What Can Linear Interpolation of Neural Network Loss Landscapes Tell Us?

Tiffany Vlaar (University of Edinburgh) and Jonathan Frankle (MosaicML)

ICML 2022

Tiffany.Vlaar@ed.ac.uk
jonathan@mosaicml.com

Research Question

Does the shape of the loss along linear path from initial to final state relate to the "success" of optimization?

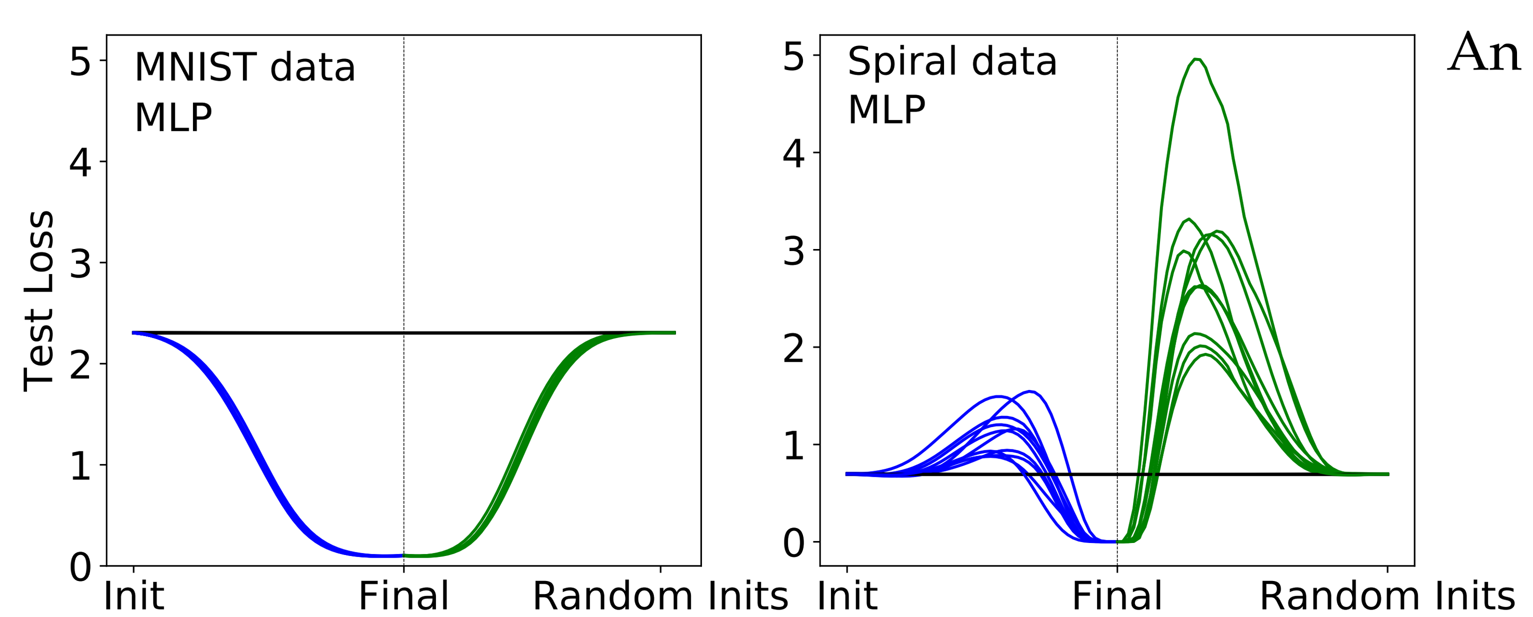
We study the influence of optimizer and architecture design choices on the
1) Shape of the linear path AND 2) Test accuracy of the final model.

Background

Linear interpolation is *"a simple and lightweight method to probe neural network loss landscapes"*
– Lucas et al., 2021

Linear interpolation path between θ_i (initial) and θ_f (final) parameter state:

$$\theta_\alpha = (1-\alpha)\theta_i + \alpha\theta_f \quad \text{for } \alpha \in [0,1] \qquad \text{(Goodfellow et al., 2015)}$$



An absence of barriers along the linear path

⇒ *"tasks are relatively easy to optimize"*
– Goodfellow et al., 2015

⇒ *"Though dimension is high, the space is in some sense simpler than we thought: [...] the walk could just as well have taken a straight line without encountering any obstacles"*
– Li et al., 2018

Linear Interpolation Revisited

In modern neural network architectures: *"Loss plateaus and error remains at the level of random chance ... until near the optimum."*
– Frankle, 2020

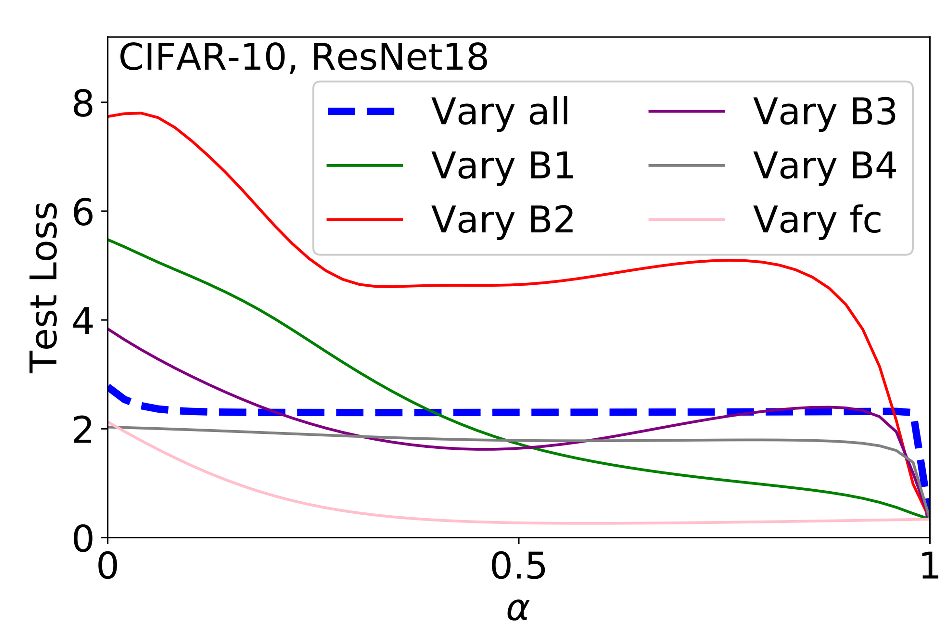
"Networks violating the [Monotonic Linear Interpolation] property can be produced systematically, by encouraging the weights to move far from initialization."
– Lucas et al., 2021

Layer-wise Linear Interpolation

Vary a single layer (or convolutional block) ℓ from initial to final state. Keep all other parameters fixed at their final state.

$$\theta_\alpha^{(\ell)} = (1-\alpha)\theta_0^{(\ell)} + \alpha\theta_f^{(\ell)},$$

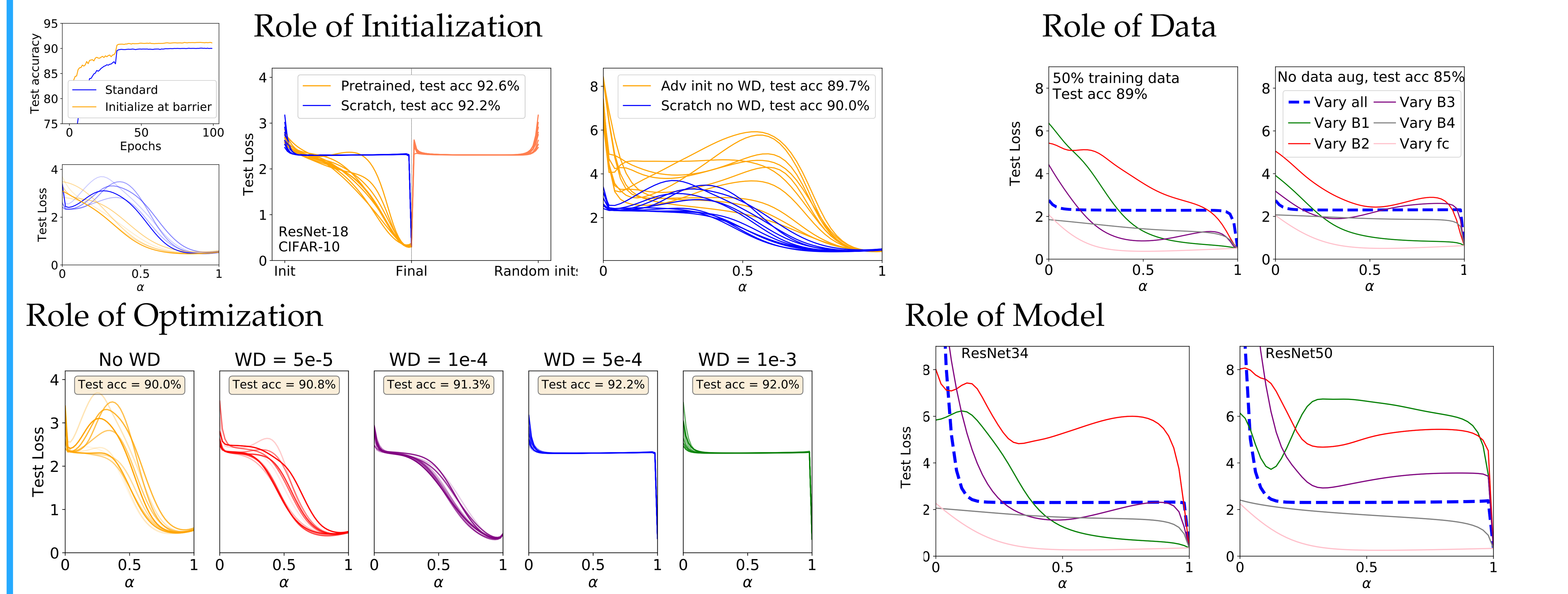
$$\theta_\alpha^{(k)} = \theta_f^{(k)},\ k \neq \ell \qquad \text{(Chatterji et al., 2020)}$$



Base Model: ResNet-18 architecture, CIFAR-10 data.

Findings

The shape of the linear path from initial to final parameter state is NOT a reliable indicator of test accuracy.



Does the shape of loss along the linear path relate to other aspects of optimization?

- Pre-training on ImageNet consistently removes the presence of barriers for ResNet architectures, whereas adversarial initialization on random labels increases barriers.
- Distance between initial and final model state is *not* a reliable indicator of non-monotonic behaviour along linear path.

"pre-trained weights guide the optimization to a flat basin of the loss landscape."
– Neyshabur et al., 2020

"Large distances moved in weight space encourage non-monotonic interpolation"
– Lucas et al., 2021

The Adversarial Effect of Partial Pre-Training

Set model to trained (T) state.

- Re-set specific layer/convolutional block to random initialization (RI).
- Re-train whole model.

⇒ Worse test accuracy!

Model: ResNet-18, *Data:* CIFAR-10.

Method	Test accuracy (%)
Train from scratch	92.2 ±0.2
T-All but RI-1	91.8 ±0.2
T-All but RI-2	91.8 ±0.2
T-All but RI-3	92.4 ±0.2
T-All but RI-4	91.0 ±0.3

Future Work

- Revisit study for attention-based models.
- Further exploration of layer-wise sensitivity to initialization and optimizer hyperparameter settings.
If interested, also see Vlaar & Leimkuhler, Multirate Training of Neural Networks, ICML 2022.