Appendices

A. AutoNEB insertion

New pivots are inserted at locations where the loss values resulting from evaluations rise higher than a certain threshold above the the estimate given the adjacent pivots, see Figure A.1. More formally, the loss curve between pivots i and $i+1$ can be approximated by interpolating the pivot values:

$$L^{i,i+1}_{\text{guess}}(\alpha) = L(p_i)(1-\alpha) + L(p_{i+1})\alpha$$

where $\alpha \in [0,1]$ interpolates between the pivots. The true loss value a the same position is:

$$L^{i,i+1}(\alpha) = L\big(p_i(1-\alpha) + p_{i+1}\alpha\big).$$

Denote the difference as:

$$\Delta L^{i,i+1}(\alpha) = L^{i,i+1}(\alpha) - L^{i,i+1}_{\text{guess}}(\alpha)$$

If the true loss rises too high above the estimated loss, a new pivot should be inserted. It is beneficial to first insert pivots at the highest residuum: Each new pivot requires more expensive gradient evaluations. The deviation between true loss and estimate is evaluated at discrete positions $\alpha \in \mathcal{A} \subset (0,1)$, $|\mathcal{A}| < \infty$. The differences $\Delta L^{i,i+1}(\alpha)$ are normalised to the range of values of $L(p_i)$. When this normalised deviation rises above a certain threshold ϑ, a new pivot is inserted, i.e. insert a pivot between i and $i+1$ when:

$$\vartheta > \frac{\Delta L^{i,i+1}(\alpha)}{\max_i L(p_i) - \min_i L(p_i)} =: \Delta l^{i,i+1}(\alpha) \qquad (4)$$

Only one pivot is inserted per line segment per AutoNEB iteration. If several α for one line segment fulfil the above condition, only the position with the highest residuum is chosen. Additionally, the total number of pivots to insert per iteration is limited so that the highest deviations are prioritised.

B. Quantitative minimum and saddle point losses

Table B.1 shows the full list of results. Here, we compare the numbers in detail to characteristic metrics of a neural network, with error margins below 10% for all energy and error rate values:

The *loss of an untrained network* amounts to $-\log(0.1) = 2.3$ on CIFAR10 and $-\log(0.01) = 4.6$ on CIFAR100.

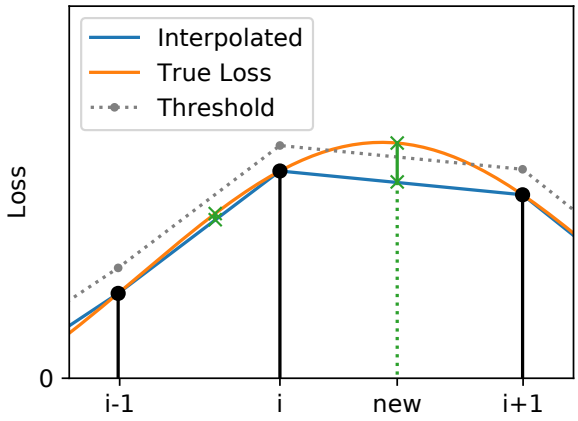

Figure A.1. New items are inserted in each cycle of AutoNEB when the true energy at an interpolated position between two points rises too high compared to the interpolated energy. Between i and $i+1$, a new pivot is inserted. Between $i-1$ and i, the difference is small enough that no additional pivot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

The saddle point energies are between $1/3$ to $1/20$ of the initial loss for the shallow networks (all CNNs and ResNet-8) and *about two orders of magnitude smaller* for the deep residual networks.

The *test loss* at the saddle points of the smallest one-layer CNNs comes close to the test loss of the minima. The wider and especially deeper the CNN, the closer the saddle loss approaches the minima. The saddle point energies of the deep residual networks (not ResNet-8) on CIFAR10 are about one order of magnitude smaller than the average minimum loss 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*.

The *training loss* at the saddle points is at most eight times as large as the training loss of the minima. This ratio, reported as "factor" in Table B.1, is hard to interpret directly as it can approach zero when the network fits the data perfectly. Instead, we report that all saddle losses are closer to the training than the test loss for all but the smallest three basic CNNs. For all deeper architectures, the saddle loss is *much closer to the training than to the test loss*, see Figure 5.

The *classification performance* does not decrease significantly along the paths. 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.

Table B.1. Quantitative results. "Min." denotes the average value at the minima. For the saddle point values ("Sadd."), the maximum value of each metric along the local MEPs 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 $\gamma = 10^{-3}$. Basic CNNs and ResNets are trained for 136 epochs, DenseNets for 266 epochs. The "factor" is the ratio between average saddle point and minima loss. The standard deviations of all values at the minima and saddle are smaller than 10% when averaged over the instances or over the mini-batches.

		TRAIN ENERGY				TEST ENERGY		TEST ERROR RATE [%]		
		MIN.	SADD.	FACTOR	EPOCH	MIN.	SADD.	MIN.	SADD.	Δ
DATASET	ARCHITECTURE									
C10+	CNN-12	0.5428	0.6381	1.2	78	0.63	0.69	21.4	23.6	2.2
	CNN-24	0.4403	0.5390	1.2	84	0.59	0.64	19.8	21.8	2.0
	CNN-36	0.3982	0.4814	1.2	91	0.57	0.61	19.0	20.8	1.8
	CNN-48	0.3750	0.4331	1.2	**103**	0.56	0.59	18.6	19.9	1.2
	CNN-96	0.3324	0.3900	1.2	**103**	0.55	0.57	17.9	19.3	1.4
	CNN-48X2	0.1402	0.1564	1.1	**136**	0.45	0.46	13.4	14.2	0.8
	CNN-48X3	0.0918	0.1164	1.3	**110**	0.46	0.50	13.4	15.2	1.7
	RESNET-8	0.2045	0.2086	1.0	**136**	0.37	0.38	12.0	12.5	0.4
	RESNET-20	0.0162	0.0324	2.0	**104**	0.36	0.38	8.5	8.9	0.5
	RESNET-32	0.0057	0.0097	1.7	**107**	0.36	0.37	7.5	7.8	0.3
	RESNET-44	0.0031	0.0060	1.9	**122**	0.36	0.36	7.1	7.4	0.3
	RESNET-56	0.0022	0.0141	6.5	85	0.35	0.36	6.9	7.4	0.5
	DENSENET-40-12	0.0008	0.0046	5.9	**205**	0.25	0.25	5.6	6.0	0.3
	DENSENET-100-12-BC	0.0005	0.0026	4.8	**205**	0.21	0.22	4.9	5.1	0.2
C100+	CNN-12	1.6167	1.9029	1.2	69	2.00	2.11	51.0	54.2	3.2
	CNN-24	1.3854	1.6930	1.2	70	1.90	1.99	48.2	51.5	3.3
	CNN-36	1.2670	1.5801	1.2	71	1.87	1.95	47.2	50.2	3.0
	CNN-48	1.1977	1.5002	1.3	73	1.87	1.94	46.6	49.5	2.9
	CNN-96	1.0549	1.3304	1.3	82	1.85	1.91	45.6	48.4	2.8
	CNN-48X2	0.6579	0.8372	1.3	91	1.71	1.73	41.0	42.6	1.6
	CNN-48X3	0.4393	0.6124	1.4	91	1.88	1.89	43.7	45.5	1.8
	RESNET-8	1.0894	1.1547	1.1	**103**	1.46	1.49	39.6	40.6	1.0
	RESNET-20	0.3528	0.5442	1.5	79	1.42	1.48	33.3	34.7	1.4
	RESNET-32	0.1422	0.3576	2.5	77	1.53	1.57	31.5	33.7	2.2
	RESNET-44	0.0753	0.2117	2.8	85	1.60	1.62	30.8	32.5	1.7
	RESNET-56	0.0428	0.2978	7.0	71	1.64	1.66	30.3	32.4	2.0
	DENSENET-40-12	0.0101	0.0808	8.0	166	1.30	1.31	26.3	27.7	1.4
	DENSENET-100-12-BC	0.0050	0.0223	4.4	**205**	1.12	1.13	23.7	24.6	0.8