Indian Institute of Science
Bangalore, India
भारतीय विज्ञान संस्थान
बंगलौर, भारत

NEURAL INFORMATION PROCESSING SYSTEMS

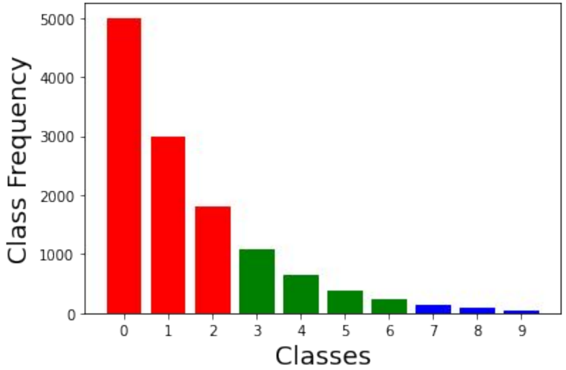
Escaping Saddle Points for Effective Generalization on Class-Imbalanced Data

Harsh Rangwani*, **Sumukh K Aithal***, Mayank Mishra, R. Venkatesh Babu

Indian Institute of Science, Bengaluru

Long-Tailed Learning

Training data

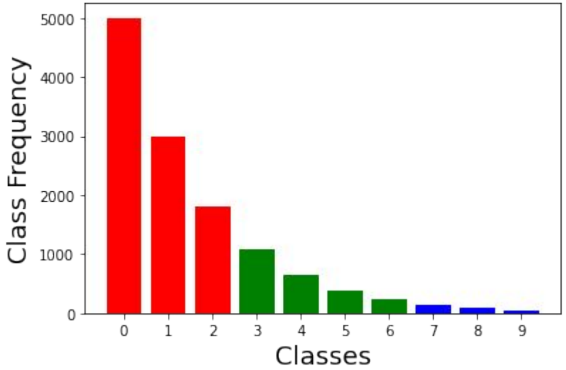


Long tailed distribution

Natural datasets are often imbalanced in terms of the frequency of samples in each class.

Long-Tailed Learning

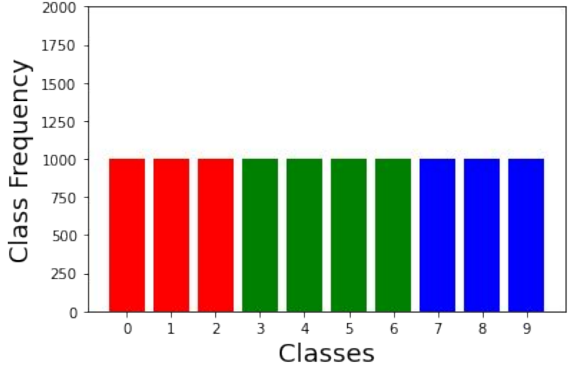
Training data



Long tailed distribution

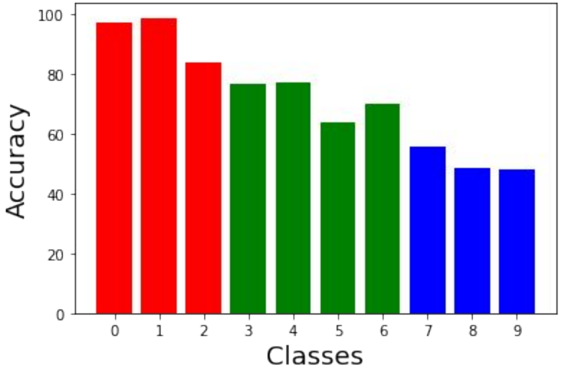
Model

Test data



Balanced distribution

Class-Wise Accuracy on Balanced Test Set



The performance of neural networks degrades significantly on the minority class samples.

Long-Tailed Learning

In this work, we primarily focus on analyzing the nature of solutions of loss manipulation methods.

1. **Cross-Entropy + Deferred-Reweighting (CE+DRW)[1]**
   Re-weight the CE loss based on the inverse of number of samples in each class. (Minority class samples are given more weight)

[1] *Cao, Kaidi, et al. "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss." NeurIPS 2019.*

Long-Tailed Learning

In this work, we primarily focus on analyzing the nature of solutions of loss manipulation methods.

1. **Cross-Entropy + Deferred-Reweighting (CE+DRW)[1]**
   Re-weight the CE loss based on the inverse of number of samples in each class. (Minority class samples are given more weight)

2. **LDAM (Margin Based Loss)[1]**
   Regularize the minority class samples more (Larger margin) compared to the majority class samples.

3. **Vector Scaling Loss (VS)[2]**
   Combination of Multiplicative (Class Dependent Temperature) and Additive Adjustments (Logit Adjustment)

[1]*Cao, Kaidi, et al. "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss." NeurIPS 2019.*

[2]*Kini, Ganesh Ramachandra, et al. "Label-imbalanced and group-sensitive classification under overparameterization." NeurIPS 2021*

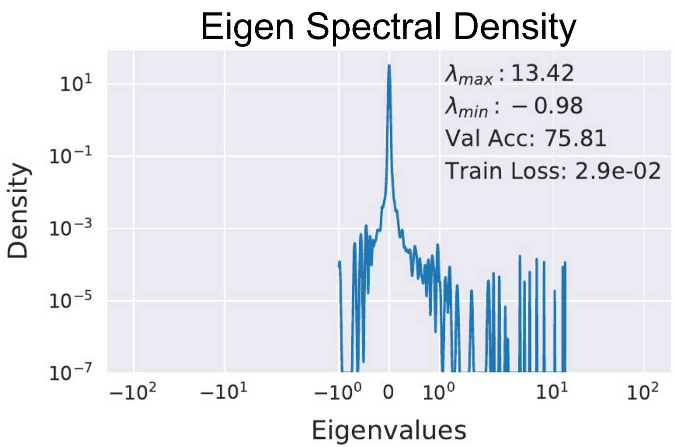
Loss Landscape

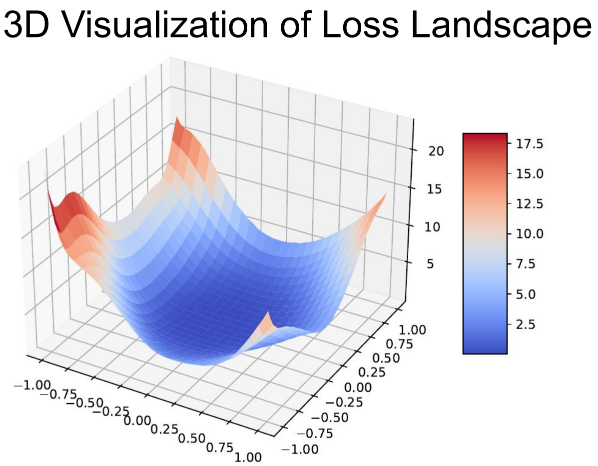
- The Hessian of the training loss can be used to analyze the nature of converged solution and the dynamics of optimization in deep neural networks.

Loss Landscape

- The Hessian of the training loss can be used to analyze the nature of minima and the dynamics of optimization in deep neural networks.

- Eigenvalues of the Hessian (Eigen Spectral Density) characterize the local curvature of the loss at the solution.





Loss Landscape

- The Hessian of the training loss can be used to analyze the nature of minima and the dynamics of optimization in deep neural networks.
- Eigenvalues of the Hessian (Eigen Spectral Density) characterize the local curvature of the loss at the solution.
- Geometry of the loss landscape is correlated with generalization. For example, flat minima generalizes better than sharp minima.[1]

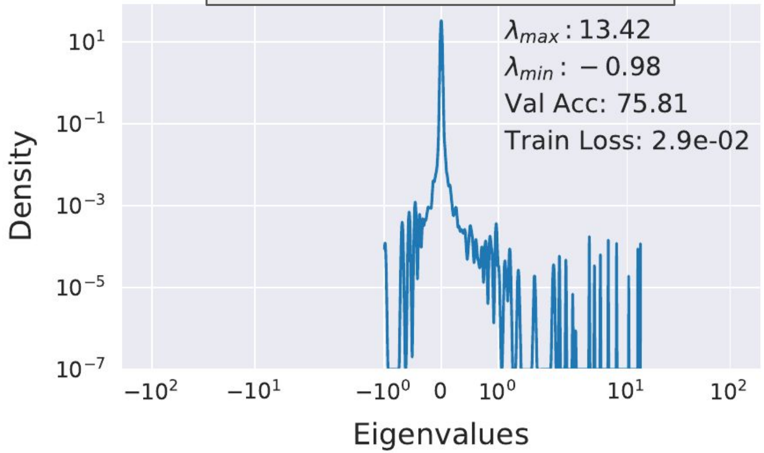
[1] *Keskar, Nitish Shirish, et al. "On large-batch training for deep learning: Generalization gap and sharp minima." ICLR 2017.*

Class-Wise Loss Landscape Analysis in Imbalanced Datasets

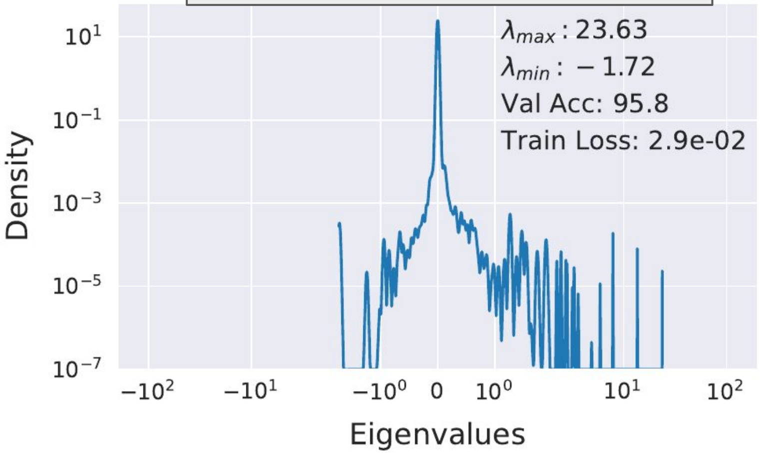
- **Prior Work**: Hessian of the average loss (Eigen Spectral Density) is used to characterize the nature of minima.

- On imbalanced datasets, this analysis is not very useful as it indicates convergence to **local minima** and **imitates the head class**.

Overall Spectral Density



Head Class Spectral Density



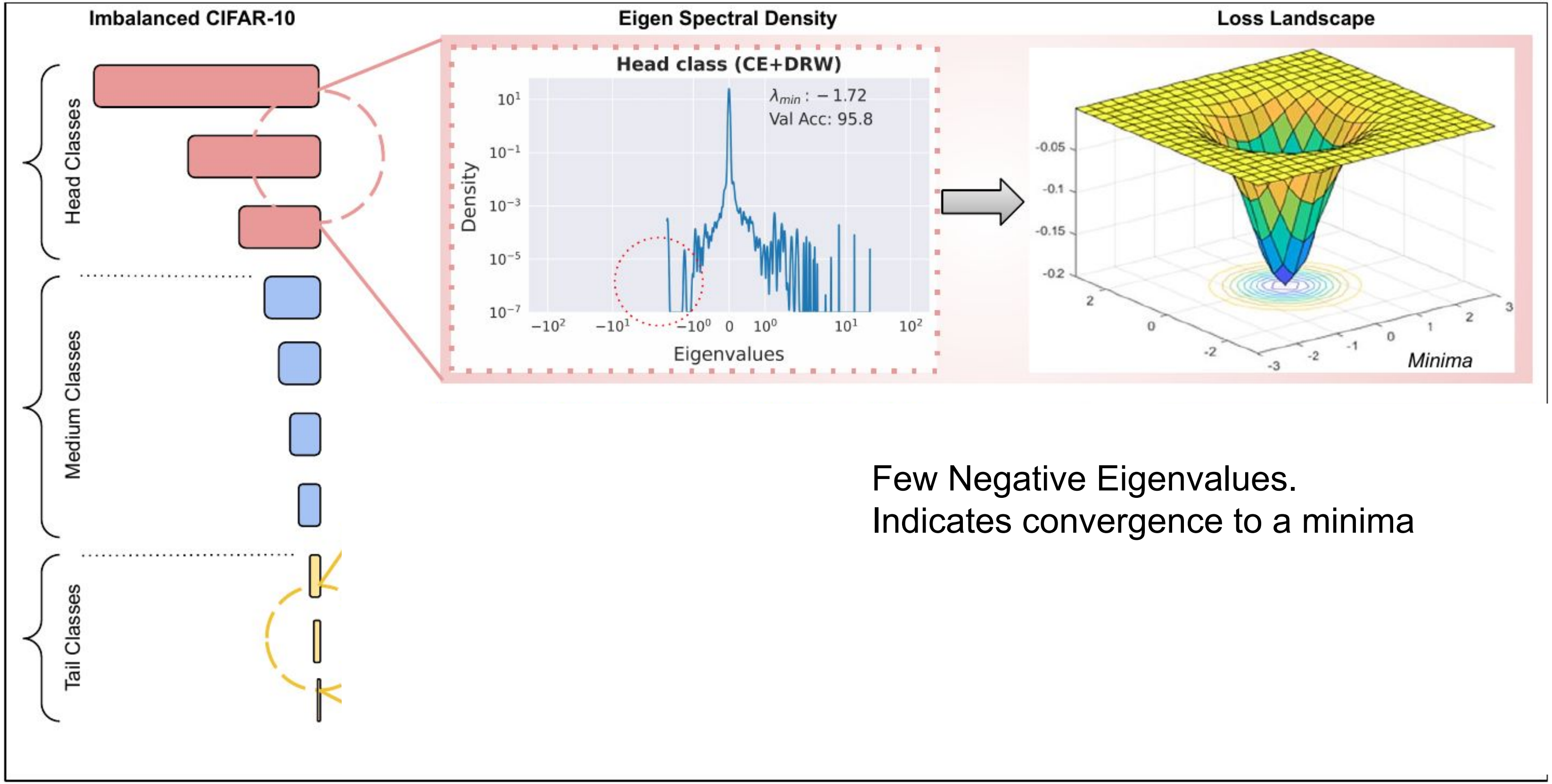
Class-Wise Loss Landscape Analysis in Imbalanced Datasets

- **Prior Work**: Hessian of the average loss (Eigen Spectral Density) is used to characterize the nature of minima.

- On imbalanced datasets, this analysis is not very useful as it indicates converges to **local minima** and **imitates the head class**.

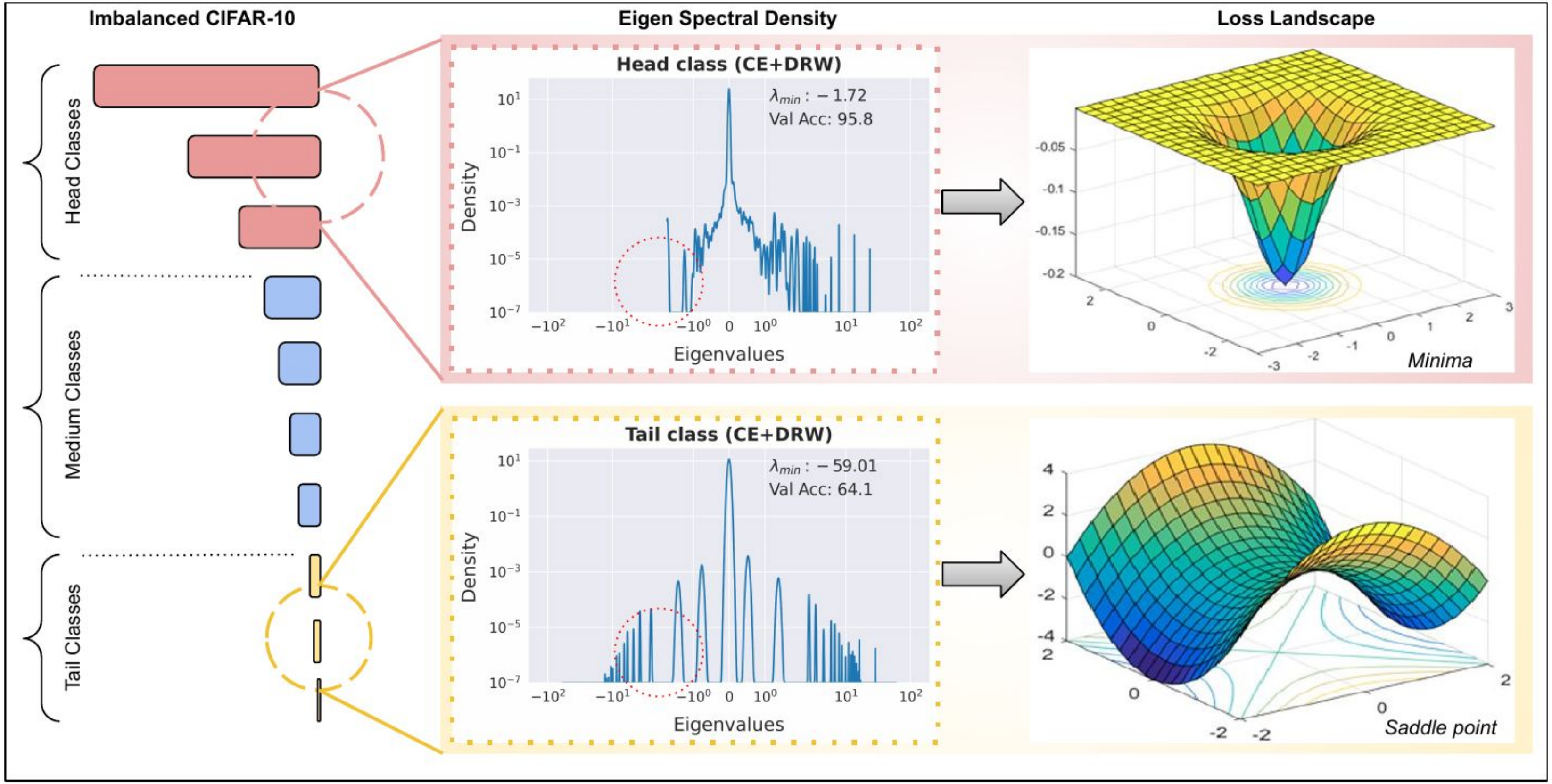
Our work*: Class-Wise Analysis of loss landscape on imbalanced datasets uncovers interesting insights.*

Convergence to Saddle Points in Tail Class Loss Landscape



Few Negative Eigenvalues.
Indicates convergence to a minima

Convergence to Saddle Points in Tail Class Loss Landscape



Escaping Saddle Points Improves Generalization

- Due to the occurrence of saddle points, we observe that the network suffers from poor generalization on minority classes.

- Prior work on escaping saddle points includes methods like Perturbed Gradient Descent (PGD)[1] which are not commonly used to train deep neural networks.

[1]Jin, Chi, et al. "How to escape saddle points efficiently." International Conference on Machine Learning. PMLR, 2017.

Escaping Saddle Points Improves Generalization

- Due to the occurrence of saddle points, we observe that the network suffers from poor generalization on minority classes.
- Prior work on escaping saddle points includes methods like Perturbed Gradient Descent (PGD)[1] which are not commonly used to train deep neural networks.
- Sharpness-Aware Minimization (SAM)[2] is a recently proposed optimizer with an objective to explicitly find a flat minima with a low loss.
- We show that **Sharpness-Aware Minimization (SAM)** can also escape saddle points and lead to improved generalization particularly on the tail classes.

[1] *Jin, Chi, et al. "How to escape saddle points efficiently." International Conference on Machine Learning. PMLR, 2017.*

[2] *Foret, Pierre, et al. Sharpness-aware minimization for efficiently improving generalization. In ICLR, 2021*

Analysis of SAM for Escaping Saddle Points

f : Objective Function
ρ: Neighborhood size

SAM: $$\min_{w} \max_{||\epsilon|| \leq \rho} f(w+\epsilon)$$

First step: Find a sharp maximal point ϵ in the neighborhood of the weights w.

Second step: Minimize the loss at this sharp maximal point.

$$\hat{\epsilon}(w) \approx \arg\max_{||\epsilon|| \leq \rho} f(w) + \epsilon^T \nabla f(w) \;= \rho \nabla f(w) / ||\nabla f(w)||_2$$

Analysis of SAM for Escaping Saddle Points

f : Objective Function
ρ: Neighborhood size

SAM:

$$\min_{w} \max_{||\epsilon|| \leq \rho} f(w+\epsilon)$$

First step: Find a sharp maximal point ϵ in the neighborhood w.

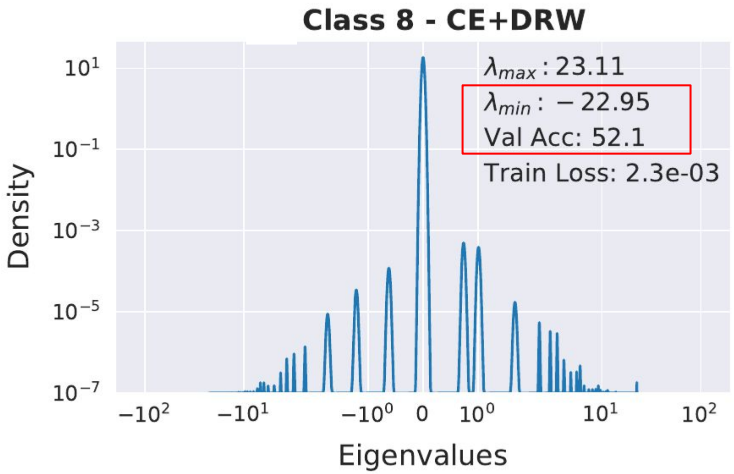
Second step: Minimize the loss at this sharp maximal point.

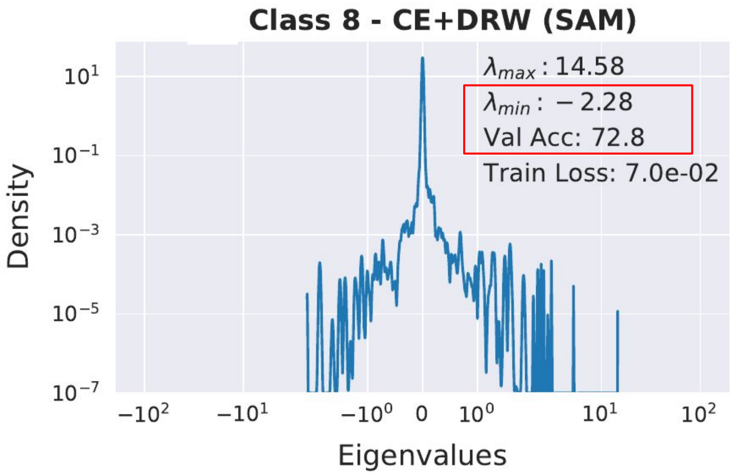
$$\hat{\epsilon}(w) \approx \arg\max_{||\epsilon|| \leq \rho} f(w) + \epsilon^T \nabla f(w) \ = \rho \nabla f(w) / ||\nabla f(w)||_2$$

Informal Theoretical Explanation:
We theoretically show that the SAM amplifies the gradient component along the negative curvature. This helps SAM to effectively escape saddle points.

Escaping Saddle Points Improves Generalization

With SAM (high ρ), the large negative eigenvalues present in the loss landscape of the tail class get suppressed. (i.e no more saddle point)

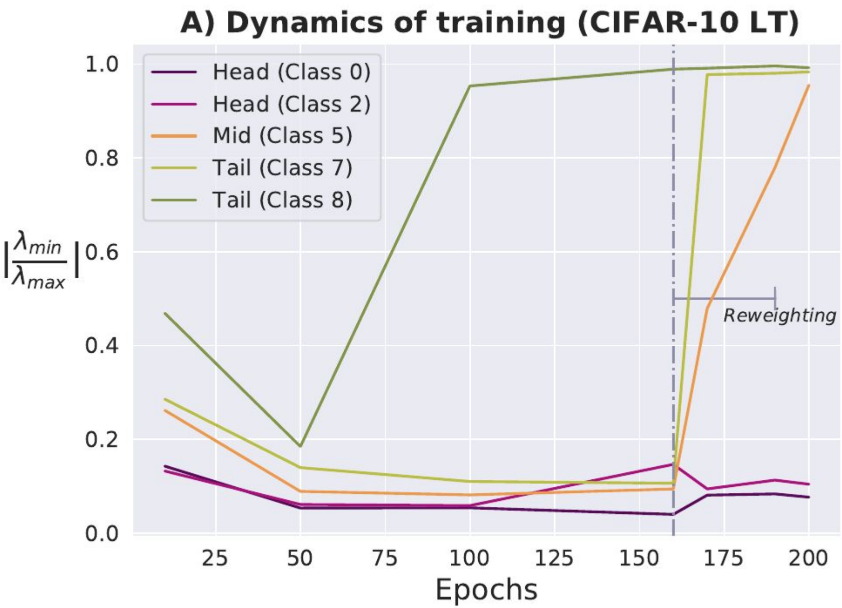




Eigen Spectral Density of **Tail Classes** with SGD (left) and SAM (right)

Dynamics of Training on Long-Tailed datasets

With SGD, network converges to non-convex regions with negative curvature for tail classes.

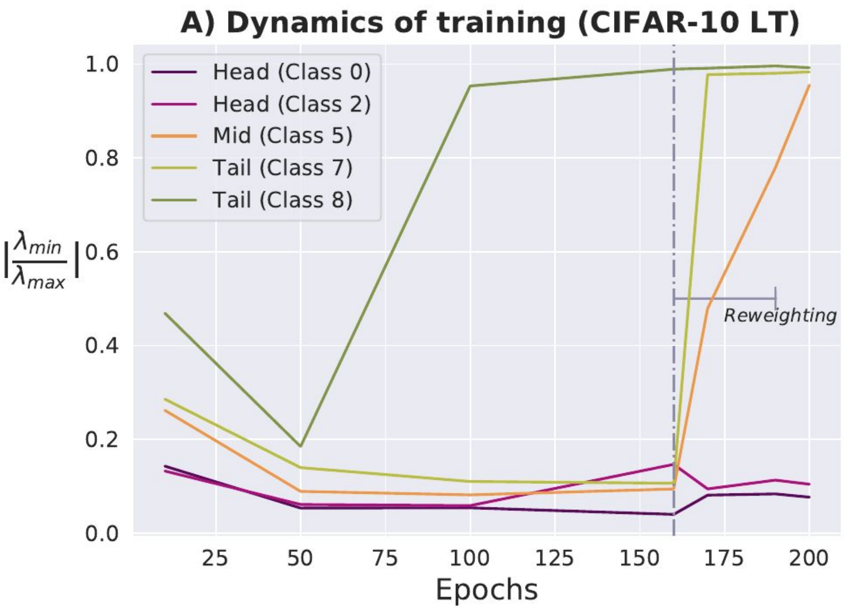


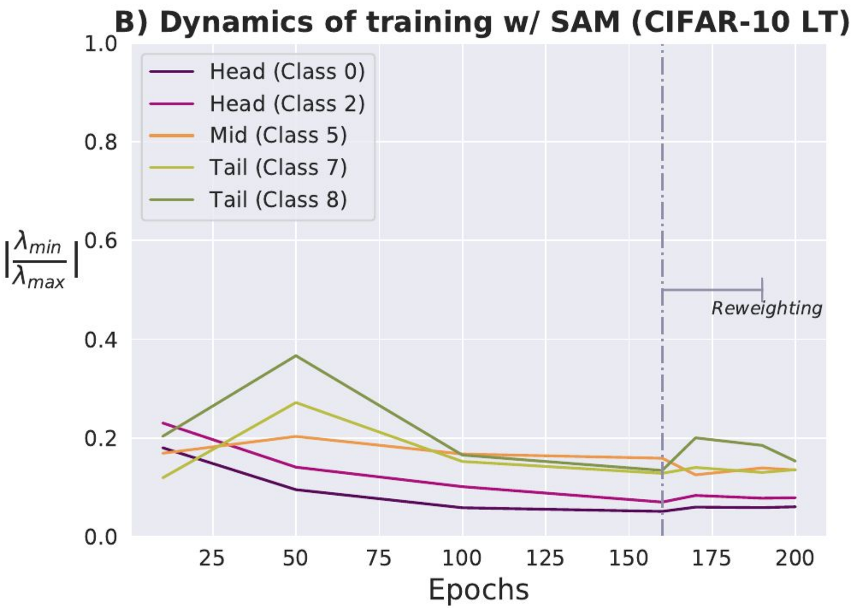
We find that with reweighting and margin enhancement is the main culprit, which forces model into non-convex regions leading to saddle points.

$|\frac{\lambda_{min}}{\lambda_{max}}|$: Measure of non-convexity of the loss landscape. (High value indicates non-convex regions)

Dynamics of Training on Long-Tailed datasets

SAM does not allow the tail classes to reach a region of high non-convexity.





$|\frac{\lambda_{min}}{\lambda_{max}}|$: Measure of non-convexity of the loss landscape. (High value indicates non-convex regions)

Results on CIFAR-10 LT and CIFAR-100 LT

	CIFAR-10 LT				CIFAR-100 LT			
	Acc	Head	Mid	Tail	Acc	Head	Mid	Tail
CE	$71.7_{\pm 0.1}$	$90.8_{\pm 3.6}$	$71.9_{\pm 0.4}$	$52.3_{\pm 3.7}$	$38.5_{\pm 0.5}$	$64.5_{\pm 0.7}$	$36.8_{\pm 1.0}$	$8.2_{\pm 1.0}$
CE + SAM	$73.1_{\pm 0.3}$	$93.3_{\pm 0.2}$	$74.1_{\pm 0.6}$	$51.7_{\pm 1.0}$	$39.6_{\pm 0.6}$	$66.5_{\pm 0.7}$	$38.1_{\pm 1.1}$	$8.0_{\pm 0.6}$
CE + DRW [8]	$75.5_{\pm 0.2}$	$91.6_{\pm 0.4}$	$74.1_{\pm 0.4}$	$61.4_{\pm 0.9}$	$41.0_{\pm 0.6}$	$61.3_{\pm 1.3}$	$41.7_{\pm 0.5}$	$14.7_{\pm 0.9}$
CE + DRW + SAM	$80.6_{\pm 0.4}$	$91.4_{\pm 0.3}$	$78.0_{\pm 0.4}$	$73.1_{\pm 0.9}$	$44.6_{\pm 0.4}$	$61.2_{\pm 0.8}$	$47.5_{\pm 0.6}$	$20.7_{\pm 0.6}$

5.1% ↑   3.6% ↑

DRW + SAM improves upon the overall performance of CE+DRW by 5.1% on CIFAR-10 LT and 3.6% on CIFAR-100 LT datasets, with the tail class accuracy increasing by 11.7% and 7.7% respectively.

Results on CIFAR-10 LT and CIFAR-100 LT

	CIFAR-10 LT				CIFAR-100 LT			
	Acc	Head	Mid	Tail	Acc	Head	Mid	Tail
CE	$71.7_{\pm0.1}$	$90.8_{\pm3.6}$	$71.9_{\pm0.4}$	$52.3_{\pm3.7}$	$38.5_{\pm0.5}$	$64.5_{\pm0.7}$	$36.8_{\pm1.0}$	$8.2_{\pm1.0}$
CE + SAM	$73.1_{\pm0.3}$	$93.3_{\pm0.2}$	$74.1_{\pm0.6}$	$51.7_{\pm1.0}$	$39.6_{\pm0.6}$	$66.5_{\pm0.7}$	$38.1_{\pm1.1}$	$8.0_{\pm0.6}$
CE + DRW [8]	$75.5_{\pm0.2}$	$91.6_{\pm0.4}$	$74.1_{\pm0.4}$	$61.4_{\pm0.9}$	$41.0_{\pm0.6}$	$61.3_{\pm1.3}$	$41.7_{\pm0.5}$	$14.7_{\pm0.9}$
CE + DRW + SAM	$80.6_{\pm0.4}$	$91.4_{\pm0.3}$	$78.0_{\pm0.4}$	$73.1_{\pm0.9}$	$44.6_{\pm0.4}$	$61.2_{\pm0.8}$	$47.5_{\pm0.6}$	$20.7_{\pm0.6}$
LDAM + DRW [8]	$77.5_{\pm0.5}$	$91.1_{\pm0.8}$	$75.7_{\pm0.7}$	$66.4_{\pm0.2}$	$42.7_{\pm0.3}$	$61.8_{\pm0.6}$	$42.2_{\pm1.5}$	$19.4_{\pm0.9}$
LDAM + DRW + SAM	$81.9_{\pm0.4}$	$91.0_{\pm0.2}$	$79.2_{\pm0.5}$	$76.4_{\pm1.1}$	$45.4_{\pm0.1}$	$64.4_{\pm0.3}$	$46.2_{\pm0.2}$	$20.8_{\pm0.3}$
VS [30]	$78.6_{\pm0.3}$	$90.6_{\pm0.4}$	$75.8_{\pm0.5}$	$70.3_{\pm0.5}$	$41.7_{\pm0.5}$	$54.4_{\pm0.2}$	$41.1_{\pm0.6}$	$26.8_{\pm1.0}$
VS + SAM	$82.4_{\pm0.4}$	$90.7_{\pm0.0}$	$79.6_{\pm0.5}$	$78.0_{\pm01.2}$	$46.6_{\pm0.4}$	$56.4_{\pm0.4}$	$48.8_{\pm0.6}$	$31.7_{\pm0.1}$

Integrating SAM with state-of-the-art techniques for long-tailed learning (LDAM, VS) leads to significant gains in overall accuracy primarily due to the major gain in the accuracy on the tail classes.

Results on Large Scale Datasets

- Problem of saddle points also exists in large datasets.
- SAM is easily generalizable to large-scale imbalanced datasets without any changes.

Method		iNaturalist 2018				ImageNet-LT			
	Two stage	Acc	Head	Mid	Tail	Acc	Head	Mid	Tail
CE	✗	60.3	72.8	62.7	54.8	42.7	62.5	36.6	12.5
cRT [27] †	✓	68.2	73.2	68.8	66.1	50.3	62.5	47.4	29.5
LWS [27] †	✓	69.5	71.0	69.8	68.8	51.2	61.8	48.6	33.5
MiSLAS [57]	✓	**71.6**	73.2	**72.4**	70.4	52.7	61.7	51.3	**35.8**
DisAlign [55]	✓	69.5	61.6	70.8	69.9	52.9	61.3	**52.2**	31.4
DRO-LT [44]	✗	69.7	**73.9**	70.6	68.9	**53.5**	**64.0**	49.8	33.1
CE + DRW	✗	63.0	59.8	64.4	62.3	44.9	57.9	42.2	21.6
CE + DRW + SAM	✗	65.3	60.5	66.2	65.5	47.1	56.6	45.8	28.1
LDAM + DRW	✗	67.5	63.0	68.3	67.8	49.9	61.1	48.2	28.3
LDAM + DRW + SAM	✗	70.1	64.1	70.5	**71.2**	53.1	62.0	52.1	34.8

Summary and Conclusion

1. Training on imbalanced datasets can lead to convergence to points with sufficiently **large negative curvature** in the loss landscape for **minority classes**.

Summary and Conclusion

1. Training on imbalanced datasets can lead to convergence to points with sufficiently **large negative curvature** in the loss landscape for **minority classes**.

2. We propose to use **SAM** with a high regularization factor ρ as an effective method to escape regions of negative curvature and **enhance the generalization performance**.

Summary and Conclusion

1. Training on imbalanced datasets can lead to convergence to points with sufficiently **large negative curvature** in the loss landscape for **minority classes**.

2. We propose to use **SAM** with a high regularization factor ρ as an effective method to escape regions of negative curvature and **enhance the generalization performance**.

3. Results on various datasets with different long-tail learning methods indicate that the proposed method is **generic** and **improves base method** significantly.

Thank You