("Sadd."), the maximum value of each metric along each final path is computed and the results are averaged. The "epoch" is measured at the point where the loss falls below the saddle point loss for the first time. It is noted in **bold** if it belongs to the third part of training with learning rate $\gamma = 10^{-4}$. ResNets are trained for 136 epochs, DenseNets for 266 epochs.

		Train energy				Test energy		Test error rate [%]		
		Min.	Sadd.	Factor	Epoch	Min.	Sadd.	Min.	Sadd.	Δ
Dataset	Architecture									
C10+	ResNet-20	0.016	0.032	2	**104**	0.36	0.38	8.5	8.9	0.5
	ResNet-32	0.006	0.012	2	**107**	0.36	0.37	7.5	7.9	0.4
	ResNet-44	0.003	0.006	2	**122**	0.36	0.37	7.1	7.5	0.4
	ResNet-56	0.002	0.015	7	83	0.35	0.36	6.9	7.5	0.6
	DenseNet-40-12	0.001	0.005	6	**205**	0.25	0.25	5.6	6.0	0.4
	DenseNet-100-12-BC	0.001	0.003	5	**205**	0.21	0.22	4.9	5.2	0.3
C100+	ResNet-20	0.353	0.560	2	79	1.42	1.48	33.3	34.8	1.5
	ResNet-32	0.142	0.358	3	77	1.53	1.57	31.5	33.7	2.2
	ResNet-44	0.075	0.215	3	85	1.60	1.62	30.8	32.6	1.8
	ResNet-56	0.043	0.298	7	71	1.64	1.66	30.3	32.4	2.0
	DenseNet-40-12	0.010	0.081	8	166	1.30	1.31	26.3	27.7	1.4
	DenseNet-100-12-BC	0.005	0.023	5	**205**	1.12	1.13	23.7	24.6	0.9

1. The average loss for an untrained network. For the cross-entropy loss, it is $-\log(0.1) = 2.3$ on CIFAR10 and $-\log(0.01) = 4.6$ on CIFAR100.
   The saddle point energies on both training sets are *about two orders of magnitude smaller* than the loss at the initialisation of the network.

2. The loss of the test set at the minima.
   All saddle point energies on CIFAR10 are about one order of magnitude smaller than the average minimum energy on the test set. On CIFAR100, the saddle point energies of the ResNets are smaller than a third of the value on the test set. For the DenseNets, they are at least one order of magnitude smaller.

3. The loss of the training set at the minima.
   The loss at the saddle points is 2-8 times as large as the mean loss of the minima. These ratios are noisy because the denominator can approach zero when the network fits the data perfectly (Zhang et al., 2017). Keep this in mind reading the reported factor between saddle point energies and minima.

The test error rate at the saddle point gives an intuition on how much information was lost on the saddle point. On the ResNets, the error rises by maximally 0.7% on CIFAR10 and 2.9% on CIFAR100. For the DenseNets, the error rises by up to 0.4% on CIFAR10 and 1.5% on CIFAR100. These differences are small compared to the error rate at the minima.

We also measure the epoch at which the loss crosses the saddle point loss during training, listed in Table 1. This procedure is visualised for DenseNet-100-BC and ResNet-56 on CIFAR100 in Figure 7. We find the learning curve to fall below the saddle point energy only after the learning rate was reduced for the first time and the loss dropped at least by a half. For some architectures, it is even after the second decay.

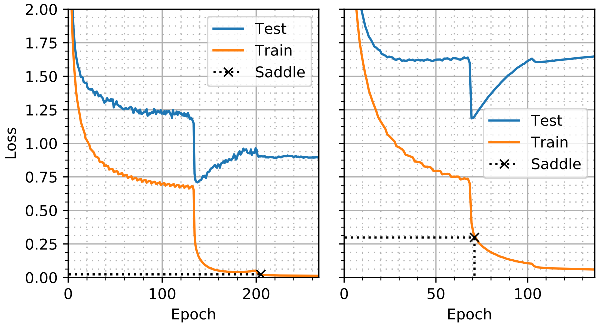

Figure 7. Learning curve of DenseNet-100-12-BC (left) and ResNet-56 (right) on CIFAR100. The training loss passes the mean saddle point energy at about 205 epochs or 78% of training respectively 72 epochs or 52%. Among all architectures considered, the average saddle point crosses the training loss last for DenseNet-100-12-BC and earliest for ResNet-56, both on CIFAR100. The crossing for the DenseNet had to be identified in a log plot.

We conclude that the saddle points have surprisingly low loss with respect to all metrics above.

5. Discussion

We have pointed out an intriguing property of the loss surface of current-day deep networks, by upper-bounding the saddle points between the parameter sets that result from stochastic gradient descent, a.k.a. "minima". These empirical upper bounds are astonishingly close to the loss at the minima themselves. At this point, we cannot give a formal characterization of the regime in which this finding holds.

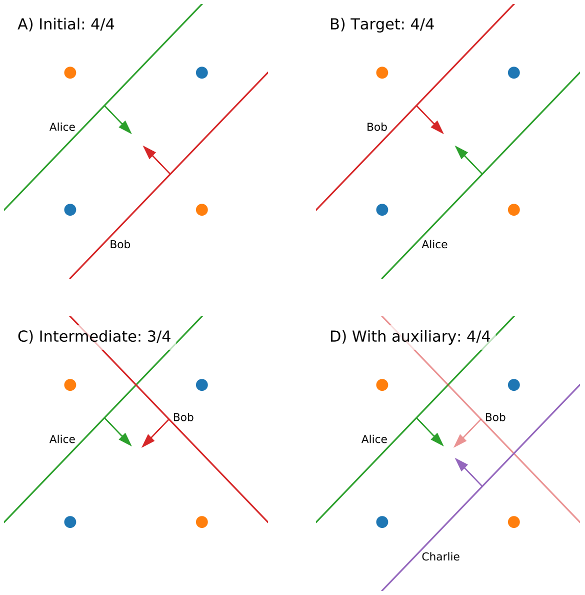

Figure 8. Network capacity for XOR dataset: The continuous transition from *(A)* one minimum to *(B)* another minimum is not possible without *(C)* misclassifying at least one instance. *(D)* Adding one helper neuron makes the transition possible while always predicting the right class for all data points, i.e. by turning off the outgoing weight of Bob.

A formal proof is also complicated by the fact that the loss surface is a function not only of the parameters and the architecture, but also of the training set; and the distribution of real-world structured data such as images or sentences does not lend itself to a compact mathematical representation. That said, we want to make two related arguments that may help explain why we observe no substantial barrier between minima.

5.1. Resilience

State of the art neural networks have dozens or hundreds of neurons / channels per layer, and skip connections between non-adjacent layers. Assume that by training, a parameter set with low loss has been identified. Now if we perturb a single parameter, say by adding a small constant, but leave the others free to adapt to this change to still minimise the loss, it may be argued that by adjusting somewhat, the myriad other parameters can “make up” for the change imposed on only one of them. After this relaxation, the procedure and argument can be repeated, though possibly with the perturbation of a different parameter.

This type of resilience is exploited and encouraged by procedures such as Dropout (Srivastava et al., 2014) or ensembling (Hansen & Salamon, 1990). It is also the reason why neural networks can be greatly condensed before a substantial increase in loss occurs (Liu et al., 2017).

5.2. Redundancy

Consider the textbook example of a two-layer perceptron that can fit the XOR problem. The two neurons traditionally used in the first hidden layer – let’s call them Alice and Bob – are shown in Figure 8(A). We can obtain an equivalent network by exchanging Alice and Bob (and permuting the weights of the neuron in the second hidden layer, not shown). This network, also corresponding to a minimum of the loss surface, is shown in Figure 8(B). Now, any path between these two minima will entail parameter sets such as the one in Figure 8(C) that incur high loss.

If, on the other hand, we introduce an auxiliary neuron, Charlie, we can play a small choreography: Enter Charlie. Charlie stands in for Bob. Bob transitions to Alice’s role. Alice takes over from Charlie. Exit Charlie. If the neuron in the second hidden layer adjusts its weights so as to disregard the output from the neuron-in-transition, the entire network incurs no higher loss than at the two original minima. We have constructed a perfect minimum energy path.

6. Conclusion

We find that the loss surface of deep neural networks contains paths with constantly low loss. We put forth two closely related arguments in the above. Both hold only if the network has some extra capacity, or degrees of freedom, to spare. Empirically, this seems to be the case for modern-day architectures applied to standard problems. We argue that due to the width of each layer, the network heavily replace parameters while producing an output with low loss.

This has the profound implication that low Hessian eigenvalues exist apart from the eigenvectors with analytically zero eigenvalues due to scaling.

The method opens the door to further empirical research on the energy landscape neural networks. When the hyperparameters of AutoNEB are further refined, we expect to find even lower paths up to the level where the true saddle points are recovered. It is then interesting to see if certain minima have a higher barrier between them than others. This makes it possible to recursively form clusters of minima, i.e. using single-linkage clustering. This analysis is yet not possible for the large error bars that we find. In the traditional energy landscape literature, this kind of clustering is done in disconnectivity graphs (Wales et al., 1998).

For practical applications, we imagine using the resulting paths as a large ensemble of neural networks, especially given that we observe practically lower test loss along the path.

References

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Ballard, A. J., Das, R., Martiniani, S., Mehta, D., Sagun, L., Stevenson, J. D., and Wales, D. J. Energy landscapes for machine learning. *Physical Chemistry Chemical Physics (Incorporating Faraday Transactions)*, 19:12585–12603, 2017. doi: 10.1039/C7CP01108C.

Ballard, Andrew J., Stevenson, Jacob D., Das, Ritankar, and Wales, David J. Energy landscapes for a machine learning application to series data. *J. Chem. Phys.*, 144 (12):124119, Mar 2016. ISSN 1089-7690. doi: 10.1063/1.4944672. URL http://dx.doi.org/10.1063/1.4944672.

Choromanska, Anna, Henaff, Mikael, Mathieu, Michaël, Arous, Gérard Ben, and LeCun, Yann. The loss surface of multilayer networks. *CoRR*, abs/1412.0233, 2014. URL http://arxiv.org/abs/1412.0233.

Ciresan, Dan C, Meier, Ueli, Masci, Jonathan, Maria Gambardella, Luca, and Schmidhuber, Jürgen. Flexible, high performance convolutional neural networks for image classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, pp. 1237. Barcelona, Spain, 2011.

Dauphin, Yann, Pascanu, Razvan, Gülçehre, Çaglar, Cho, Kyunghyun, Ganguli, Surya, and Bengio, Yoshua. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *CoRR*, abs/1406.2572, 2014. URL http://arxiv.org/abs/1406.2572.

Dinh, Laurent, Pascanu, Razvan, Bengio, Samy, and Bengio, Yoshua. Sharp minima can generalize for deep nets. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/dinh17b.html.

Freeman, C. D. and Bruna, J. Topology and Geometry of Half-Rectified Network Optimization. *ArXiv e-prints*, November 2016.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *ArXiv e-prints*, February 2018.

Gower, J. C. and Ross, G. J. S. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1): 54–64, 1969. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346439.

Graves, Alex, Mohamed, Abdel-rahman, and Hinton, Geoffrey. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649. IEEE, 2013.

Hansen, Lars Kai and Salamon, Peter. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henkelman, Graeme and Jónsson, Hannes. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics*, 113(22):9978–9985, 2000.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Huang, Gao, Liu, Zhuang, Weinberger, Kilian Q, and van der Maaten, Laurens. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, pp. 3, 2017.

Jónsson, Hannes, Mills, Greg, and Jacobsen, Karsten W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and quantum dynamics in condensed phase simulations*, pp. 385–404. World Scientific, 1998.

Kawaguchi, Kenji. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, and Tang, Ping Tak Peter. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kolsbjerg, Esben L, Groves, Michael N, and Hammer, Bjørk. An automated nudged elastic band method. *The Journal of chemical physics*, 145(9):094107, 2016.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, Hao, Xu, Zheng, Taylor, Gavin, and Goldstein, Tom. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

Liu, Zhuang, Li, Jianguo, Shen, Zhiqiang, Huang, Gao, Yan, Shoumeng, and Zhang, Changshui. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2736–2744, 2017.

Montúfar, Guido, Pascanu, Razvan, Cho, Kyunghyun, and Bengio, Yoshua. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pp. 2924–2932, 2014.

Nguyen, Quynh and Hein, Matthias. The loss surface of deep and wide neural networks. In *ICML*, 2017.

Sheppard, Daniel, Terrell, Rye, and Henkelman, Graeme. Optimization methods for finding minimum energy paths. *The Journal of Chemical Physics*, 128(13):134106, Apr 2008. ISSN 1089-7690. doi: 10.1063/1.2841941. URL http://dx.doi.org/10.1063/1.2841941.

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *ArXiv e-prints*, May 2016.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Vinyals, Oriol and Le, Quoc V. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

Wales, David J, Miller, Mark A, and Walsh, Tiffany R. Archetypal energy landscapes. *Nature*, 394(6695):758, 1998.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423, Dec 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2017.2756440.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.