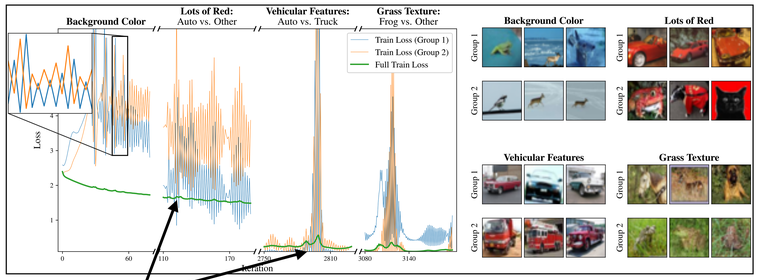
Outliers with Opposing Signals Have an Outsized Effect on Neural Network Optimization

Elan Rosenfeld Andrej Risteski

- We uncover **simple and consistent behavior** in NN optimization which naturally fits prior observations.
- It is induced by paired groups of outliers which cause **large gradients** pointing in **opposite directions**.
- We refer to them as ***Opposing Signals***.

ResNet-18 Trained with Gradient Descent on CIFAR-10



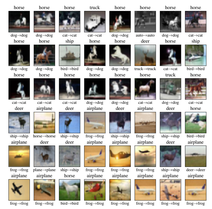
Edge of stability *exactly coincides* with outlier loss spikes

Outlier groups increase in complexity over time

Is this specific to…

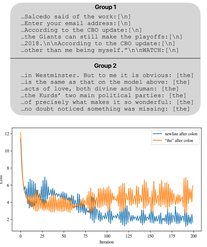
ConvNets?

No. Same occurs on Vision Transformers.



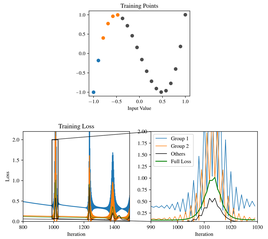
Vision?

No. Same occurs on Transformers doing next-token prediction.



Cross-Entropy?

No. Same occurs fitting a polynomial with square loss.



We identify a new phenomenon which offers a cohesive explanation and a possible ***common cause*** for:

Observations like

Edge of Stability	Progressive Sharpening
Grokking	Training Instabilities
Simplicity Bias	Spectrum Outliers

And the benefits of

Adaptive Methods	Sharpness-Aware Minimization
Batch Normalization	Large Learning Rate
Dropout	Weight Decay

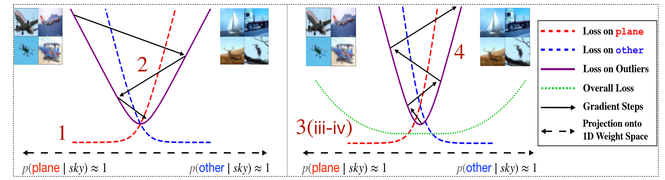
Paper:

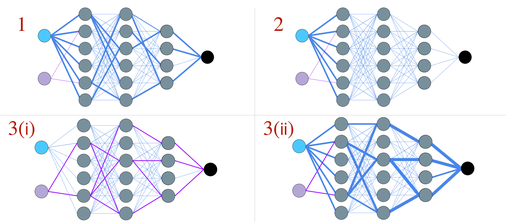
Talk:

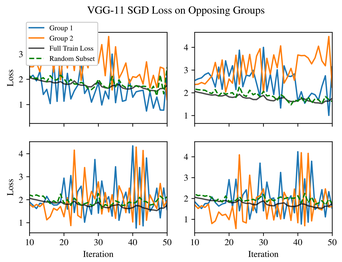
MACHINE LEARNING DEPARTMENT — Carnegie Mellon University

Our current (incomplete) understanding:

1. At initialization, activations dominated by simple, large magnitude features (e.g. "**sky background**") → large gradients.
   - **Loss on planes** points one way, **loss on non-planes** points in the **opposite direction**. Combined, they form a **narrow valley**.
2. Early optimization approaches the minimum, balancing the opposing signals, and proceeds "through the valley".
3. *(i)* **Align/amplify** subnetworks → *(ii)* **Increase magnitude** of opposing signals → *(iii)* **Losses steepen** → *(iv)* **Valley *sharpens***.
4. Once step size is too large for curvature, iterates diverge. Then:
   - Opposite group losses **grow and oscillate**;
   - Opposing signals **decrease in magnitude**, flattening the valley.







Outliers dominate SGD training dynamics

Enables precise predictions of network behavior