

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

Elan Rosenfeld

Andrej Risteski

Deep Learning: Classics and Trends

1/26/2024

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking

The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the *Grokking Phenomenon*

Vimal Thilak
vthilak@apple.com

Etai Littwin
elittwin@apple.com

Shuangfei Zhai
szhai@apple.com

Omid Saremi
osaremi@apple.com

Roni Paiss
rpaiss@apple.com

Joshua Susskind
jsusskind@apple.com

Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit

Boaz Barak
Harvard University

Benjamin L. Edelman
Harvard University

Surbhi Goel
Microsoft Research &
University of Pennsylvania

Sham Kakade
Harvard University

Eran Malach
Hebrew University of Jerusalem

Cyril Zhang
Microsoft Research

A TALE OF TWO CIRCUITS: GROKING AS COMPETITION OF SPARSE AND DENSE SUBNETWORKS

William Merrill*, Nikolaos Tsilivis* & Aman Shukla
New York University

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR

The Slingshot Mechanism: An Empirical Study of Adaptive Optimization with Gradient Descent

Vimal Thilak
vthilak@apple.com

Omid Saremi
osaremi@apple.com

THE BREAK-EVEN POINT ON OPTIMIZATION TRAJECTORIES OF DEEP NEURAL NETWORKS

Stanisław Jastrzębski¹, Maciej Szymczak², Stanislav Fort³, Devansh Arpit⁴, Jacek Tabor², Kyunghyun Cho^{1,5,6*}, Krzysztof Geras^{1*}

¹Harvard University

²Hebrew University of Jerusalem

Learning:
Computational Limit

Surbhi Goel
Microsoft Research &
University of Pennsylvania

Cyril Zhang
Microsoft Research

Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhi@andrew.cmu.edu

Colin Wei
Computer Science Department
Stanford University
colinwei@stanford.edu

Tengyu Ma
Computer Science Department
Stanford University
tengyuma@stanford.edu

CIRCUITS: GROKING AS COMPETITIVE DENSE SUBNETWORKS

HOW DOES LEARNING RATE DECAY HELP MODERN NEURAL NETWORKS?

Kaichao You *
School of Software
Tsinghua University
youkaichao@gmail.com

Mingsheng Long (✉)
School of Software
Tsinghua University
mingsheng@tsinghua.edu.cn

Jianmin Wang
School of Software
Tsinghua University
jimwang@tsinghua.edu.cn

Michael I. Jordan
Department of EECS
University of California, Berkeley
jordan@cs.berkeley.edu

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm

The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the Grokking Phenomenon

Vimal Thilak, Eyal Litvin, Shuang Li, and Ilya Sutskever

Omid Saremi, Roni Paiss, and Joshua Susskind

Towards Explaining the Initial Large Learning Rate Phenomenon in Self-Supervised Pre-training

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhi1@andrew.cmu.edu

TOWARDS UNDERSTANDING REGULARIZATION IN BATCH NORMALIZATION

Ping Luo^{1,3*}, Xinjiang Wang^{2*}, Wenqi Shao^{1*}, Zhanglin Peng²
¹The Chinese University of Hong Kong ²SenseTime Research ³The University of Hong Kong

Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction

Kaifeng Lyu, Zhiyuan Li, Sanjeev Arora
Department of Computer Science
Princeton University
{klyu, zhiyuanli, arora}@cs.princeton.edu

How Does Batch Normalization Help Optimization?

Shibani Santurkar*, Dimitris Tsipras*, Andrew Ilyas*, Aleksander Mądry
MIT MIT MIT MIT
shibani@mit.edu tsipras@mit.edu ailyas@mit.edu mardy@mit.edu

Sham Kakade, Eran Malach, Cyril Zhang
Harvard University Hebrew University of Jerusalem Microsoft Research

HELP MODERN

Mingsheng Long (鸣盛)
School of Software
Tsinghua University
mingsheng@tsinghua.edu.cn

Michael I. Jordan
Department of EECS
University of California, Berkeley
jordan@cs.berkeley.edu

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm
- Hessian Spectrum
- Outliers

The Slingshot Mechanism of Adaptive Optimizers

Vimal Thilak
vthilak@apple.com

Omid Saremi
osaremi@apple.com

THE

TORII

Stanislaw Jastrzebski¹,
Kyunghyun Cho^{1,5,6*}, K

Measurements of Three-Level Hierarchical Structure
in the Outliers in the Spectrum of Deepnet Hessians

Vardan Papyan¹

ization?

ming:

tational Limit
ander Madry
MIT
ry@mit.edu

Towards Explaining the Regularization
Initial Large Learning Rate in Training
Networks

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University

Ping Luo^{1,3*},
The Chinese University of Hong Kong

Levent Sagun
Mathematics Department
New York University
sagun@cims.nyu.edu

Léon Bottou
Facebook AI Research
New York
leon@bottou.org

Yann LeCun
Computer Science Department
New York University
yann@cs.nyu.edu

The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping
from Sharp Minima and Regularization Effects

Zhanxing Zhu^{*123} Jingfeng Wu^{*1} Bing Yu¹ Lei Wu¹ Jinwen Ma¹

Mingsheng Long (✉)
School of Software
Tsinghua University
mingsheng@tsinghua.edu.cn

Michael I. Jordan
Department of EECS
University of California, Berkeley
jordan@cs.berkeley.edu

Kaifeng Lyu
Department
Pr
{klyu, zhiyuan}

Traces of Class/Cross-Class Structure
Pervade Deep Learning Spectra

Vardan Papyan
Department of Statistics
Stanford University
Stanford, CA 94305, USA

PAPYAN@STANFORD.EDU

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm
- Hessian Spectrum
- Outliers
- Sharpening/EoS

The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the Grokking Phenomenon

Vimal Thilak
vthilak@apple.com

Omid Saremi
osaremi@apple.com

Etai Littwin

Microsoft Research

Roni Paiss

Stanislaw Jastrzebski^{1,5,*}, Marcin Szpandzik^{1,5,6*}, Krzysztof Kurtyka^{1,5,6*}

Shuangfei Zhai

Microsoft Research

Joshua Susskind

Microsoft Research

SELF-STABILIZATION: THE IMPLICIT BIAS OF GRADIENT DESCENT AT THE EDGE OF STABILITY

Alex Damian*, Eshaan Nichani* & Jason D. Lee
Princeton University
(ad27, eshnich, jasonlee)@princeton.edu

Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability

Zhouzi Li*
IIS, Tsinghua University
zhouzi188763@gmail.com

Zixuan Wang*
IIS, Tsinghua University
wangzx2019012326@gmail.com

Jian Li*
IIS, Tsinghua University
lapordge@gmail.com

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhi1@andrew.cmu.edu

Ping Li
The Chinese University of Hong Kong

The Anisotropic Noise in Stochastic Gradient Descent from Sharp Minima

Understanding Gradient Descent on the Edge of Stability in Deep Learning

Sanjeev Arora^{*1}, Zhiyuan Li^{*1}, Abhishek Panigrahi^{*1}

Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction

Zhanxing Zhu^{1,2,3}, Jingfeng Wu¹, Bing Yu¹, Lei Wu¹, Jinwen Ma¹
1 Tsinghua University, 2 Tsinghua University, 3 Tsinghua University
zhanxing_zhu@tsinghua.edu.cn

Michael I. Jordan
Department of EECS
University of California, Berkeley
jordan@cs.berkeley.edu

Beyond the Quadratic Approximation: The Multiscale Structure of Neural Network Loss Landscapes

Chao Ma^{*1}, Daniel Kunin^{†2}, Lei Wu^{‡3}, and Lexing Ying^{§1}

¹Stanford University
Stanford, CA 94305, USA

Cross-Class Structure Learning Spectra

PAPYAN@STANFORD.EDU

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm
- Hessian Spectrum
- Outliers
- Sharpening/EoS
- Simplicity Bias

DEEP LEARNING GENERALIZES BECAUSE THE PARAMETER-FUNCTION MAP IS BIASED TOWARDS SIMPLE FUNCTIONS

Guillermo Valle Pérez
University of Oxford
guillermo.valle@dtc.ox.ac.uk

Ard A. Louis
University of Oxford
ard.louis@physics.ox.ac.uk

Chico Q. Camargo
University of Oxford

SELF-STABILIZATION: THE IMPLICIT SINGULAR

ENT DESCENT AT THE EDGE OF STABILITY

Towards Explaining the Regularization Effect of Singular Value Spectrum Normalization in Deep Learning
Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhi1@andrew.cmu.edu

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhi1@andrew.cmu.edu

Ping Luo
New York University
pingluo@nyu.edu

The Anisotropic Noise in Stochastic Gradient Descent from Sharp Minima and Regularization

The Pitfalls of Simplicity Bias in Neural Networks

Harshay Shah
Microsoft Research
harshay.rshah@gmail.com

Kaustav Tamuly
Microsoft Research
ktamuly2@gmail.com

Aditi Raghunathan
Stanford University
aditir@stanford.edu

Prateek Jain
Microsoft Research
prajain@microsoft.com

Praneeth Netrapalli
Microsoft Research
praneeth@microsoft.com

Level-Set Hessian Spectrum Help Optimization?

Spectrum of Deepnet Hessians
D Trajectory: Progress in Deep Learning:
Edge of Stability
Tsipras* MIT
Andrew Ilyas* MIT
Aleksander Madry MIT
tsipras@mit.edu ailyas@mit.edu madry@mit.edu

Boaz Barak Microsoft Research & MIT
Benjamin L. Edelman MIT
Surbhi Goel Microsoft Research & MIT
barak@csail.mit.edu ben@csail.mit.edu surbhi@microsoft.com
19012526@gmail.com

SGD on Neural Networks Learns Functions of Increasing Complexity

Preetum Nakkiran
Harvard University

Gal Kaplan
Harvard University

Dimitris Kalimeris
Harvard University

Tristan Yang
Harvard University

Benjamin L. Edelman
Harvard University

Fred Zhang
Harvard University

Boaz Barak
Harvard University

Generalization Benefit of Sparsity Reduction

Michael I. Jordan
Department of EECS
University of California, Berkeley
jordan@cs.berkeley.edu

Scale-Invariant Cross-Class Structure Pervade Deep Learning Spectra

PAPYAN@STANFORD.EDU

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm
- Hessian Spectrum
- Outliers
- Sharpening/EoS
- Simplicity Bias
- Adaptive Methods

DEEP LEARNING GENERALIZES
PARAMETER-FUNCTION MAP IS
SIMPLE FUNCTIONS

The Shinghot Mechanism: An Empirical Analysis of Adaptive Optimizers and the Grokking Phenomenon
Guillermo Valle Pérez
University of Oxford
guillevalle@robots.ox.ac.uk

Toward Understanding Why Adam Converges Faster Than SGD for Transformers

Yan Pan
Yuanzhi Li
Carnegie Mellon University

YPAN2@ANDREW.CMU.EDU
YUANZHIL@ANDREW.CMU.EDU

Ard A. Loui
Guaid Saremi
Roni Paiss
Joshua Susskind
Boaz Barak
Benjamin L. Edelman
Surbhi Goel

Why are Adaptive Methods Good for Attention Models?

Jingzhao Zhang
MIT
jzhzhang@mit.edu

Sai Praneeth Karimireddy
EPFL
sai.karimireddy@epfl.ch

Andreas Veit
Google Research
aveit@google.com

Seungyeon Kim
Google Research
seungyeon@google.com

Sashank Reddi
Google Research
sashank@google.com

Sanjiv Kumar
Google Research
sanjivk@google.com

Suvrit Sra
MIT
suvrit@mit.edu

ALUPTIC PROGRESSION IN DEEP LEARNING:
ARTIFACT AND BEYOND
Jian Li
Tsinghua University
ordge@gmail.com

SGD on Neural Networks Learns Functions of Increasing Complexity
Department of Mathematics, New York University
is.nyu.edu

HOW DOES LEARNING RATE DECAY HELP MODERN TRAINING OF NEURAL NETWORKS?
Zhang, G. & Bengio, S. (2016).
Precutur Nakkiran, Gal Kaplan, Dimitris Kalimeris, Tristan Yang
Harvard University, Harvard University, Harvard University, Harvard University
Kathao You, Mingsheng Long (✉)
School of Software, Tsinghua University, School of Software, Tsinghua University
Benjamin L. Edelman, Fred Zhang, Boaz Barak
Harvard University, Harvard University, Harvard University

The Pitfalls of Simplicity
Zhanxin

Beyond the Quality of N
Harshay Shah
Microsoft Research
harshay_sah@redmond.com

Chaopraek Jai
Daniel Prajain
Microsoft Research
prajain@microsoft.com

NOISE IS NOT THE MAIN FACTOR BEHIND THE GAP BETWEEN SGD AND ADAM ON TRANSFORMERS, BUT SIGN DESCENT MIGHT BE

Frederik Kunstner, Jacques Chen, J. Wilder Lavington & Mark Schmidt†
University of British Columbia, Canada CIFAR AI Chair (Amii)†
praneeth@microsoft.com, fkd@ubc.ca 94305, USA

Partially Understood Phenomena Abound in NN Optimization

An incomplete list:

- Grokking
- Benefits of Large LR
- Batchnorm
- Hessian Spectrum
- Outliers
- Sharpening/EoS
- Simplicity Bias
- Adaptive Methods
- Unstable Training

DEEP LEARNING GENERALIZES BECAUSE OF THE
PARAMETER-FUNCTION MAP IS BIASED TOWARDS
SIMPLE FUNCTIONS
Toward Understanding Why Adam Converges Faster Than SGD for
The Slingshot Mechanism: An Empirical Study of Transformer Optimization Help
Adaptive Optimizers and the Grokking of Outliers in the Spectrum of Deepnet Hessians
THE BREAKTHROUGH: Progress Along GD Trajectory: THEORETICAL ANALYSIS OF THE COMPUTATIONAL LIMIT
Guillermo Valle Pérez
University of Oxford
guilleval@ox.ac.uk
Chito O. Camargo
Carnegie Mellon University
camargo@cmu.edu
Yipeng Pan
Carnegie Mellon University
yipeng@cmu.edu
Chitranjan Chakrabarti
Carnegie Mellon University
chitran@cmu.edu
Srispr
MIT
srispr@mit.edu
Andrew Li
MIT
aliwas@mit.edu
Madry
MIT
madr@mit.edu

An Empirical Study of Training Self-Supervised Vision Transformers

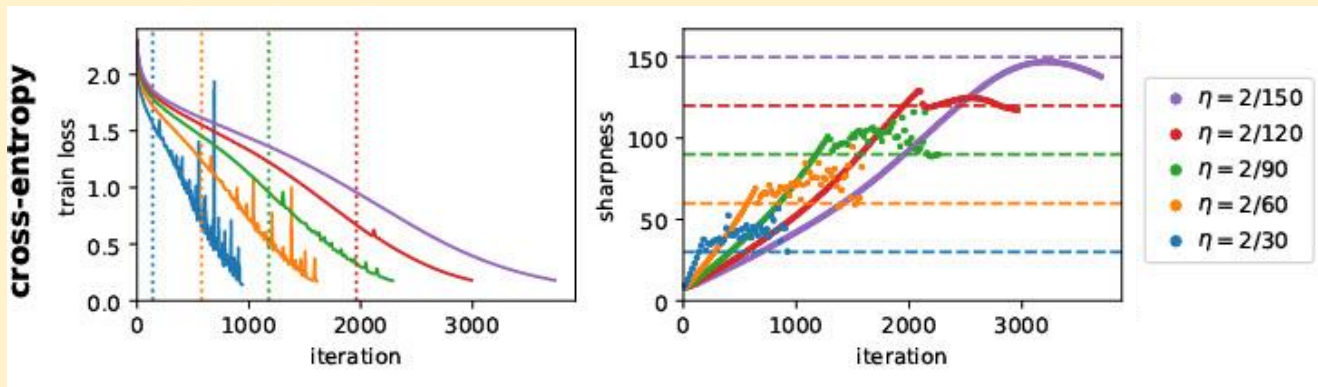
Xinlei Chen* Saining Xie* Kaiming He
Facebook AI Research (FAIR)

How to Fine-Tune Vision Models with SGD

Ananya Kumar Ruoqi Shen Sébastien Bubeck Suriya Gunasekar

Beyond the Question: Descent Might Be the Most Efficient Cross-Class Structure
The Role of New Invariant Landscapes
Harshay Shah
Microsoft Research
harshay_sah@msr.com
Rishi Hanjra
Stanford University
hanjra@stanford.edu
Saurabh Kadavath
Stanford University
kadavath@stanford.edu
Pradeep Laila
Microsoft Research
prajain@microsoft.com
Daniel Kurup
University of British Columbia, Canada
praneeth@cs.ubc.ca
Frederik Kunzner
University of British Columbia, Canada
kunzner@cs.ubc.ca
Jacques Chen
University of British Columbia, Canada
jacques@cs.ubc.ca
J. Wilder Lavington
University of British Columbia, Canada
lavington@cs.ubc.ca
Mark Schmidt
University of British Columbia, Canada
schmidt@cs.ubc.ca
Pavani
Stanford University
pavani@stanford.edu

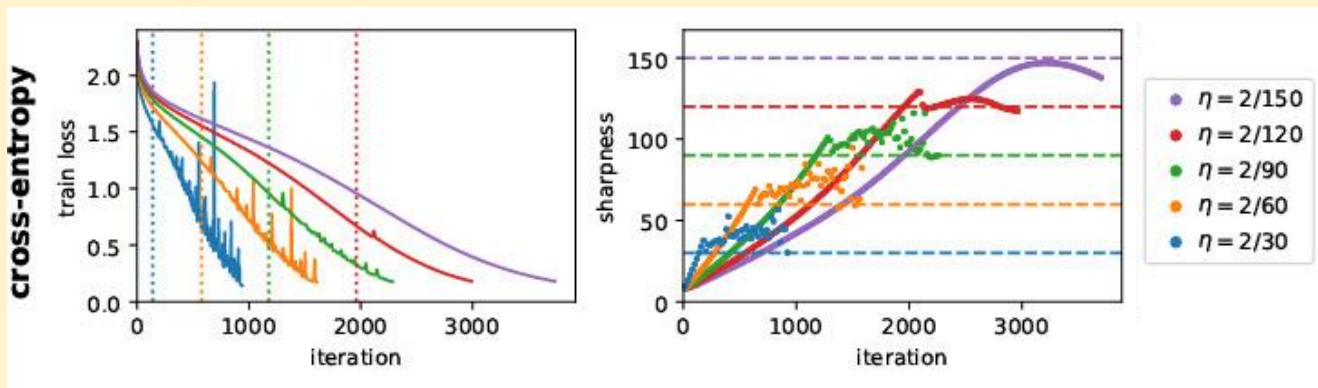
Progressive Sharpening + Edge of Stability



Meanwhile, loss decreases non-monotonically, with frequent “spikes”.

“Sharpness” = top eigenvalue of loss Hessian
First rises to $2/\eta$...
Then hovers around that value.

Progressive Sharpening + Edge of Stability



This is just more evidence that **something more** is needed to understand NN training dynamics...

Yet Another Phenomenon

I'm going to present our finding:
another interesting phenomenon in neural network optimization.

But the goal is not just to add to the growing list.

Instead, we hope it can help explain and unify
these observations via a **shared underlying cause**.*

Yet Another Phenomenon

(We also look at SGD)

Let's run the following experiment:

1. Train a neural network with **full-batch gradient descent** on CIFAR-10.
2. Track losses on each training point *individually*.
3. Fix some iteration T .
4. Calculate changes in loss on each point from step T to step $T+1$.
5. Visualize the samples with the *most positive* and *most negative* changes.

What should we expect to see?

Yet Another Phenomenon

VGG-11

ResNet-18

Largest increase in loss



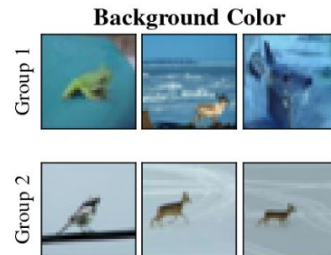
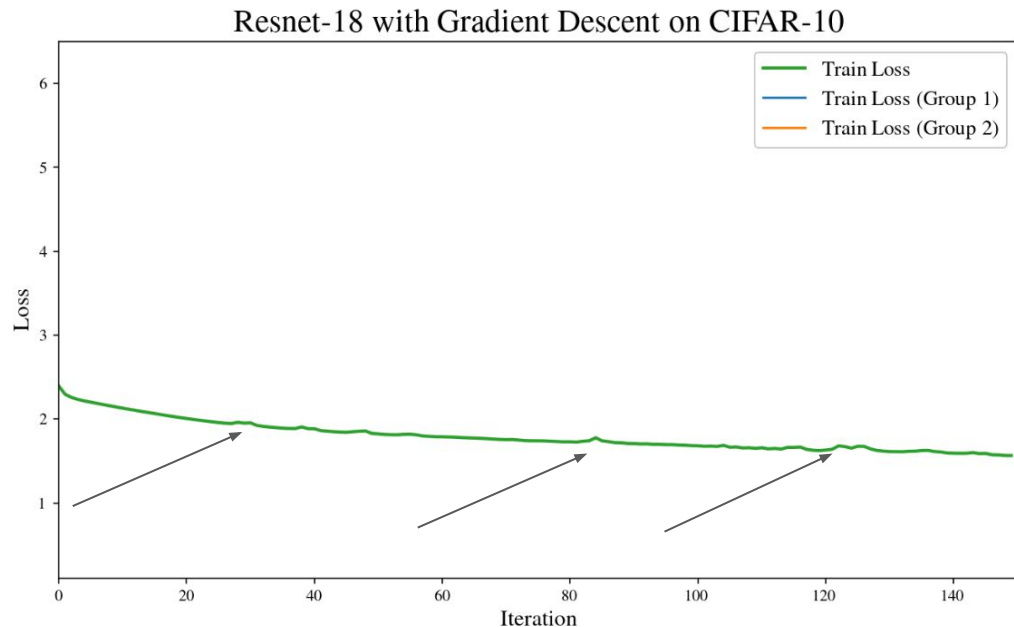
Largest decrease in loss



The precise patterns change, but this occurs *all throughout training*.

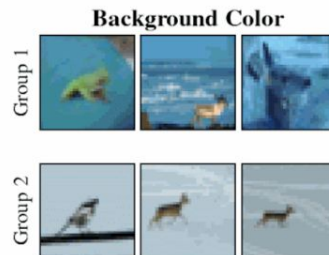
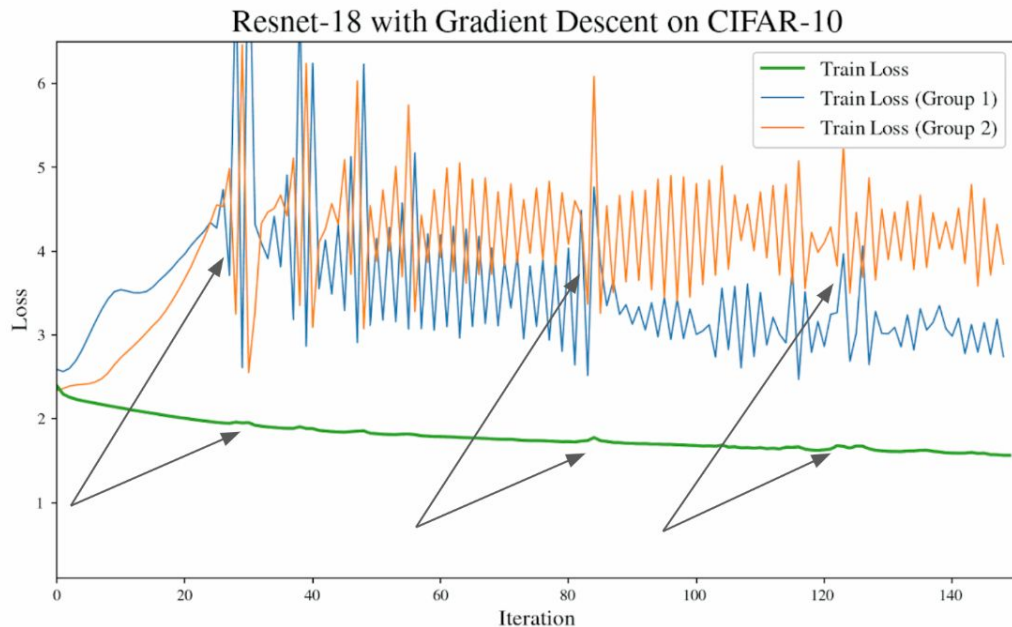
Visualizing the Group Losses

These groups are ~20 samples each.



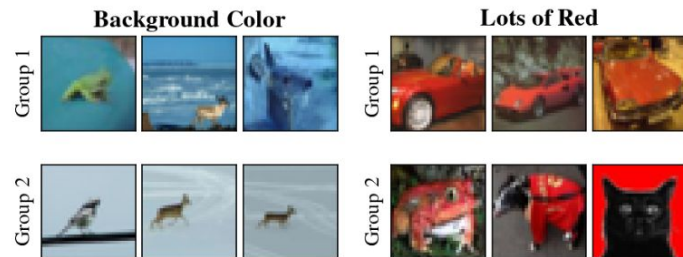
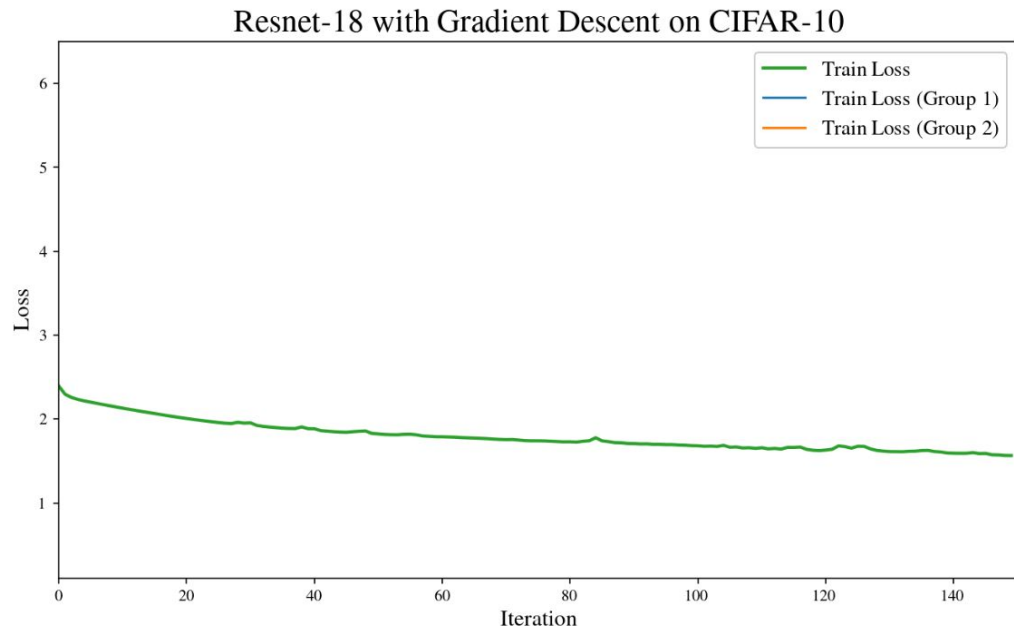
Samples were selected for largest change in loss, so we expect a “spike” somewhere.

Visualizing the Group Losses



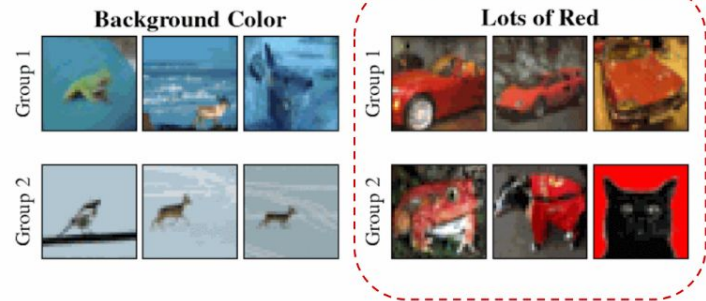
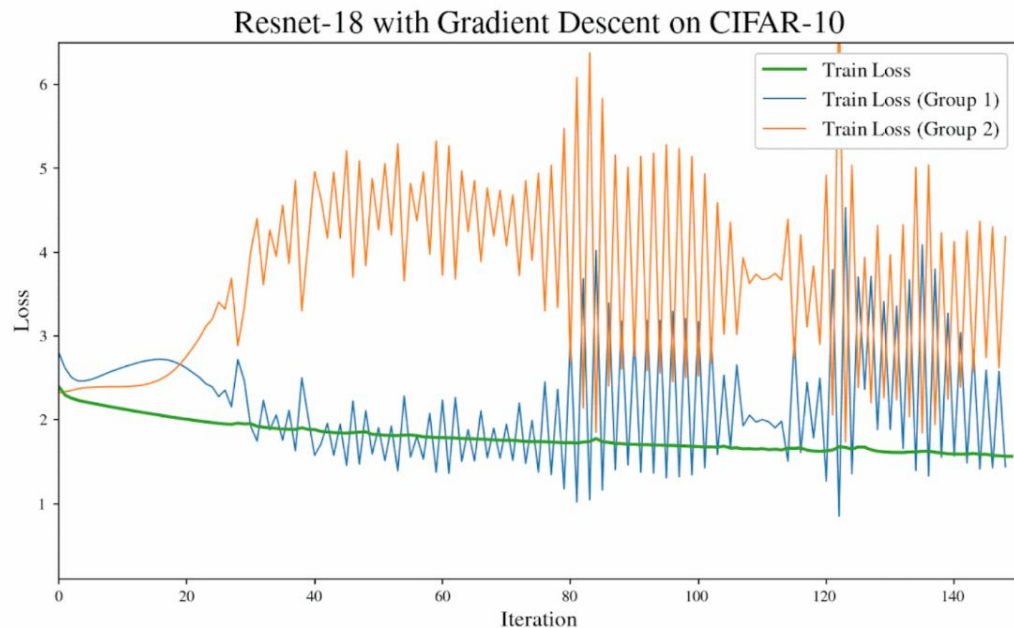
These opposing groups oscillate with large amplitude *continuously!*

Visualizing the Group Losses



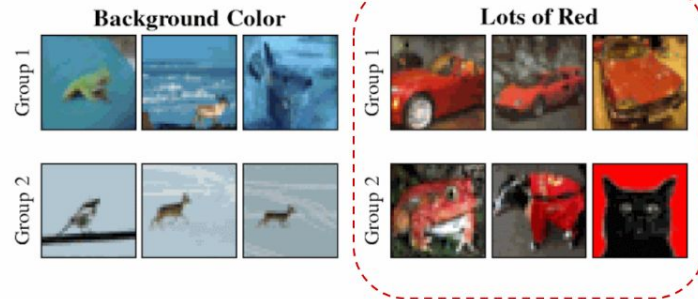
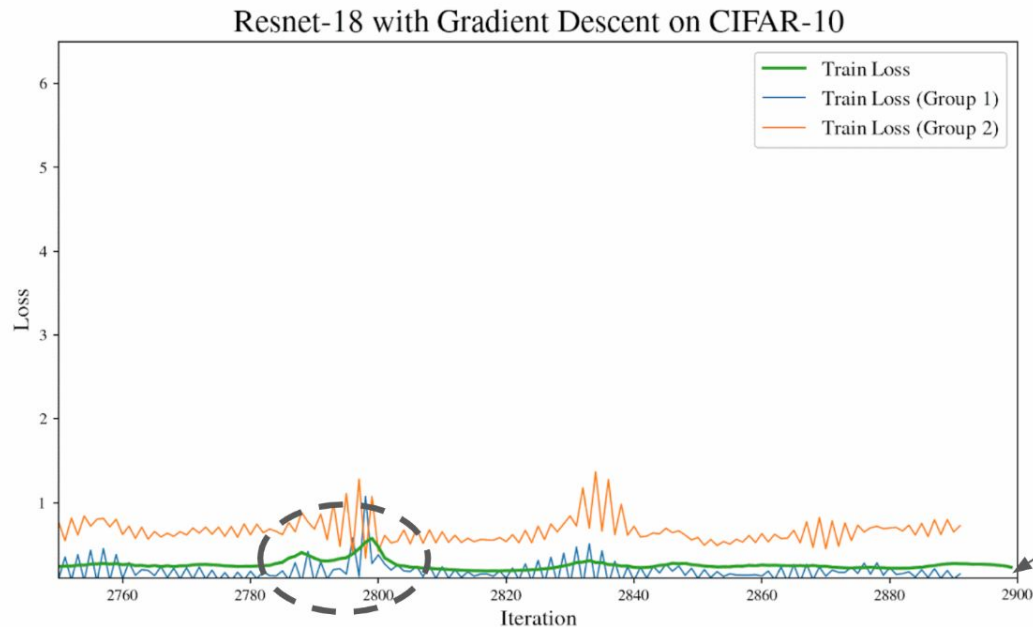
What about another group?

Visualizing the Group Losses



When we're close to interpolating, shouldn't this effect be reduced?

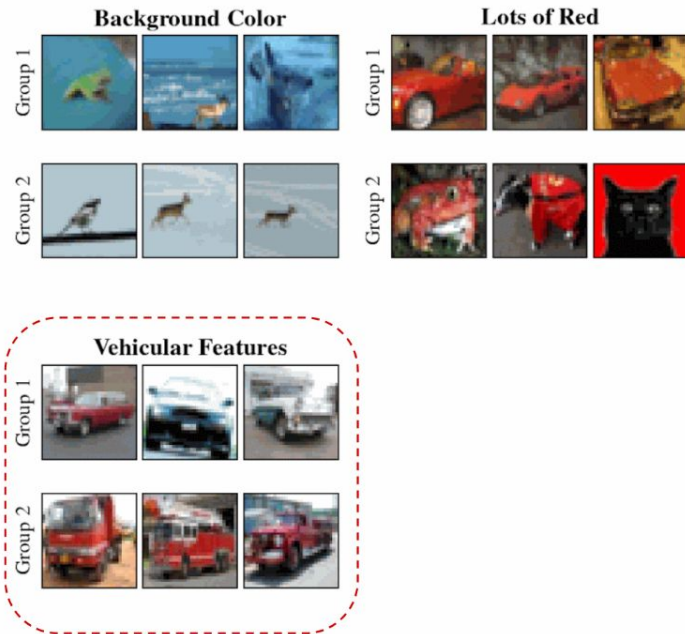
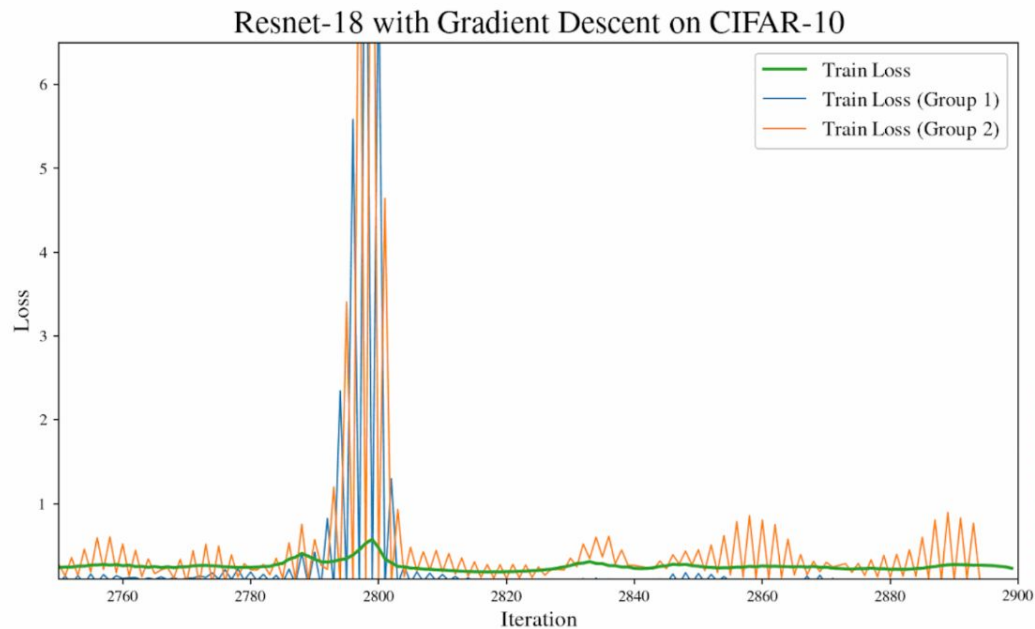
Visualizing the Group Losses



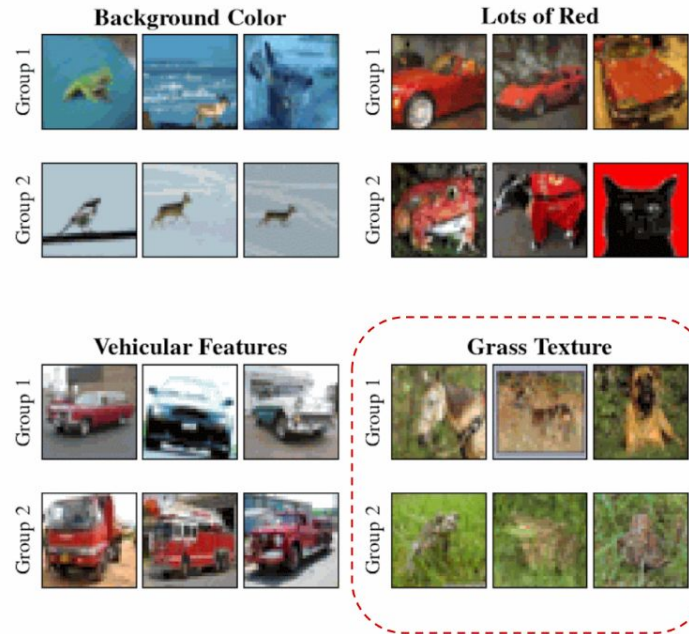
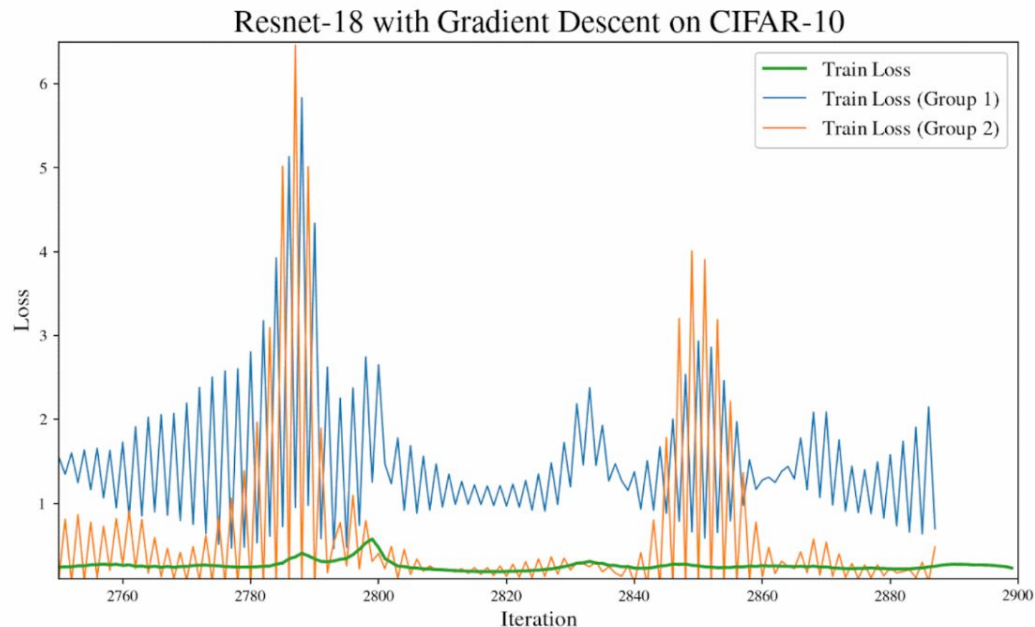
(Note the change in x-axis)

Yes, the amplitude is substantially smaller...
So what's causing these loss increases?

Visualizing the Group Losses



Visualizing the Group Losses



Even at the end stages of training, large loss swings are still occurring.

What's Going On?

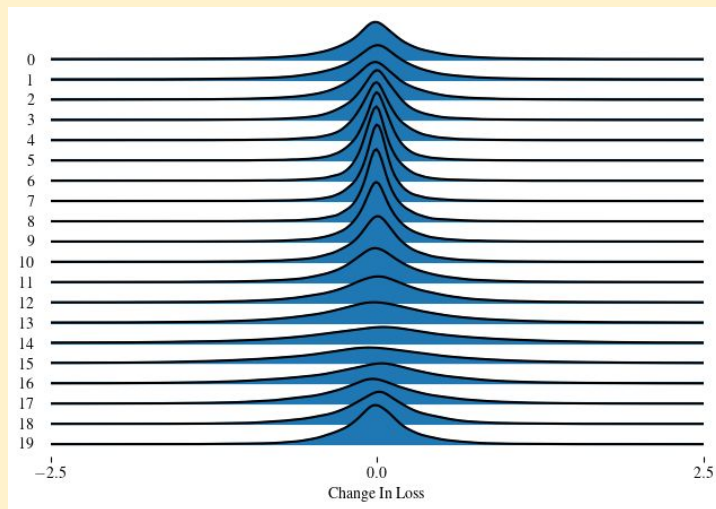
- **Prevalent features**, often with distinct colors.
 - Roughly, “prevalent” \approx “fills a lot of the image”
- **Begin simple**, become progressively more complex.
 - “Simple” \approx “available at random initialization”
- Large gradients pointing in **opposite directions**.
 - Learning “red = car” decreases loss on red cars, increases loss on red *non*-cars

We call these features—or the gradients they induce—*Opposing Signals*.

What's Going On?

Does this occur for *every* training sample?

Distribution of changes in loss:

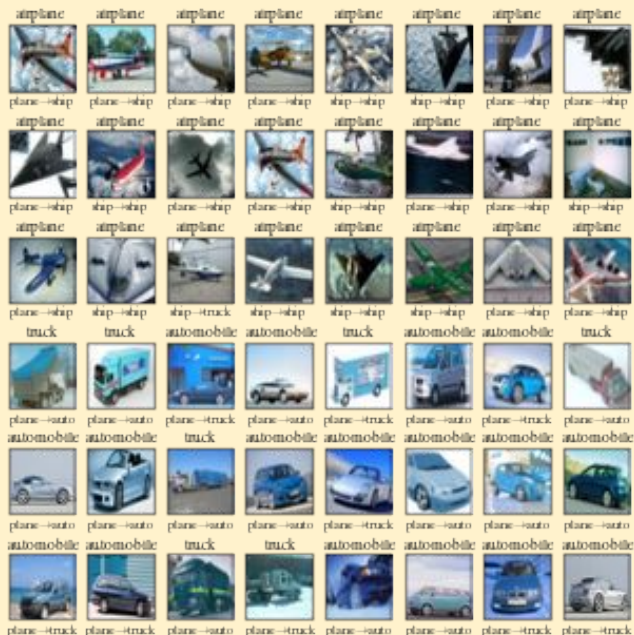


These samples are *significant* outliers.

What Causes Opposing Signals?

Is this a property of architecture (ConvNet)?

No. Same occurs in a Vision Transformer.



What Causes Opposing Signals?

Maybe it's a property of the data modality (images)?

Also no.

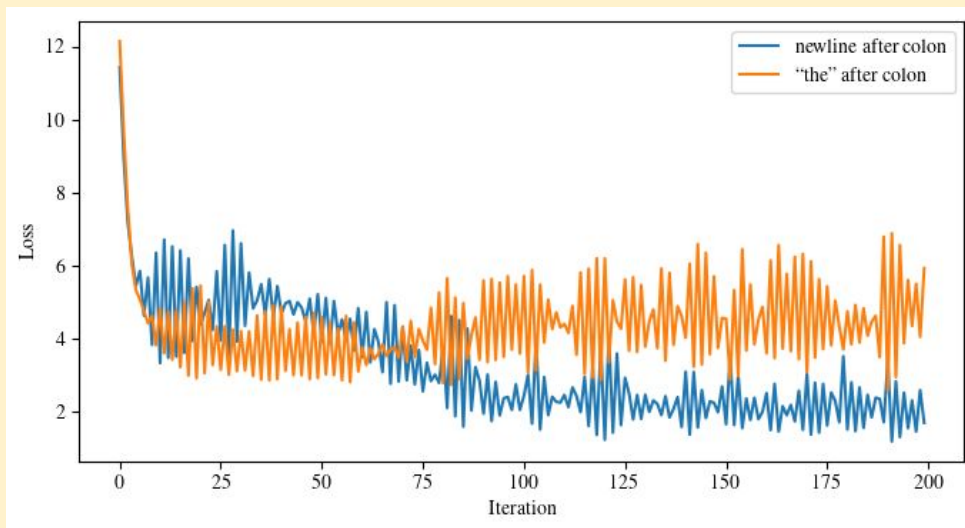
Group 1

```
Salcedo said of the work:[\n]
Enter your email address:[\n]
According to the CBO update:[\n]
Here's how the Giants can still make the
playoffs:[\n]
in early 2018.\n\nAccording to the CBO update:[\n]
other than me being myself." \n\nWATCH:[\n]
```

Group 2

```
MPs in Westminster. But to me it is obvious: [the]
The wheelset is the same as that on the model above:
[the]
all other acts of love, both divine and human: [the]
from the Kurds' two main political parties: [the]
title of precisely what makes it so wonderful: [the]
you no doubt noticed something was missing: [the]
```

(bracket is next token)



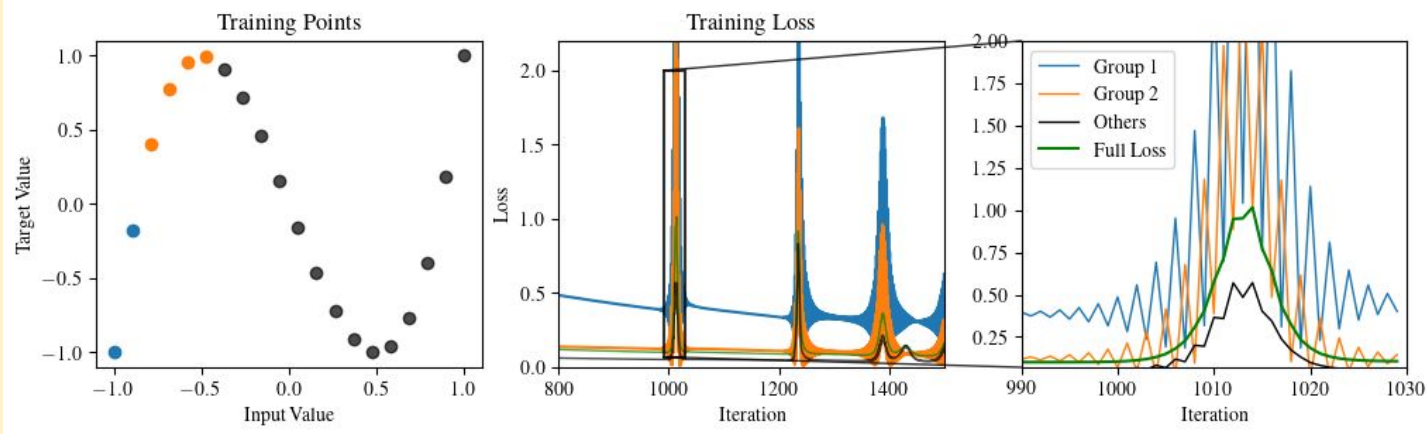
GPT-2 on OpenWebText

What Causes Opposing Signals?

Maybe it's a property of the data modality (images)?

Also no.

What about the loss (cross-entropy)?



What Causes Opposing Signals?

Remainder of this talk gives our current best understanding, with experiments.

We believe it a consequence of *depth* and *steepest descent*.

We don't fully understand the mechanism here.

- If there are parts you think aren't fully explained, you're right.
- If there are parts you think are *flat out wrong*, you could be right.

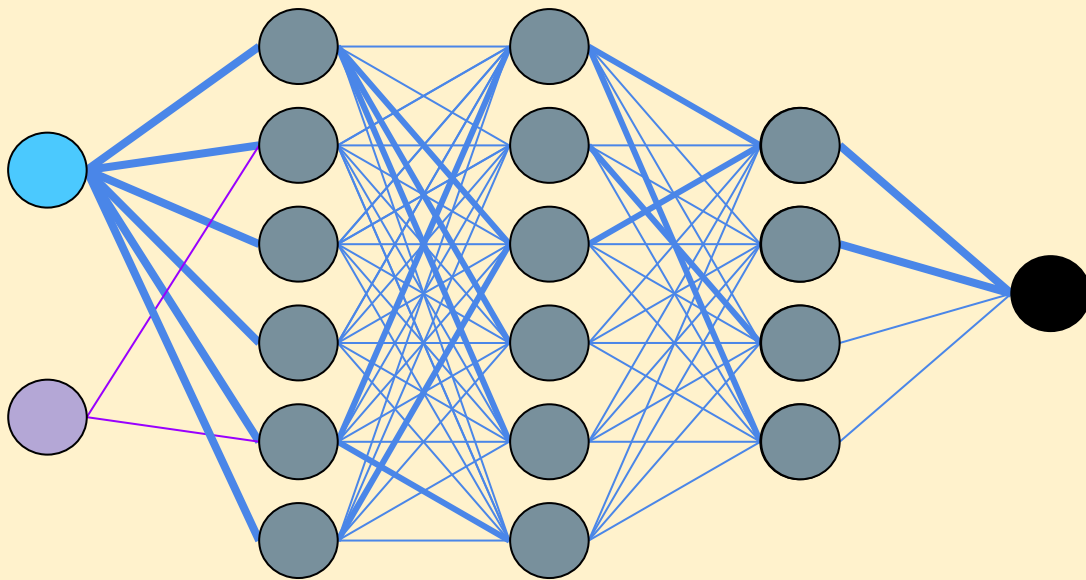
However:

- We have a reasonably descriptive high-level story...
 - and we prove this behavior for a simple model on a 2-layer linear net.*
- It enables *specific* qualitative predictions which we then verify...
 - and it naturally fits into several existing narratives of other phenomena.

A Simplified Story of Gradient Descent on Deep Neural Networks

Consider a randomly initialized MLP with two input features:

1. “Sky”: large magnitude + pervasive (propagated to all neurons).
 - Only sufficient for predicting $p(\text{class} \mid \text{“sky”})$.
2. “Shape”: small magnitude, needs to be learned.
 - But much more useful for loss reduction.

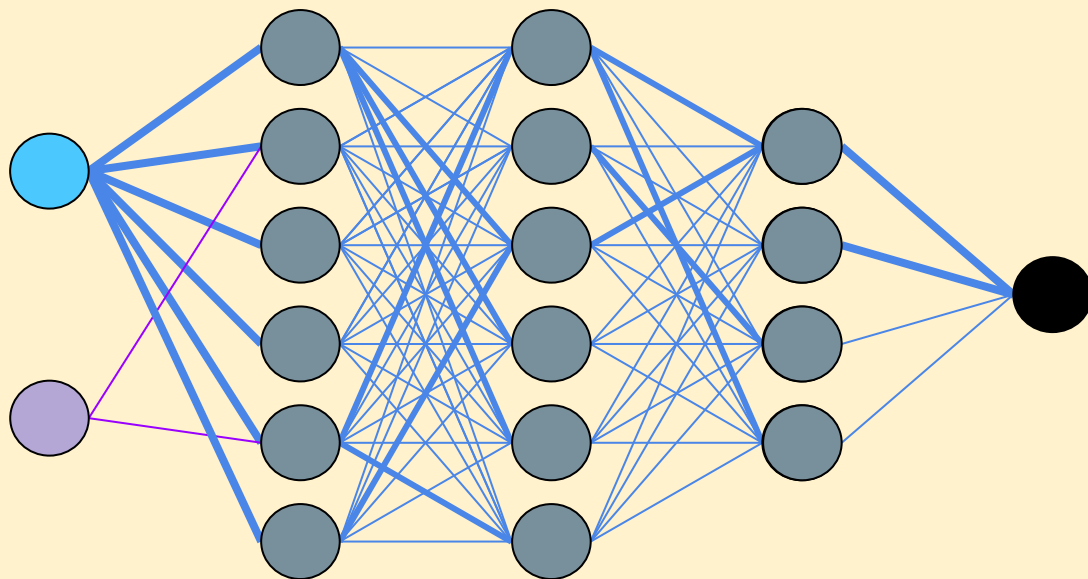
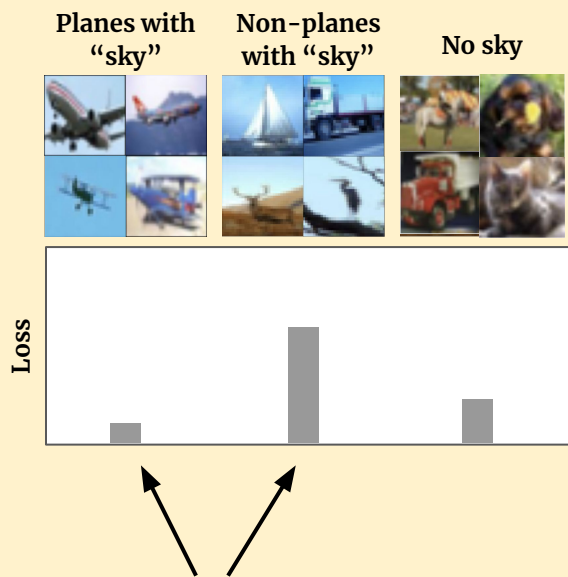


At initialization, network activations are dominated by “sky” on outliers.

- (Suppose network happens to predict “sky = plane”)

High loss \rightarrow large gradients \rightarrow rebalance towards predicting $p(\text{class} \mid \text{“sky”})$.

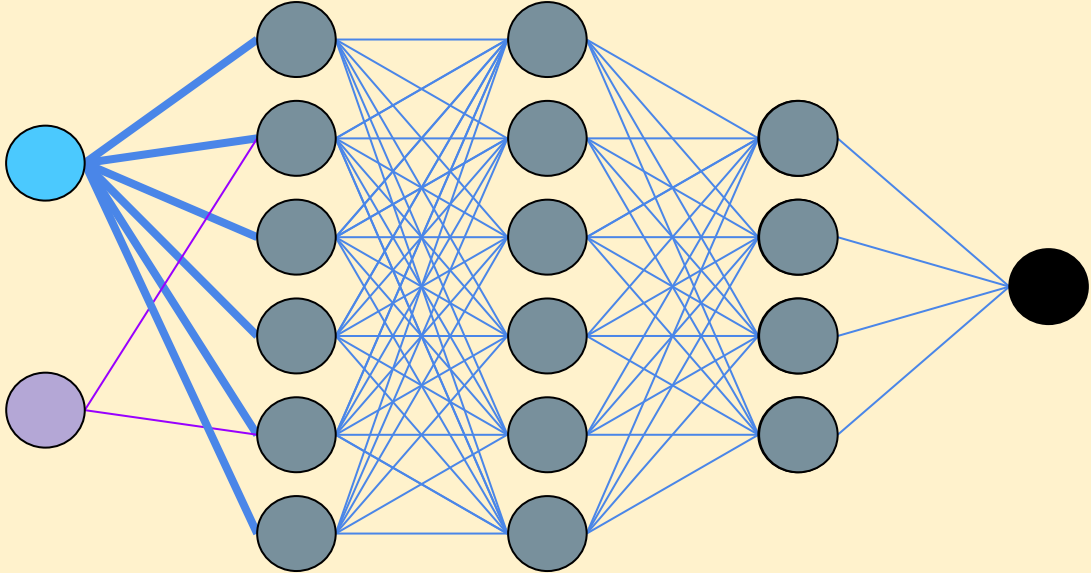
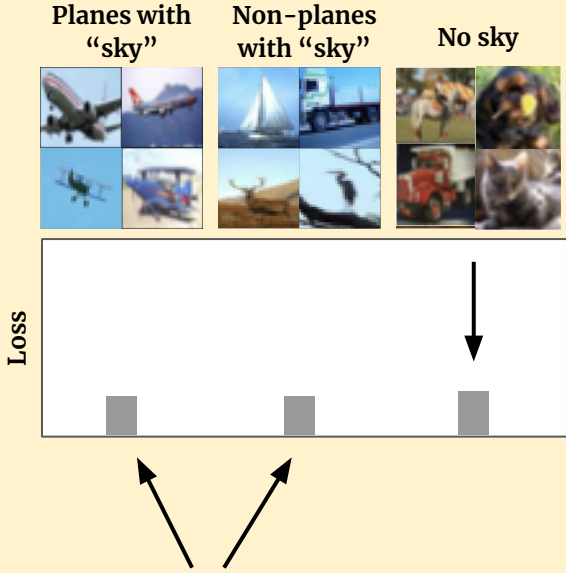
- (This “linear first” behavior has been previously observed^[1, 2])



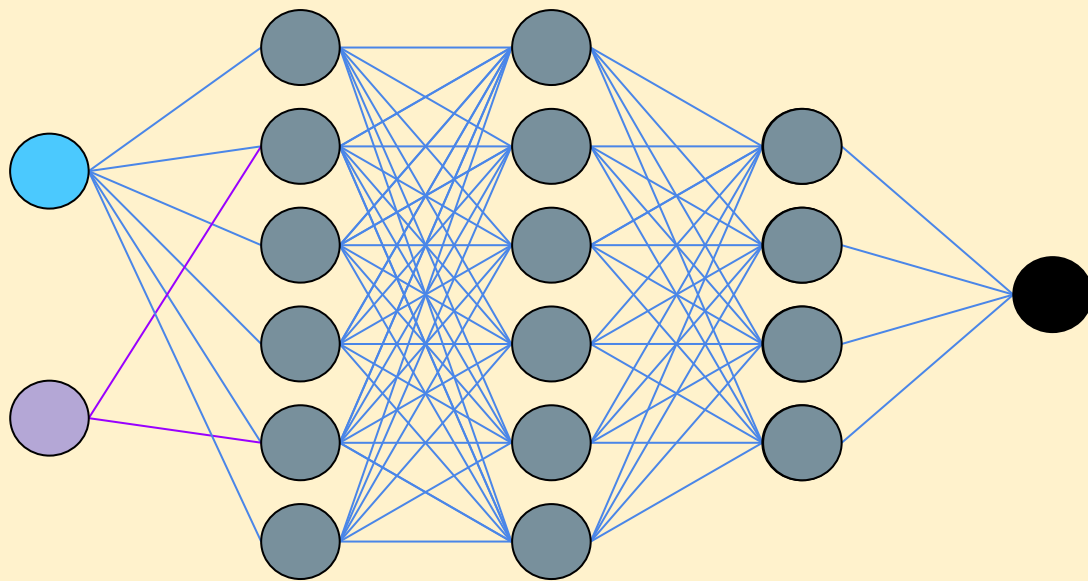
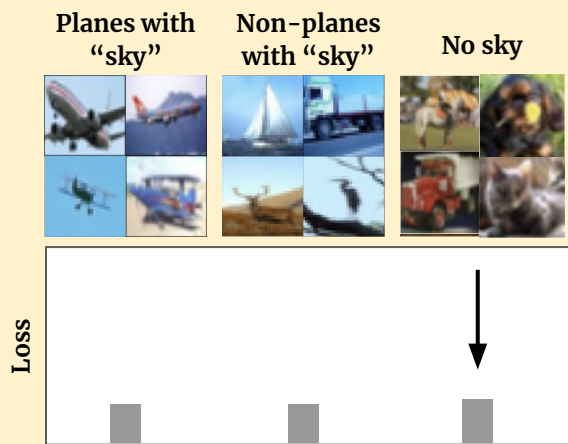
[1] SGD on Neural Networks Learns Functions of Increasing Complexity. Nakkiran et al. 2019.

[2] Do deep neural networks learn shallow learnable examples first? Mangalam and Prabhu 2019.

Once this happens, the network can now upweight the more useful “shape” feature.
Since the outliers’ loss no longer dominates the gradient, let’s visualize a non-outlier.

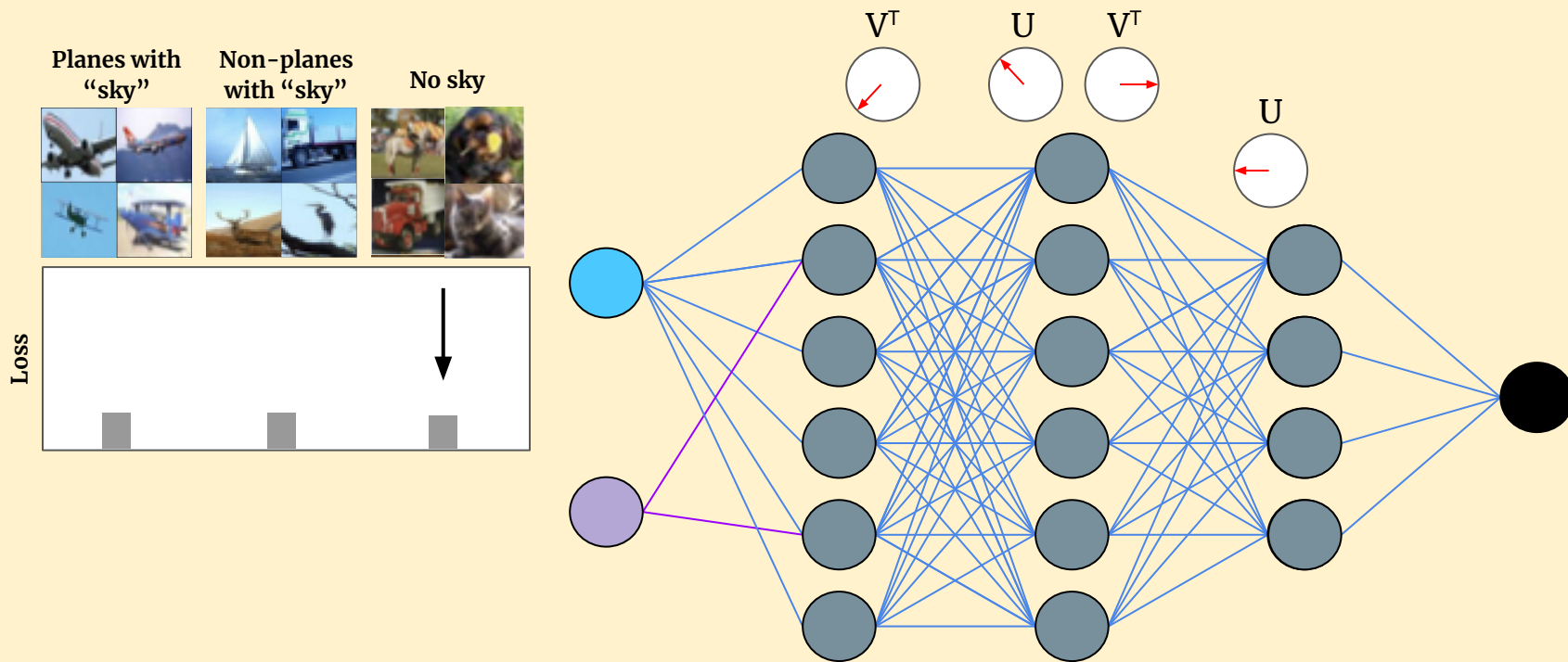


Once this happens, the network can now upweight the more useful “shape” feature.
Since the outliers’ loss no longer dominates the gradient, let’s visualize a non-outlier.



As training progresses, the top singular vectors of adjacent layers align to amplify meaningful subspaces. [3, 4]

This is how the “shape” feature gets *upweighted*.

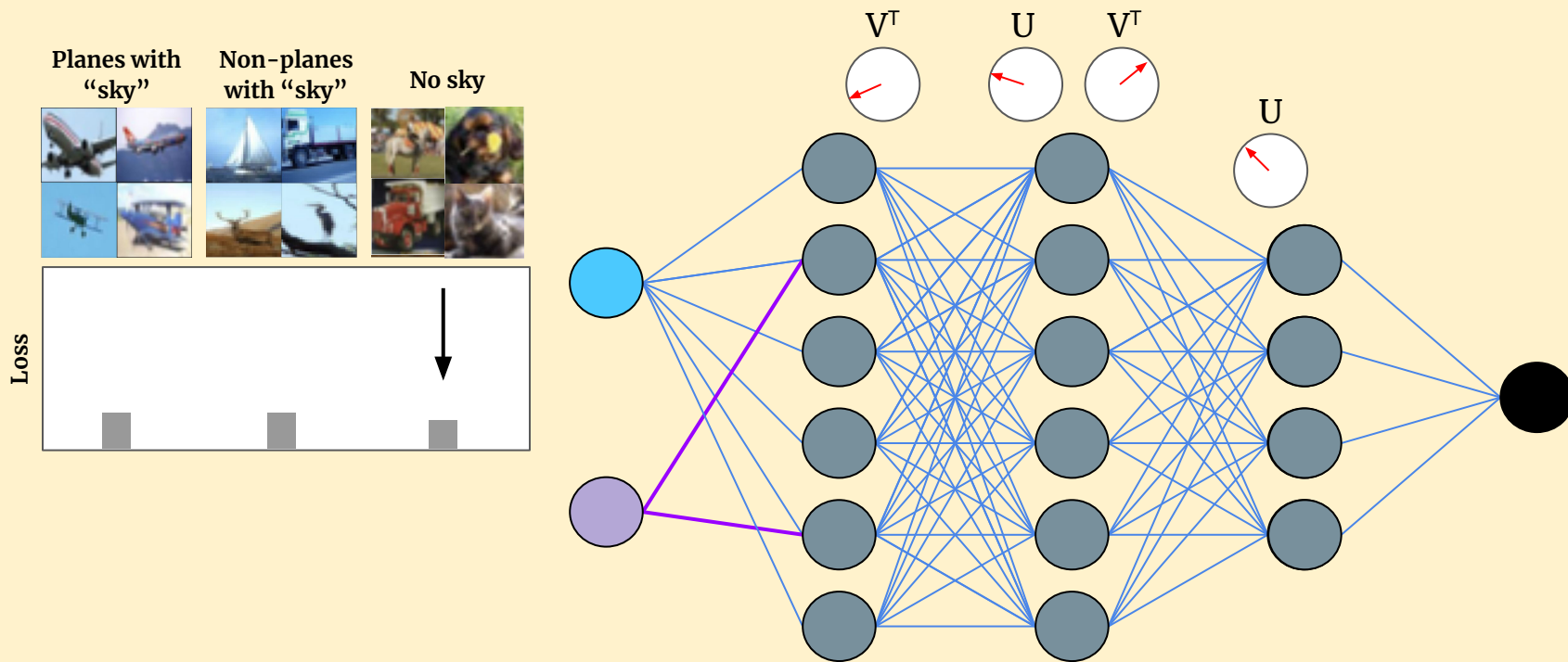


[3] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Saxe et al. 2013

[4] Unique properties of flat minima in deep networks. Muylloff and Michaeli, 2020.

As training progresses, the top singular vectors of adjacent layers align to amplify meaningful subspaces. [3, 4]

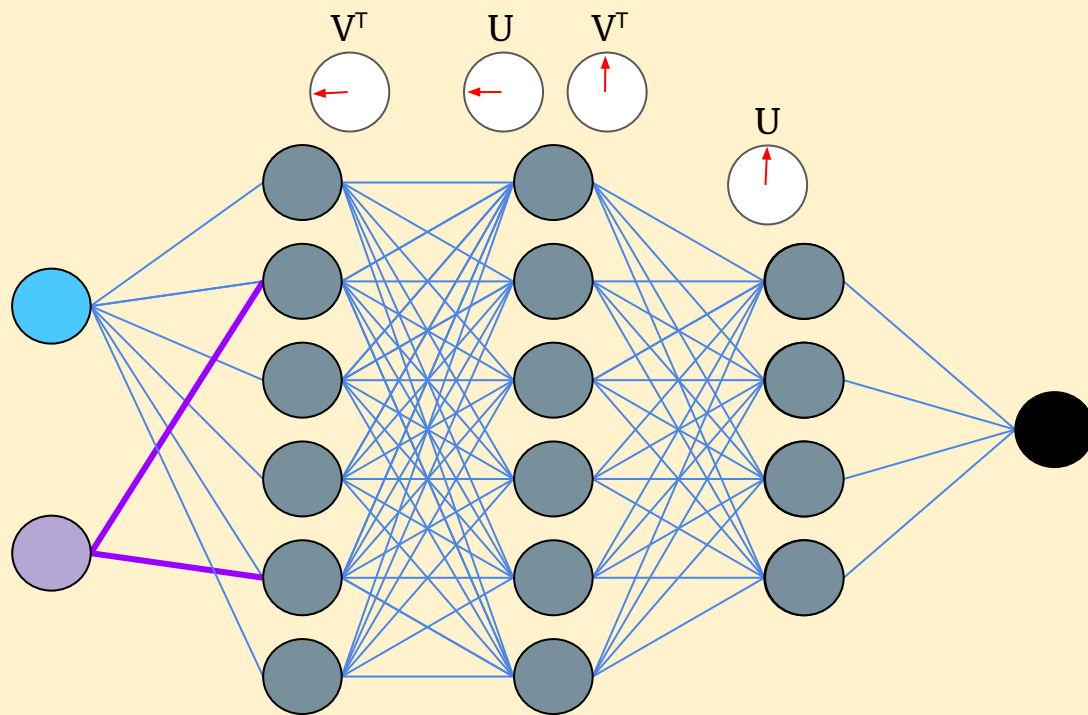
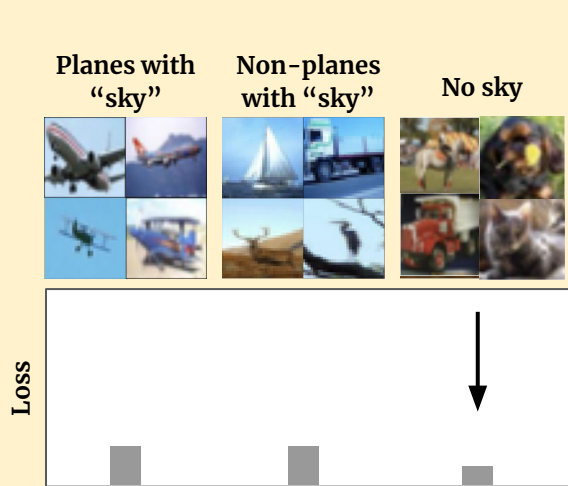
This is how the “shape” feature gets *upweighted*.



[3] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Saxe et al. 2013

[4] Unique properties of flat minima in deep networks. Muylloff and Michaeli, 2020.

This alignment has been continuously upweighting the more useful signal.



[3] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Saxe et al. 2013

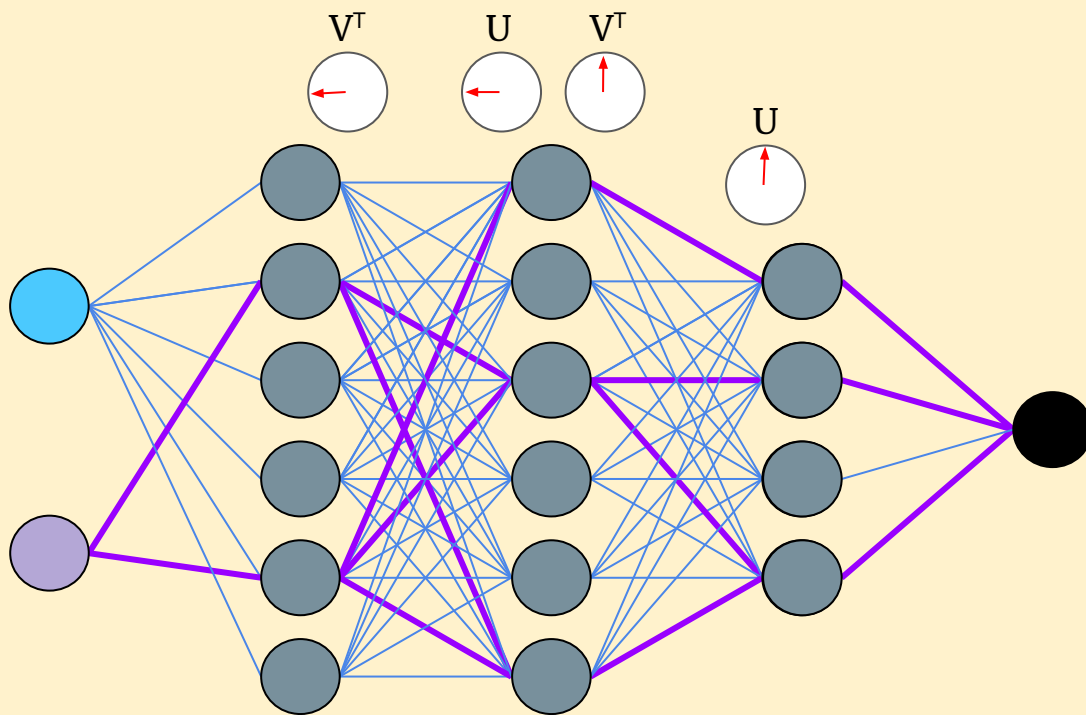
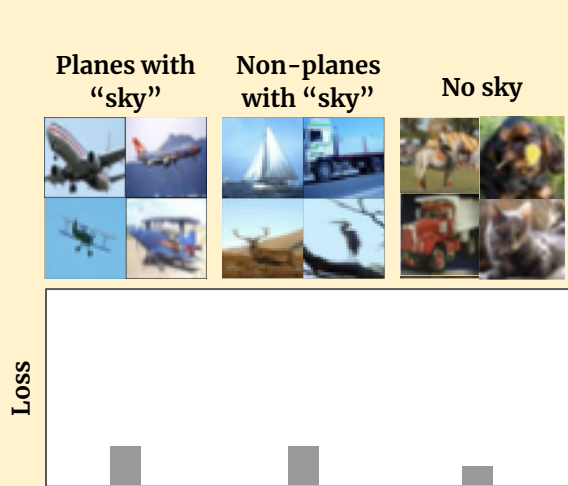
[4] Unique properties of flat minima in deep networks. Muylloff and Michaeli, 2020.

I've left one important part out of this visualization:

When "shape" is amplified, "sky" is amplified too.

This is the activation pattern for a *non-outlier*.

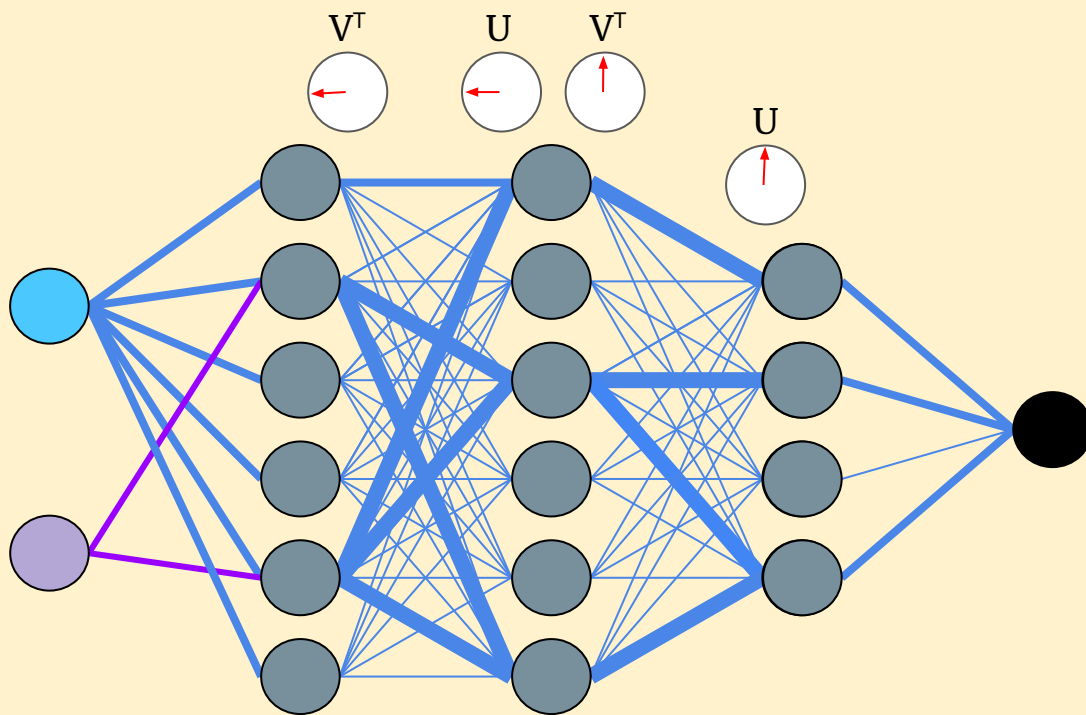
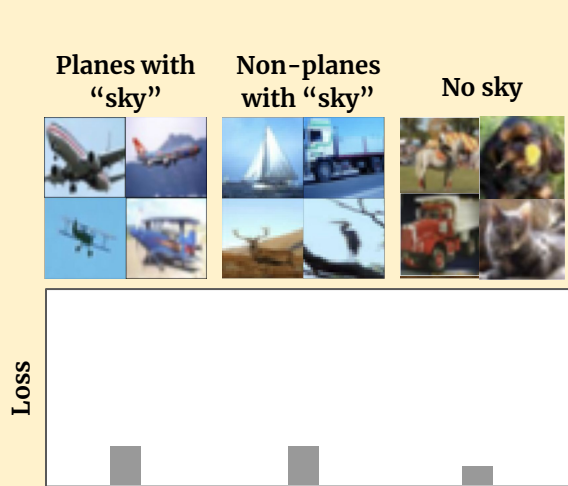
What would it look like for an outlier with a sky background?



I've left one important part out of this visualization:

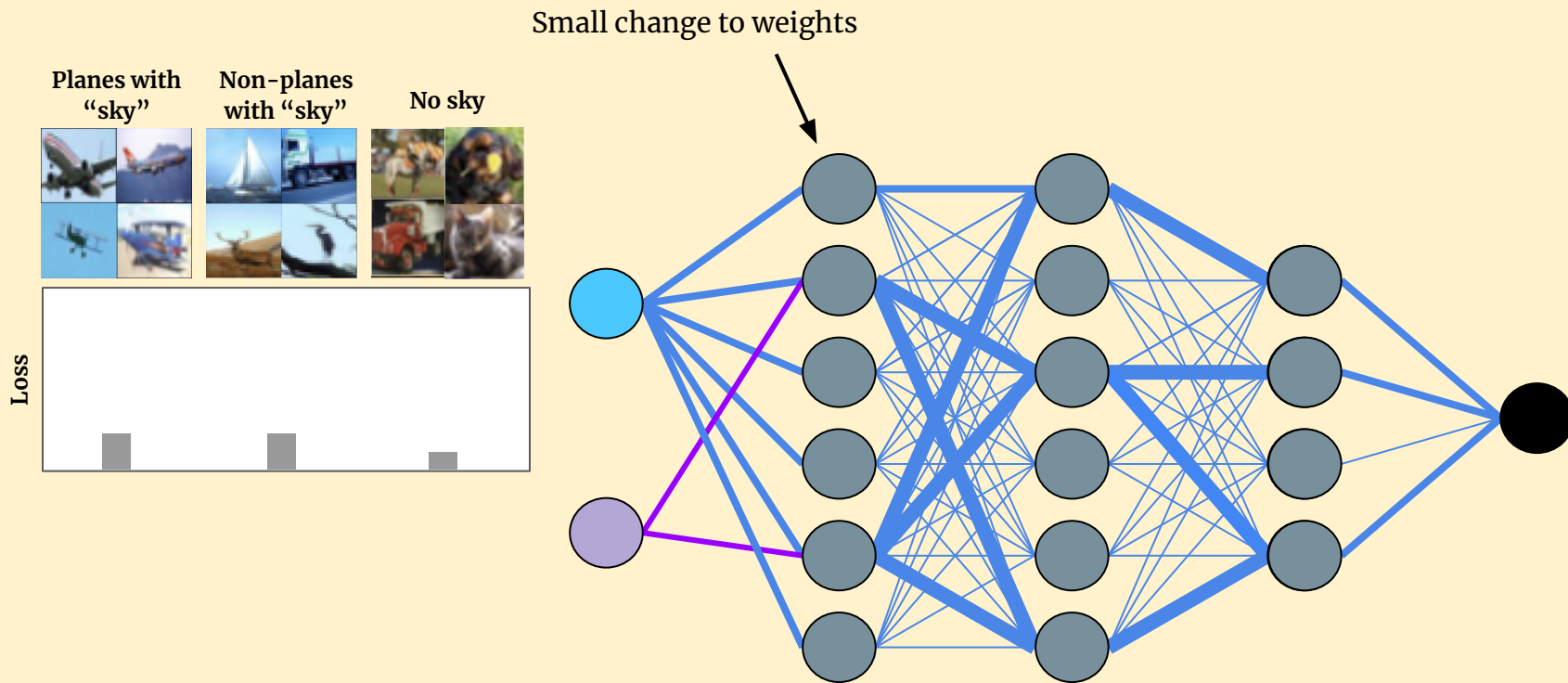
When “shape” is amplified, “sky” is amplified too.

Because it is larger + more pervasive, it still dominates the network's activations.



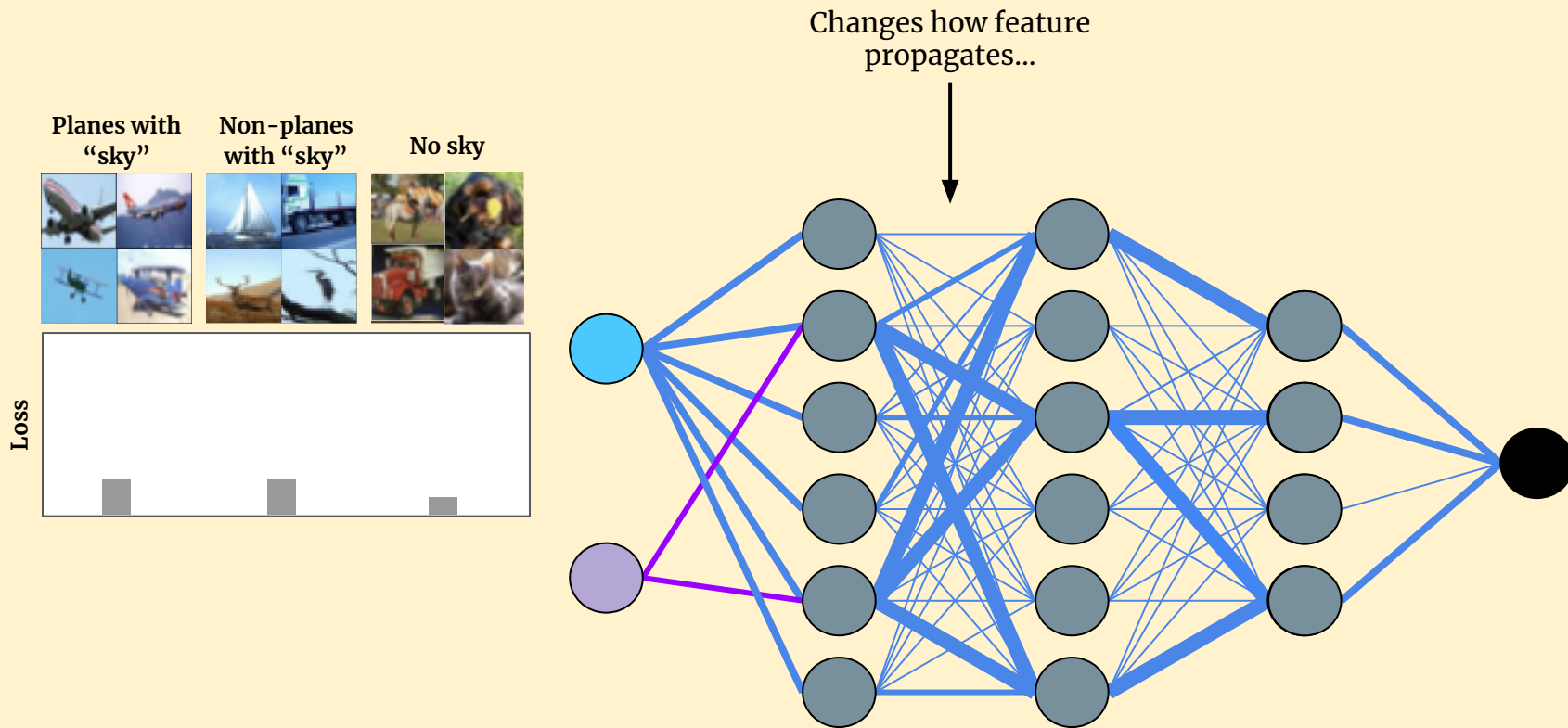
This causes large sensitivity to small changes in *how the network uses* “sky”.

- Small, targeted change to predict one group massively increases loss on the other.



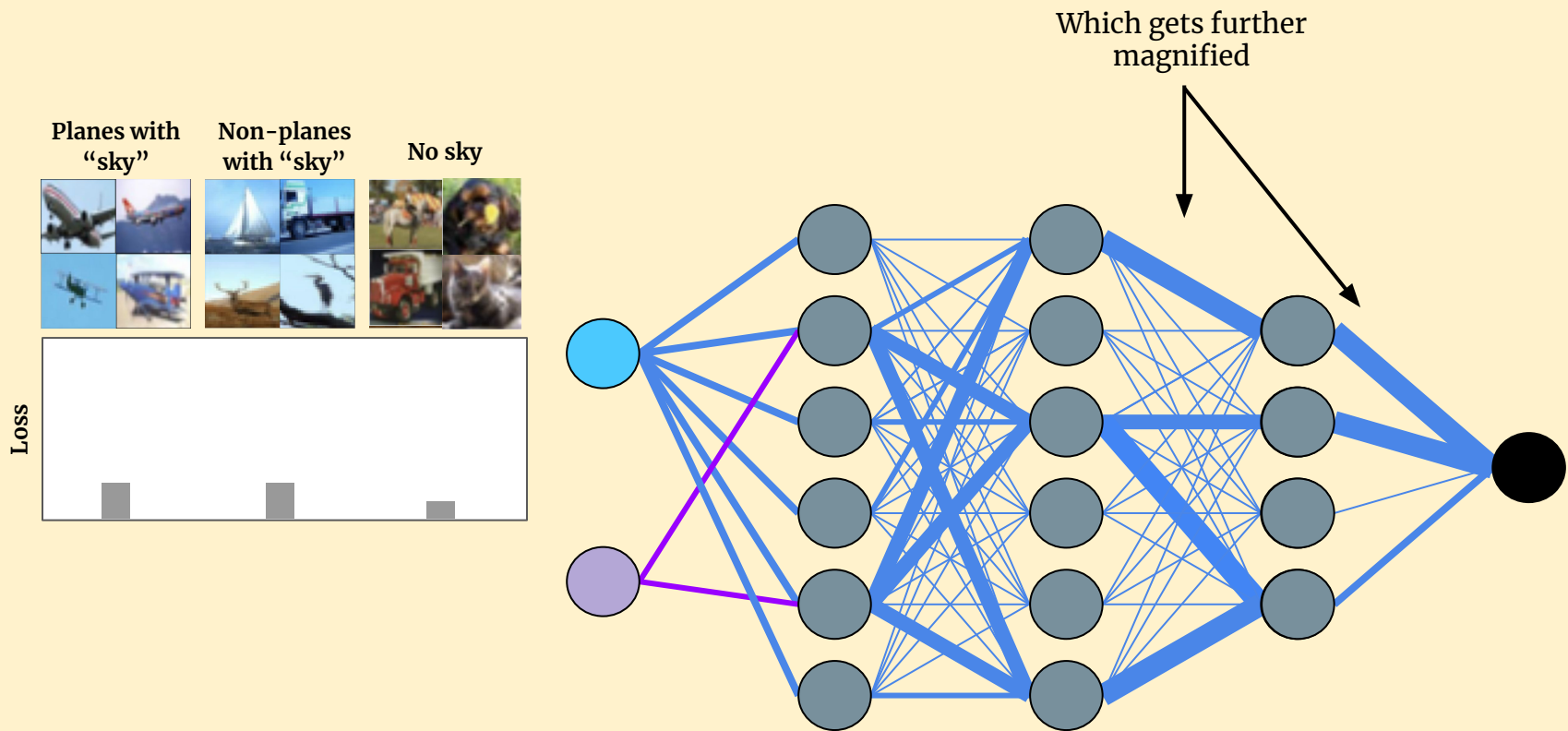
This causes large sensitivity to small changes in *how the network uses* “sky”.

- Small, targeted change to predict one group massively increases loss on the other.



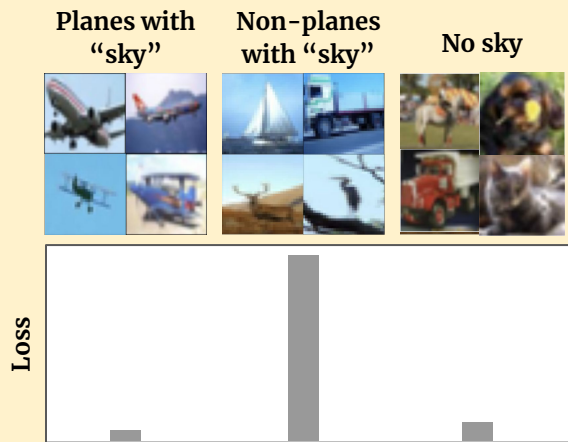
This causes large sensitivity to small changes in *how the network uses* “sky”.

- Small, targeted change to predict one group massively increases loss on the other.

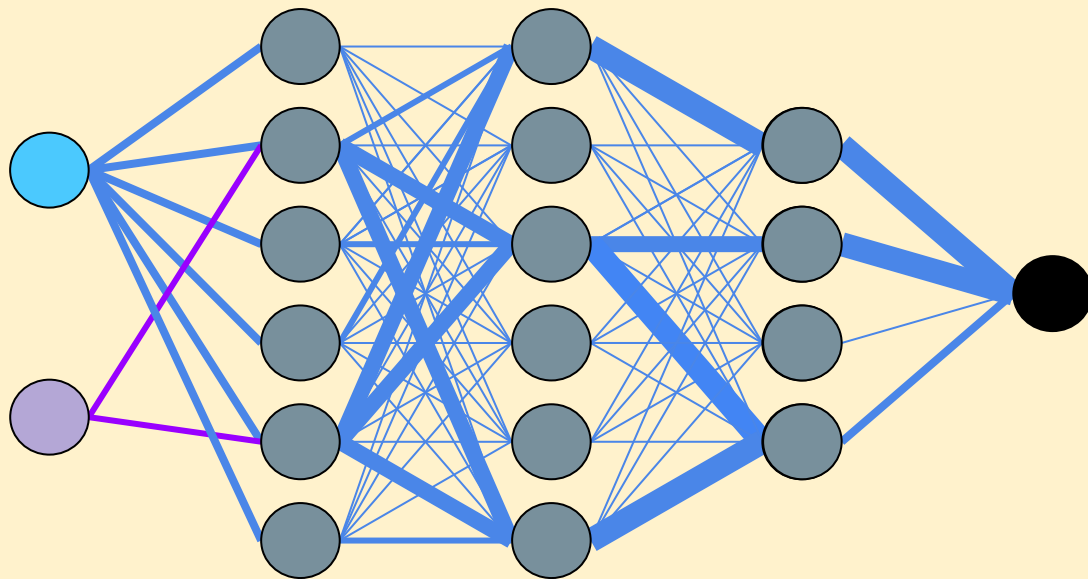


This causes large sensitivity to small changes in *how the network uses* “sky”.

- Small, targeted change to predict one group massively increases loss on the other.



And can have huge effect on loss

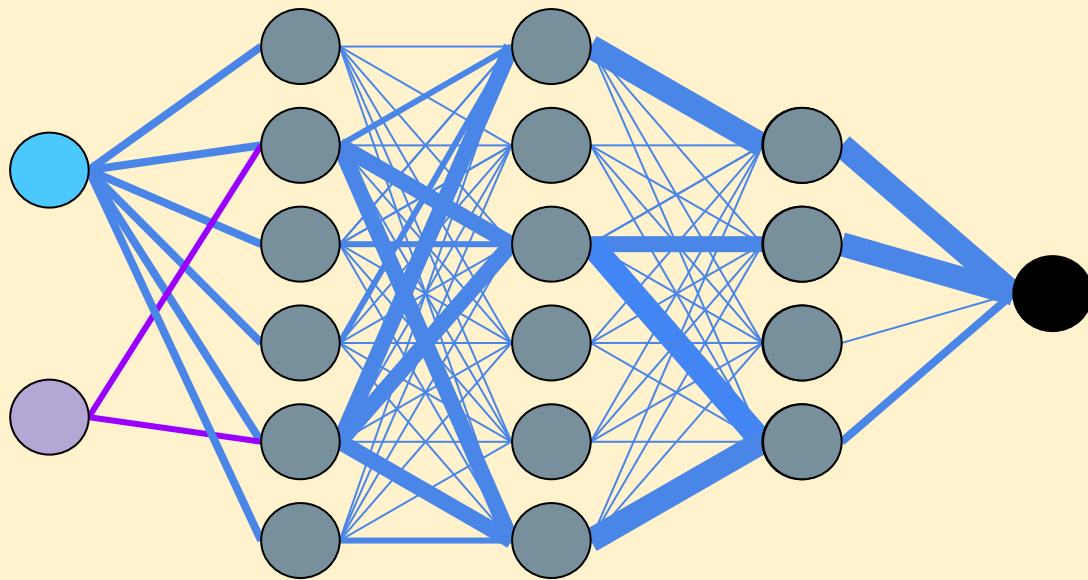
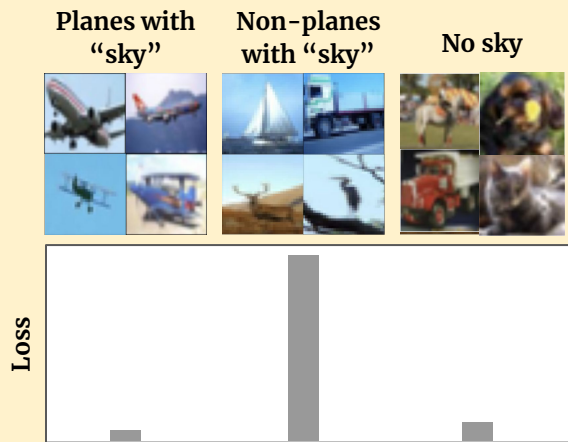


This causes large sensitivity to small changes in *how the network uses* “sky”.

- Small, targeted change to predict one group massively increases loss on the other.

In other words, **loss on outliers becomes very sharp w.r.t. parameters.**

- (“growth in sensitivity” was previously noted, e.g. weight/Jacobian norm^[5, 6])



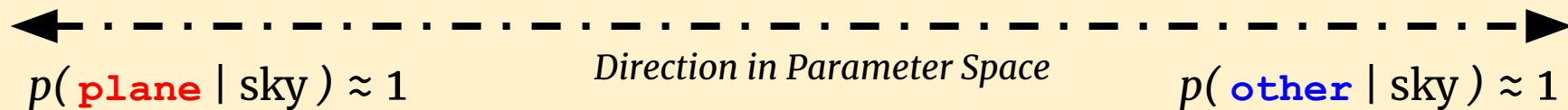
[5] On linear stability of sgd and input-smoothness of neural networks. Ma and Ying, 2021.

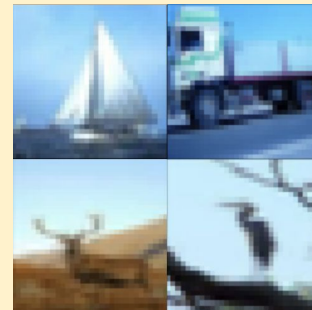
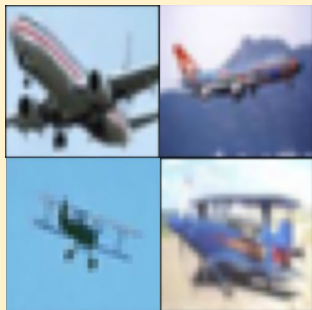
[6] On the lipschitz constant of deep networks and double descent. Gamba et al. 2023.

This story is pretty abstract.

Let's visualize something more concrete:

The (hypothetical) loss in a 1D parameter space.





How does early optimization move along this axis?

Loss on images of **planes** with sky

Loss on images of **non-planes** with sky

Optimization continues
“through the valley”^[1]

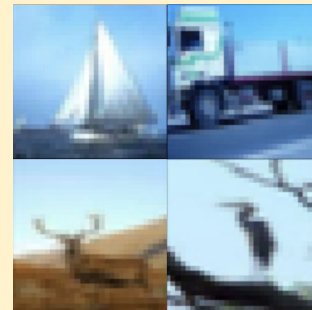
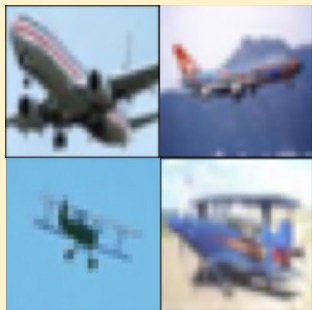


$p(\text{plane} \mid \text{sky}) \approx 1$

Direction in Parameter Space

$p(\text{other} \mid \text{sky}) \approx 1$

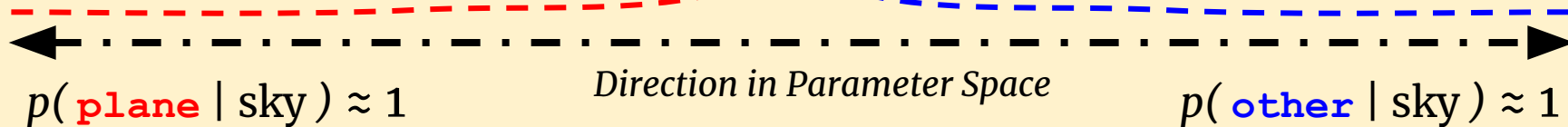
[1] A Walk with SGD. Xing et al. 2018.

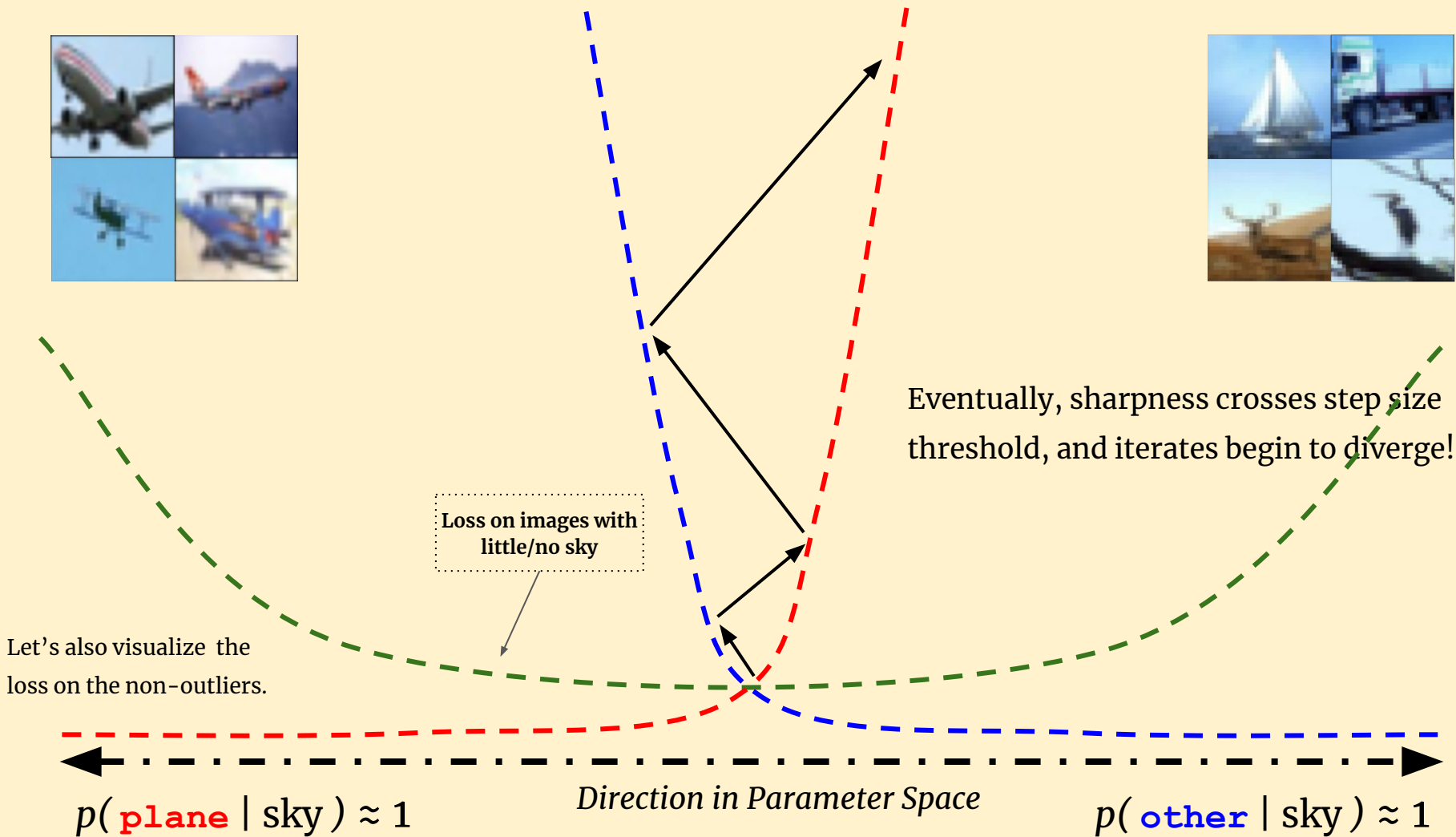
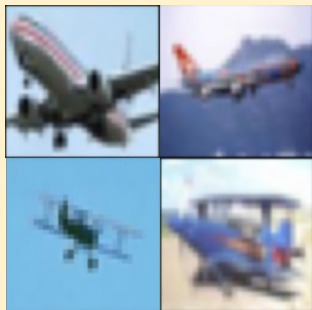


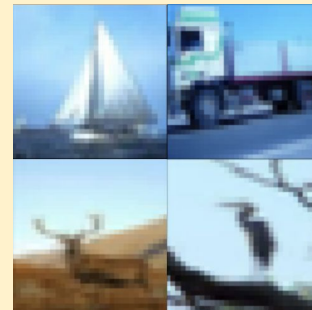
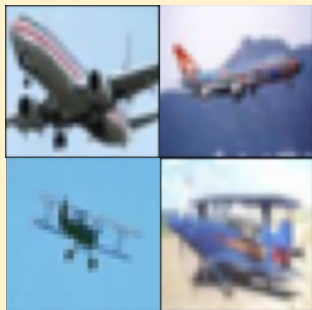
What happens when norm of “sky” grows?

Sensitivity to *how we use the sky feature* grows.

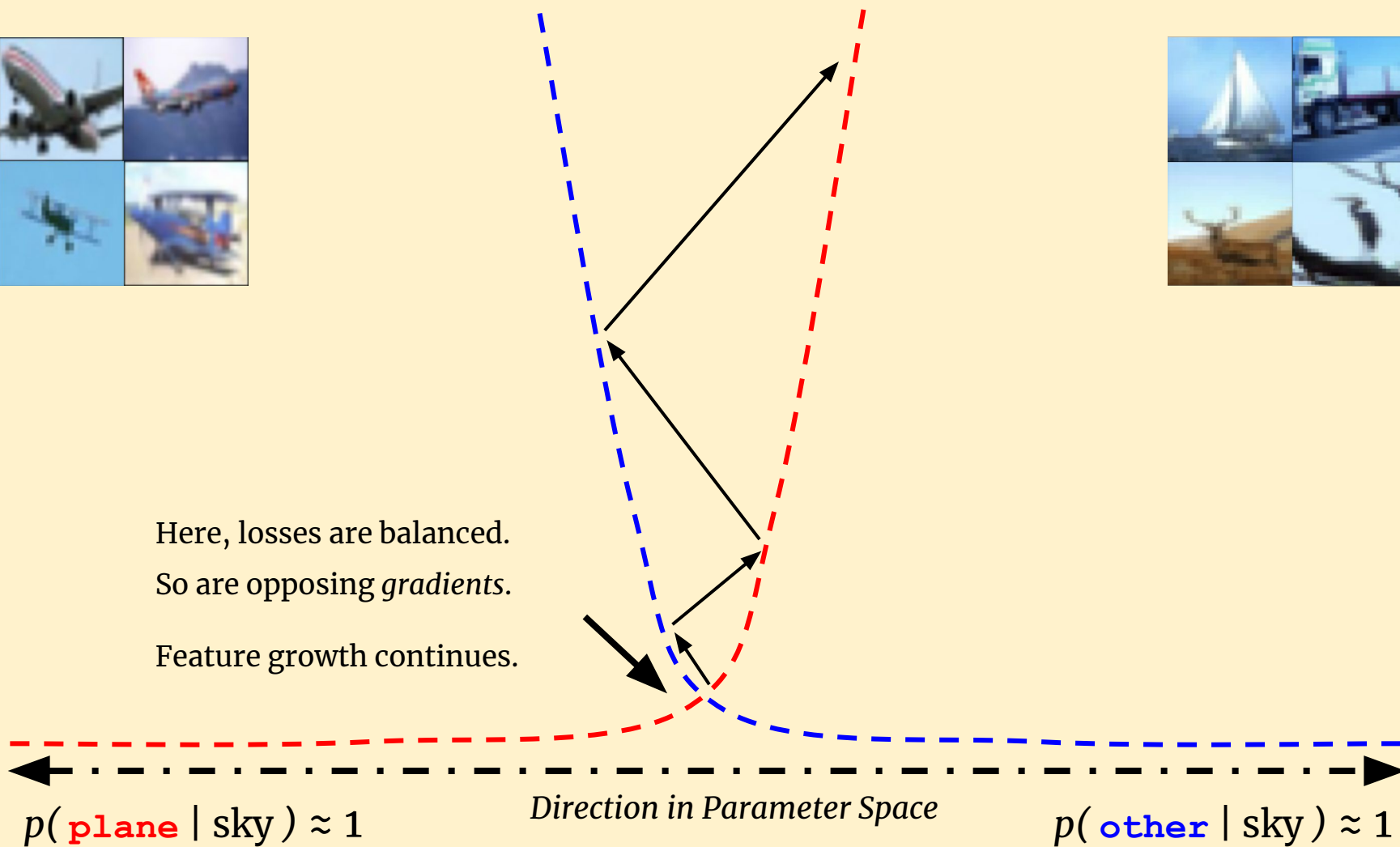
Hence, the loss **sharpens** along this direction.

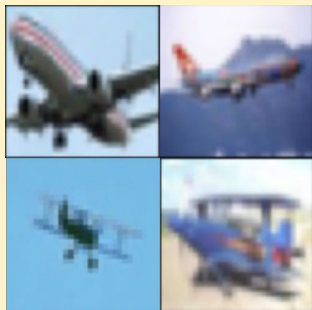




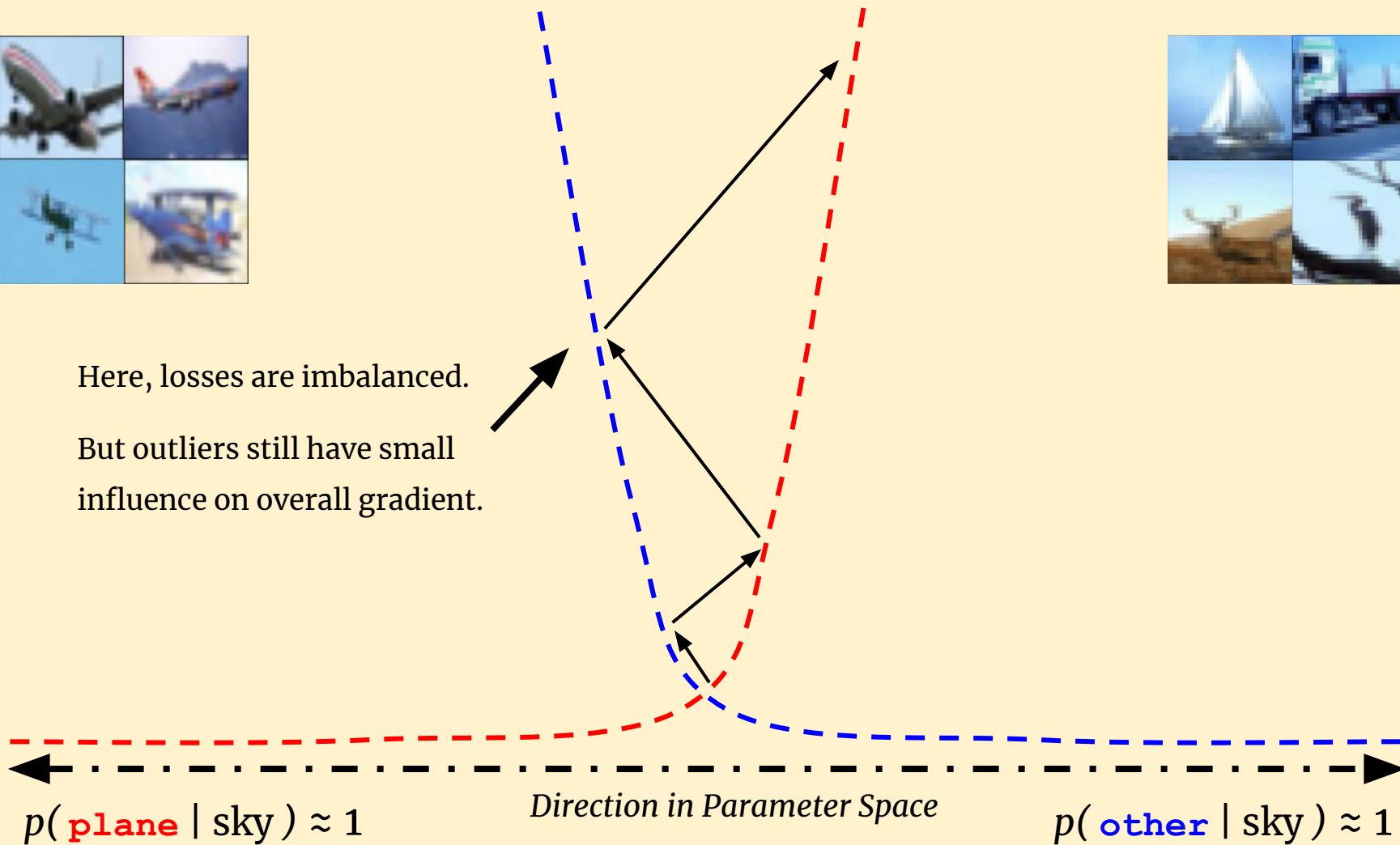


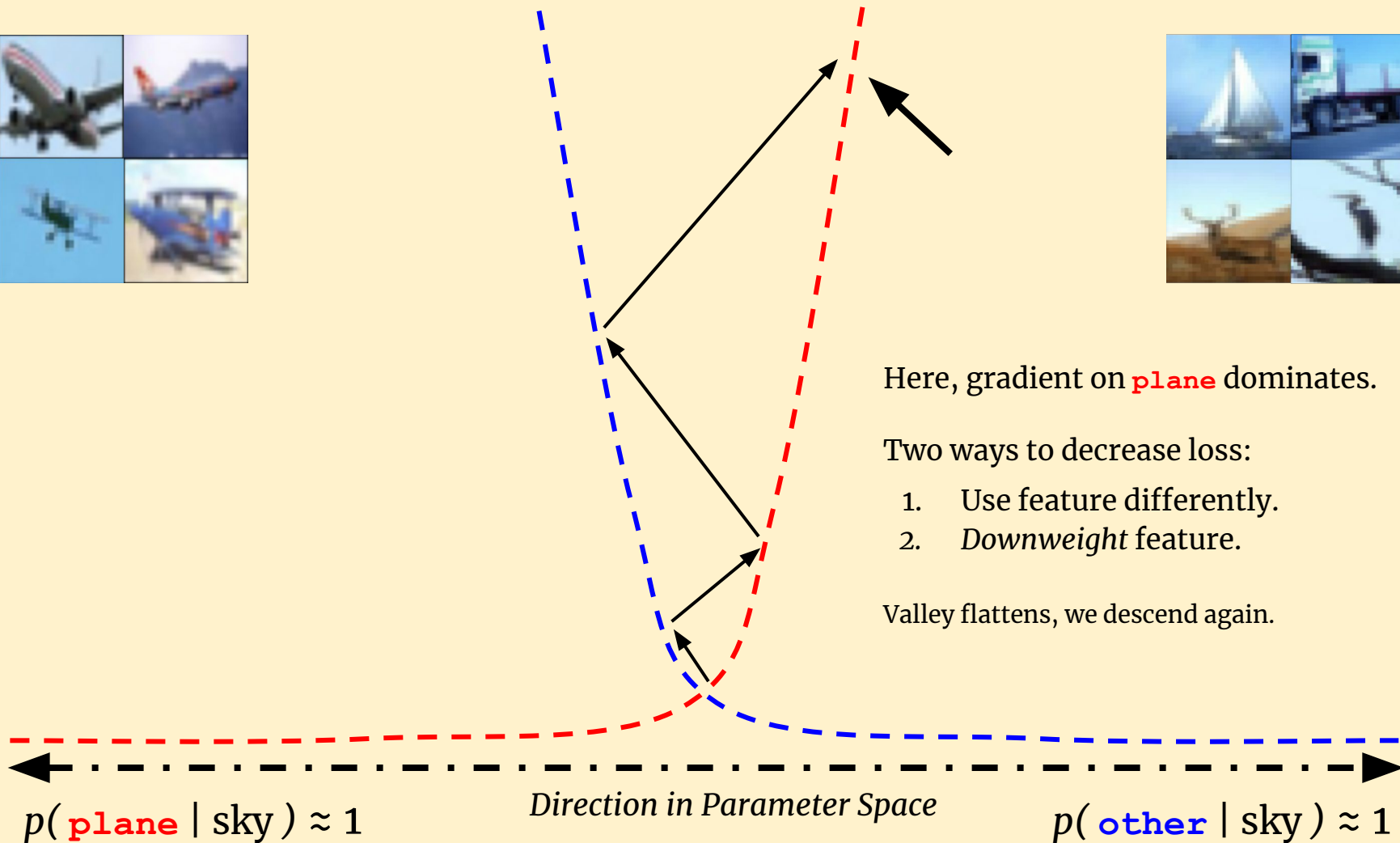
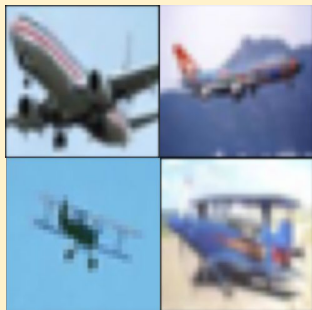
Here, losses are balanced.
So are opposing *gradients*.
Feature growth continues.





Here, losses are imbalanced.
But outliers still have small
influence on overall gradient.





Experimental Verification

The value of a theory (even a non-rigorous one) is in its ability to make predictions.

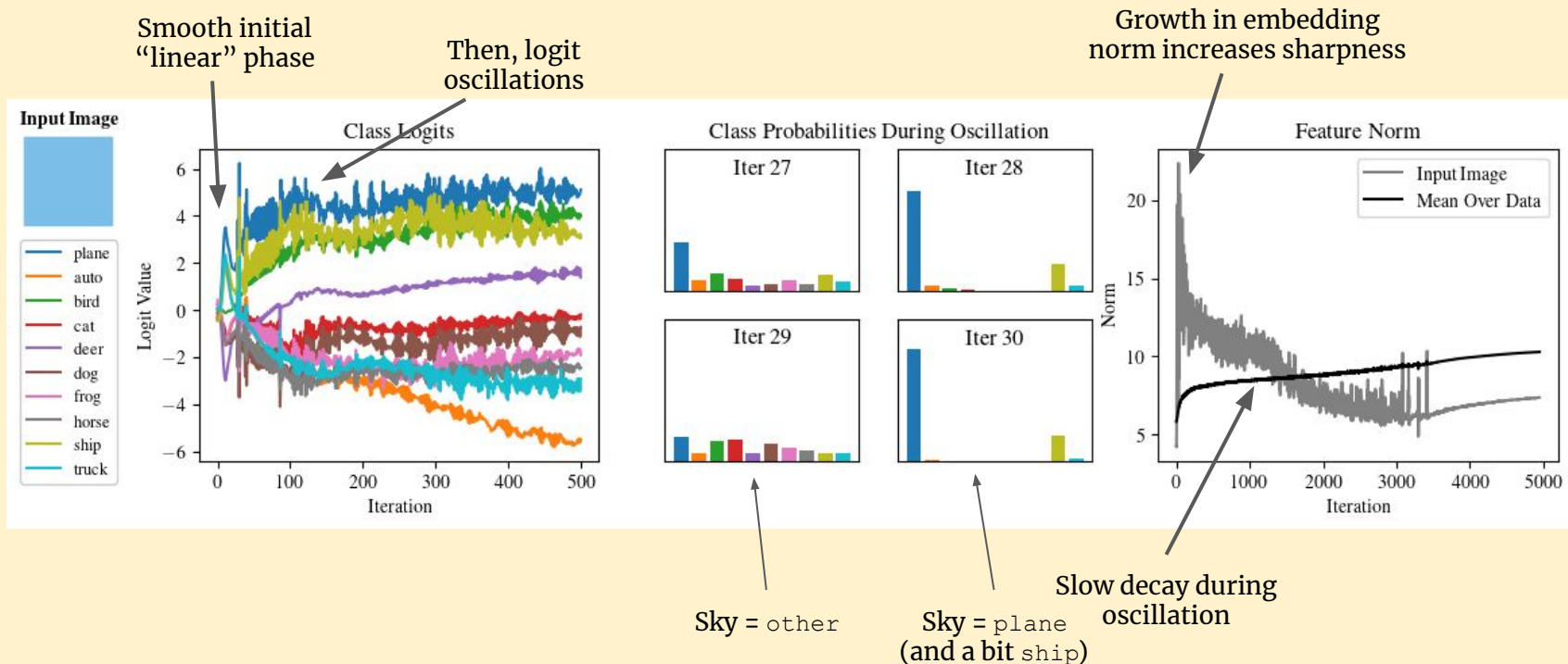
So far we've described:

1. Initial phase of fitting a “linear” model. ← (previously observed)
2. Growth in activation magnitude among images with this feature. ← (least well understood)
3. Upon reaching Edge of Stability, predictions *oscillate* between “sky = plane” and “sky = other”.
4. Oscillation results in shrinking of activation magnitude.

What does this story imply, *behaviorally*? Can we test it more directly?

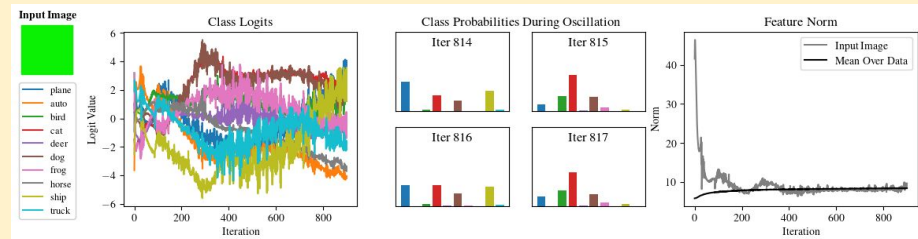
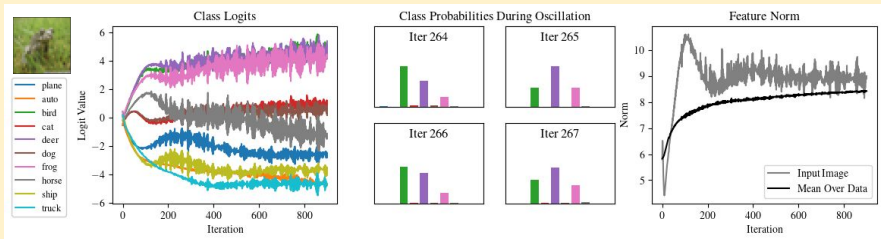
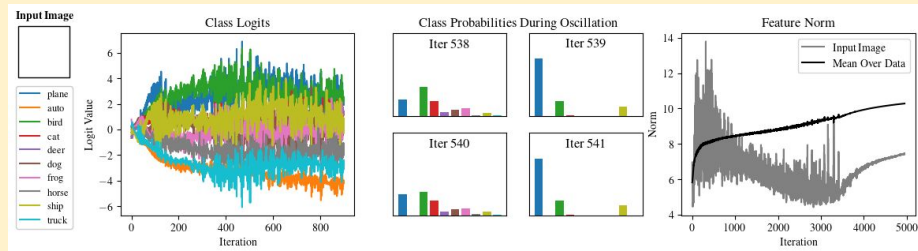
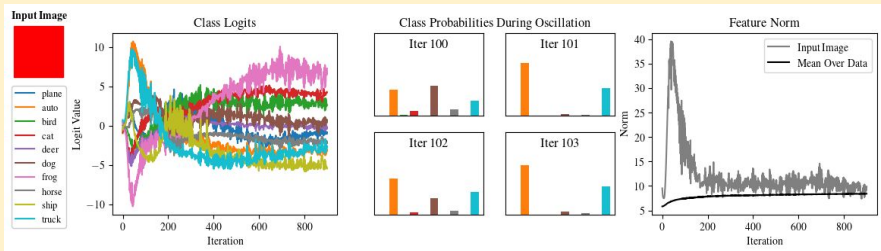
Experimental Verification

To avoid confounders, we'll pass a pure "sky" image through a ResNet-18.



Experimental Verification

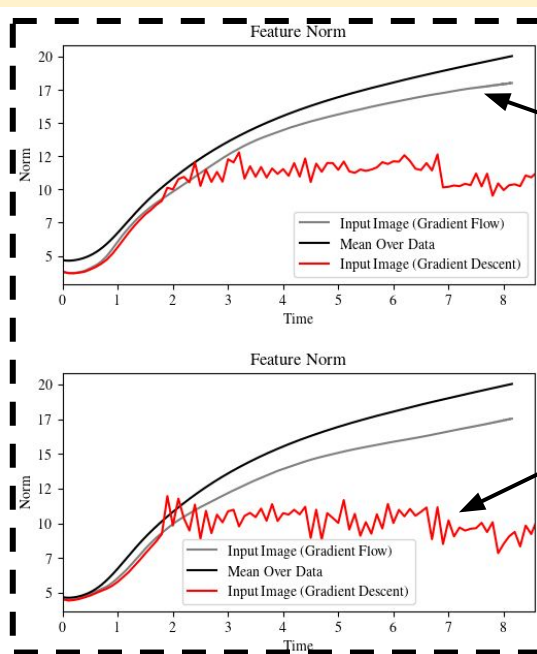
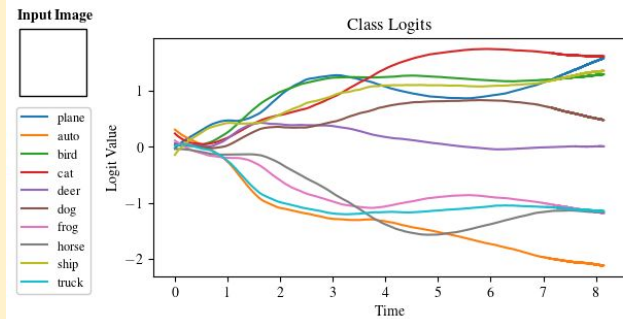
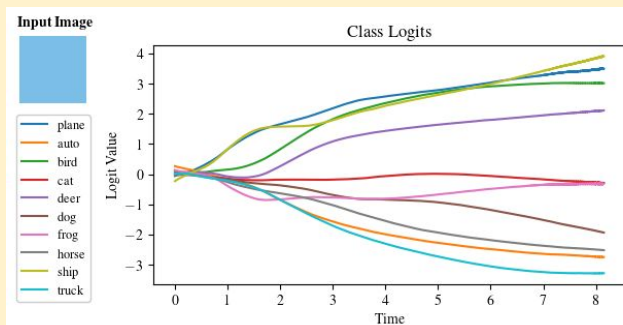
(Doesn't happen as cleanly for all archs/colors, but it's pretty consistent.)



Experimental Verification

Oscillation seems valuable for downweighting the “simple” but “incomplete” features.

- *Gradient Flow doesn't oscillate.* Maybe that's part of why it generalizes poorly?



Under gradient flow, feature norm grows continuously.

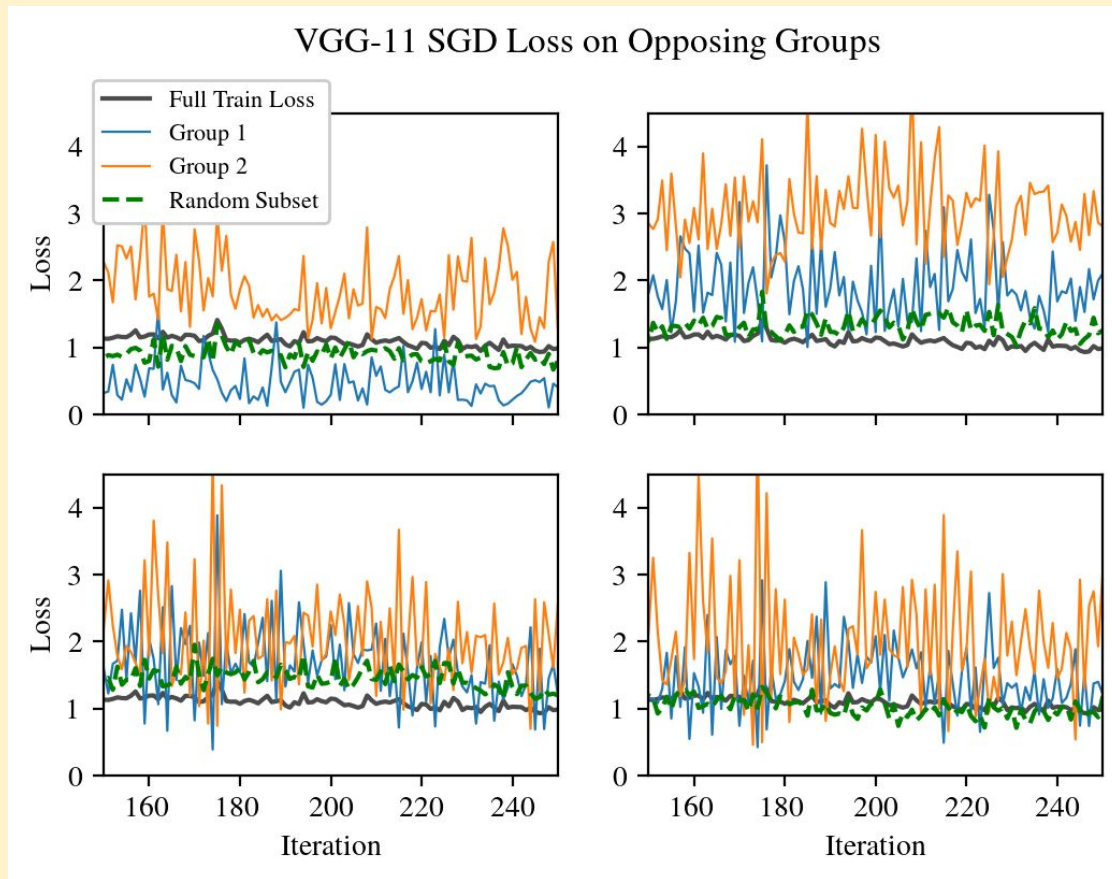
Gradient descent matches flow initially, but *norm starts decreasing* once oscillation begins

Does this Occur for SGD?

Long story short, Yes.

Alternations are not every step.

Groups are not always opposite.



Opposing Signals have clear *potential* connections to existing tools in stochastic optimization, for both training speed and generalization:

- *Batch Normalization*
- *Adaptive Gradient Methods*
- *Sharpness-Aware Minimization*
- *Large Initial Learning Rate*

Maybe these methods work because of how they handle Opposing Signals?

- **Could this help us design new improvements to SGD?**

Lots of unanswered questions.
Very happy to discuss further.

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

Elan Rosenfeld & Andrej Risteski
<https://arxiv.org/abs/2311.04163>



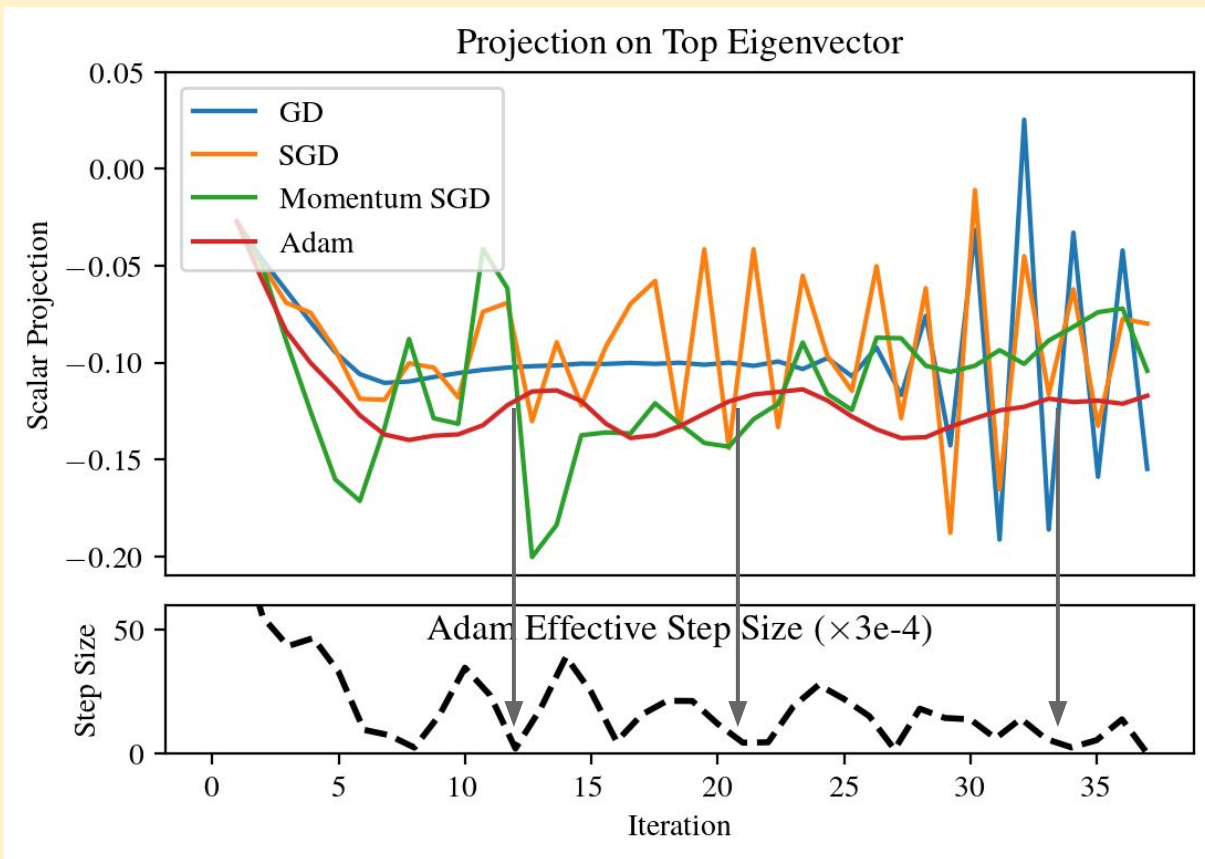
Implications for Stochastic Optimization

A Case Study of Adam vs. SGD

Adam looks markedly different!

Prevents steps that would approach the local minimum

Effective step size drops sharply when approaching valley floor.



Remember this?

