NEURAL INFORMATION PROCESSING SYSTEMS

PRINCETON UNIVERSITY

Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction

Kaifeng Lyu, Zhiyuan Li, Sanjeev Arora

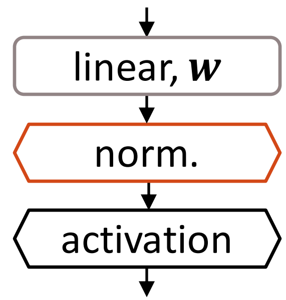
Princeton University

Normalization and Weight Decay

Common Practice: **Normalization Layers**
- BatchNorm, LayerNorm, …
- Used in ResNets, Transformers, …

Scale-invariance:

Rescaling weights $\boldsymbol{w} \mapsto c\boldsymbol{w}$ $(c > 0)$ does not change the loss



+ minor change to the architecture
$\Downarrow$
true for all params

Common Practice: **Weight Decay (WD)**
- penalizing large parameter norm
- equivalent to L^2-Regularization

- **WD** has no explicit regularization effect
- but still helps generalization [Zhang et al., 2019; Lewkowycz & Gur-Ari, 2020; Liu et al., 2020]

Gradient Descent with WD

Goal: Improve mathematical understanding of how normalization + WD improves generalization

Our Setup

- (full-batch) GD with WD

$$w_{t+1} \leftarrow (1-\eta\lambda)w_t - \eta\nabla\mathcal{L}(w_t)$$

- Scale-invariant loss wrt all params:

$$\mathcal{L}(cw) = \mathcal{L}(w), \text{ for all } c > 0$$

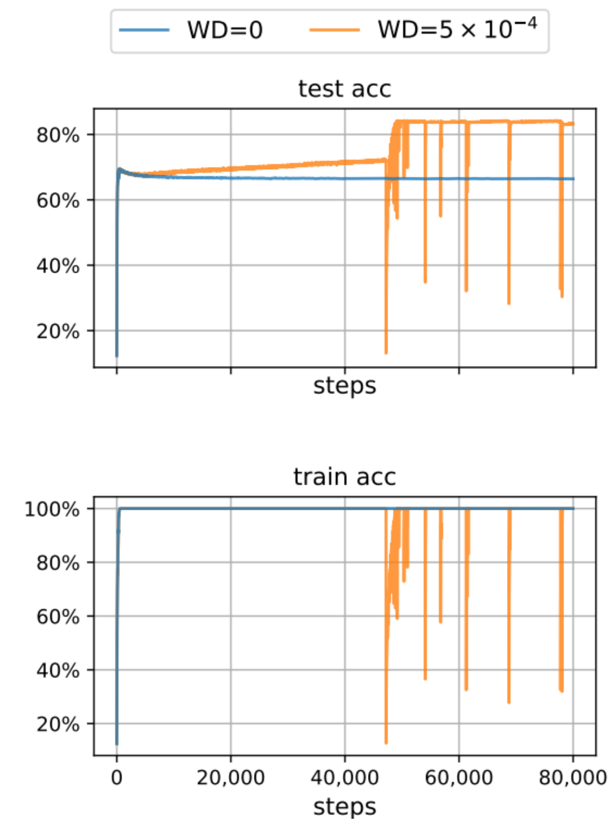
Motivating Phenomenon

Case 1: With normalization + WD:

- the net **continues to evolve** even after train acc = 100%
- test acc improves a lot
- **69.1%** → **72.0%**, and → **84.3%** after destabilization

Case 2: Removing either normalization or WD:

- test acc **does not change much** after train acc = 100%

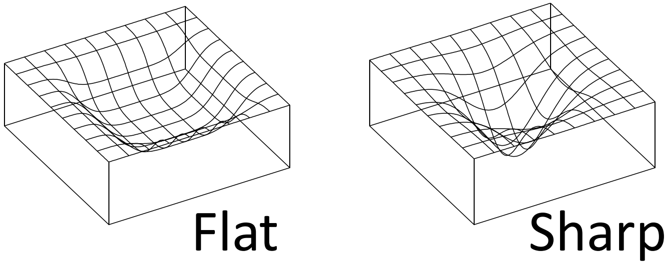


CIFAR-10 + VGG + BN (no data augmentation)

Our Theory: Sharpness Reduction

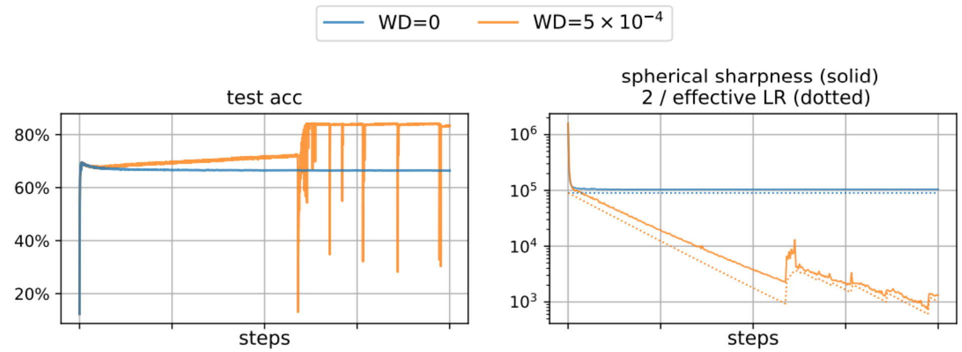
Long-held belief: flatter minima generalize better

- [Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Neyshabur et al., 2017]



Spherical Sharpness:

- A sharpness measure of the solution
- = the top eigenvalue of $\nabla^2\mathcal{L}$ after projecting the weights to $\mathbb{S}^{d-1}$



Sharpness Reduction Phenomenon: In the late phase of training, the spherical sharpness drops and test acc rises.

Our Theory: Sharpness-Reduction Flow

Our Setup

- Assume a manifold Γ of minima on $\mathbb{S}^{d-1}$
- Start our analysis near the manifold
  - Use previous analysis for loss convergence [Li et al., 2022]

Theorem 1. Eventually enters the **Edge of Stability** regime

$$\lambda_1\left(\nabla^2 \mathcal{L}(\boldsymbol{w}_t)\right) \approx 2/\eta$$

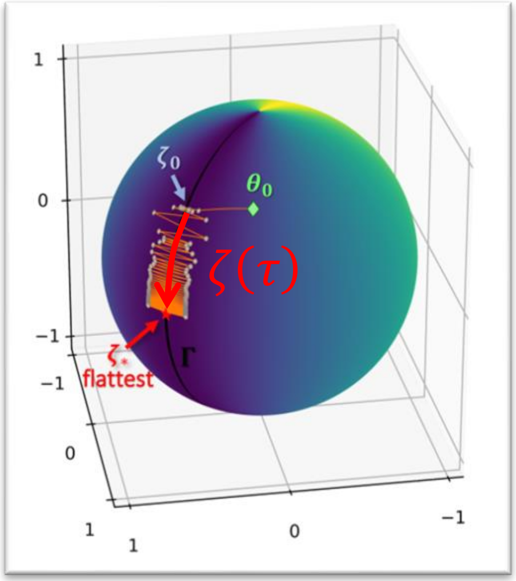
Loss no longer decreases monotonically.

Theorem 2. The projected parameter $\boldsymbol{\theta}_t := \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$

1. oscillates around the manifold
2. moves along a continuous flow on the manifold

Sharpness-Reduction Flow: $$\frac{\mathrm{d}\boldsymbol{\zeta}(\tau)}{\mathrm{d}\tau} = -\frac{2\nabla_\Gamma \log \lambda_1\left(\nabla^2 \mathcal{L}(\boldsymbol{\zeta})\right)}{4+\left\|\nabla_\Gamma \log \lambda_1\left(\nabla^2 \mathcal{L}(\boldsymbol{\zeta})\right)\right\|^2}$$

∇_Γ: Projection of gradient to the tangent space of Γ



Summary

We show that the interplay of normalization and WD results in a sharpness reduction flow that can promote generalization.

See our paper for more:

1. Understanding the **Edge of Stability** Phenomenon [Cohen et al., 2021]
2. Connecting and generalizing our results to **adaptive learning rate** methods (e.g., RMSprop)
3. Extensive empirical validation of sharpness reduction on CIFAR-10