On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay


Ekaterina Lobacheva*


Maxim Kodryan*


Nadezhda Chirkova


Andrey Malinin


Dmitry Vetrov

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

SAMSUNG Research

Y Research

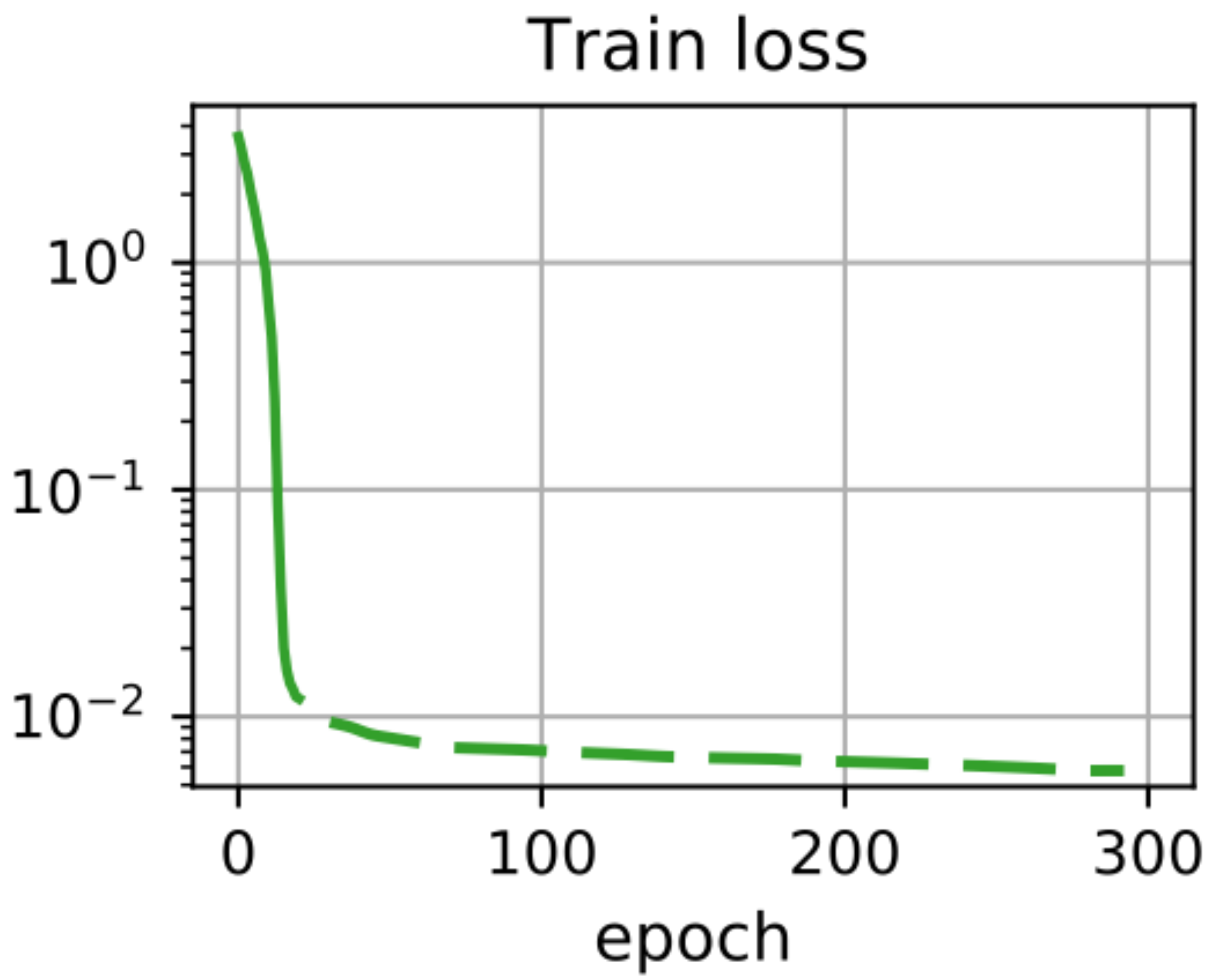
p(B|A)yesgroup.ru

* equal contribution

The beginning of the story

- ResNet on a CIFAR-100
- Training using SGD with a fixed learning rate

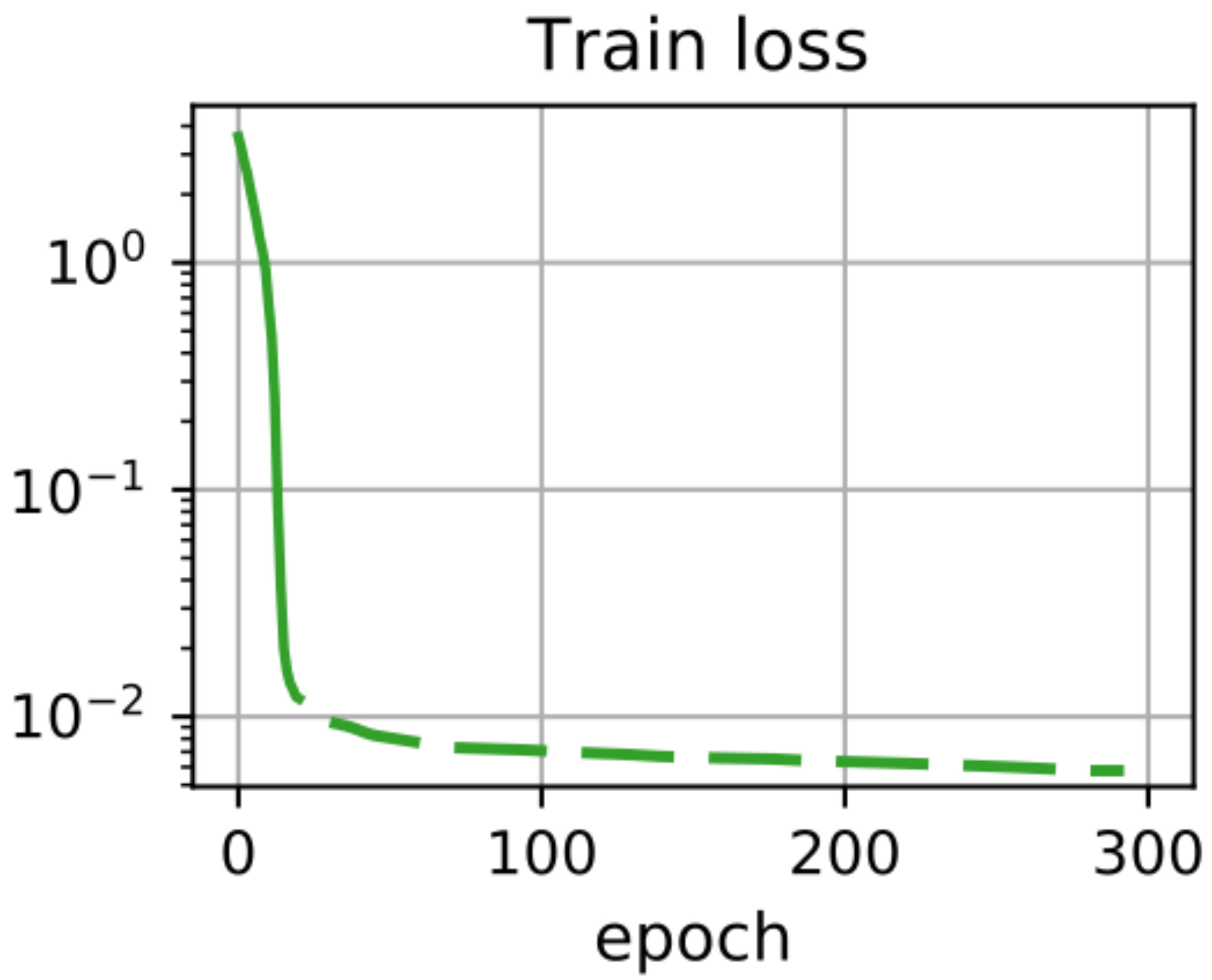
We expect convergence



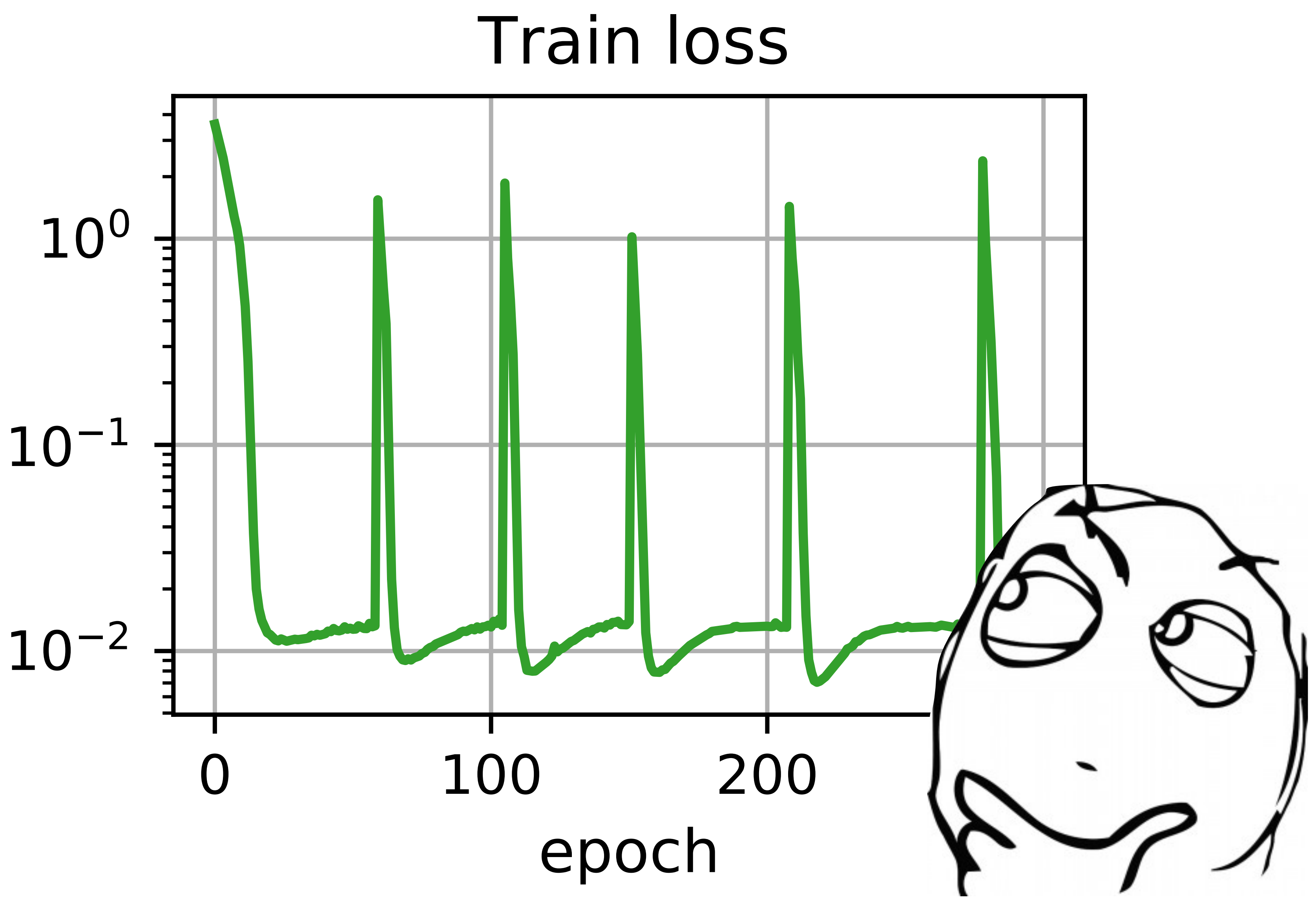
The beginning of the story

- ResNet on a CIFAR-100
- Training using SGD with a fixed learning rate

We expect convergence



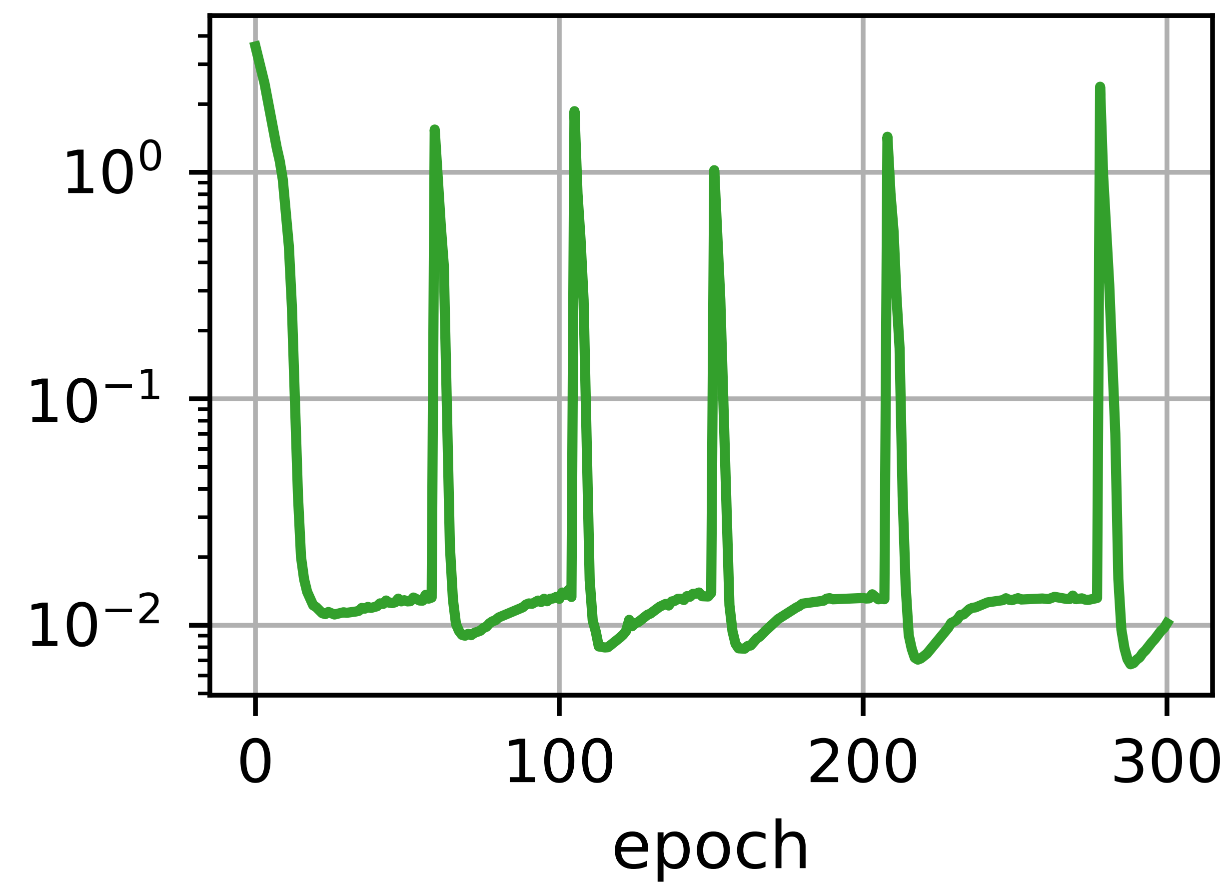
We get … periodic behavior?



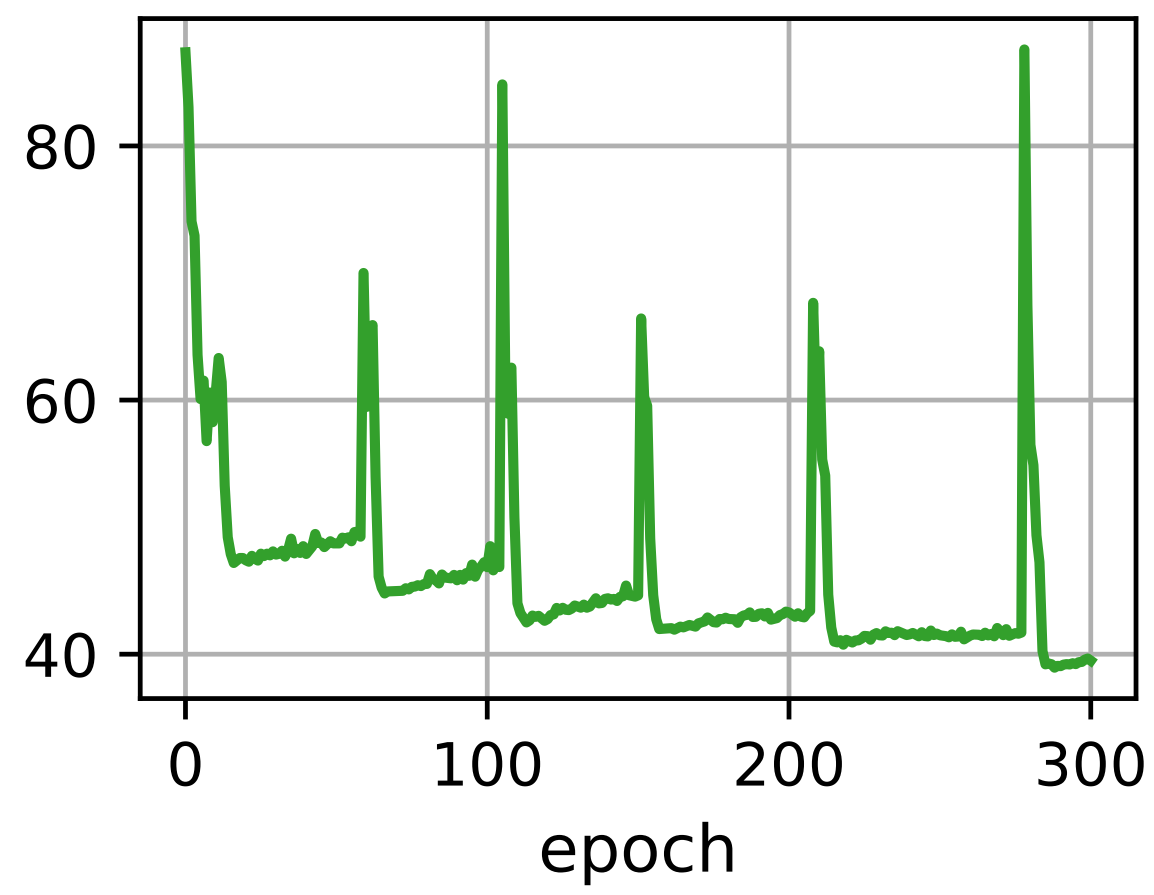
Overview

We investigate the periodic behaviour of neural networks during training

Train loss



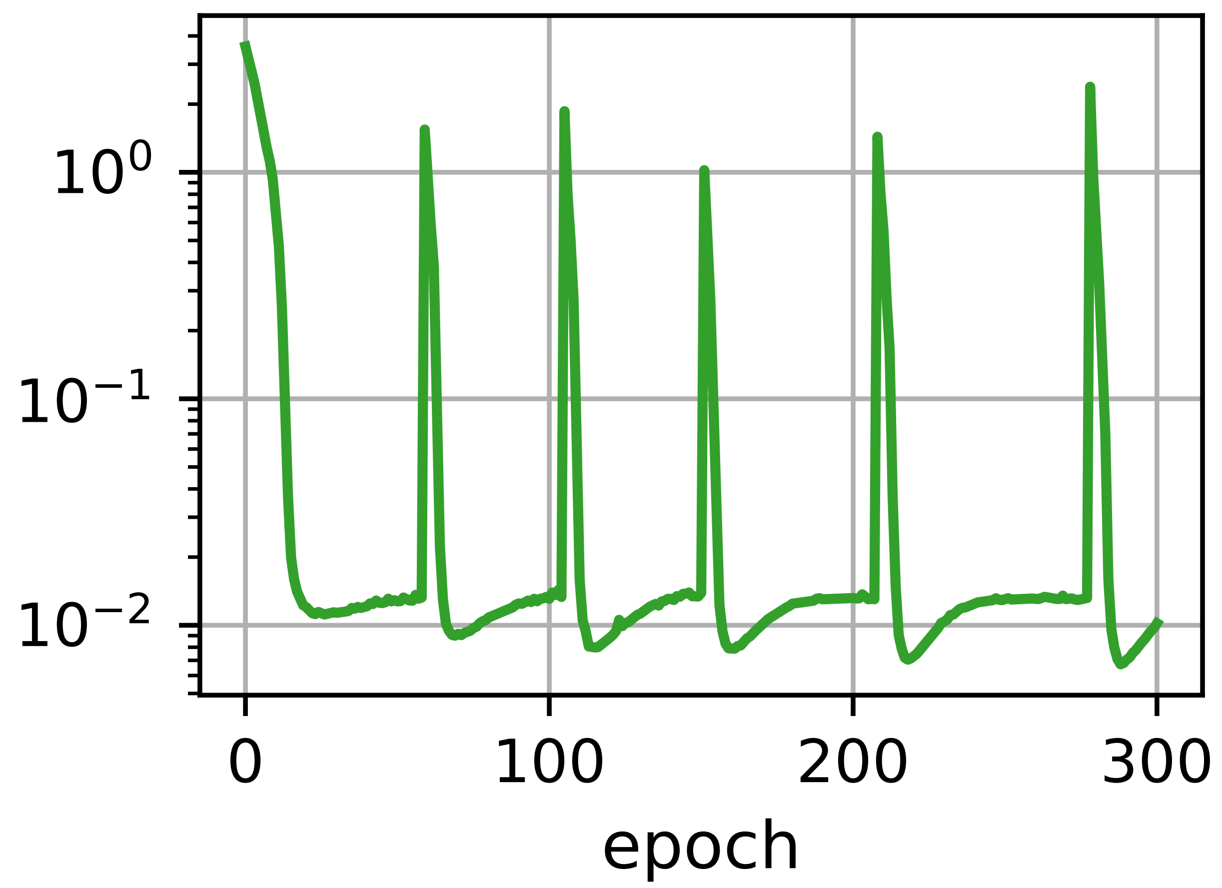
Test error, %



Overview

Goal 1. Find the reasons

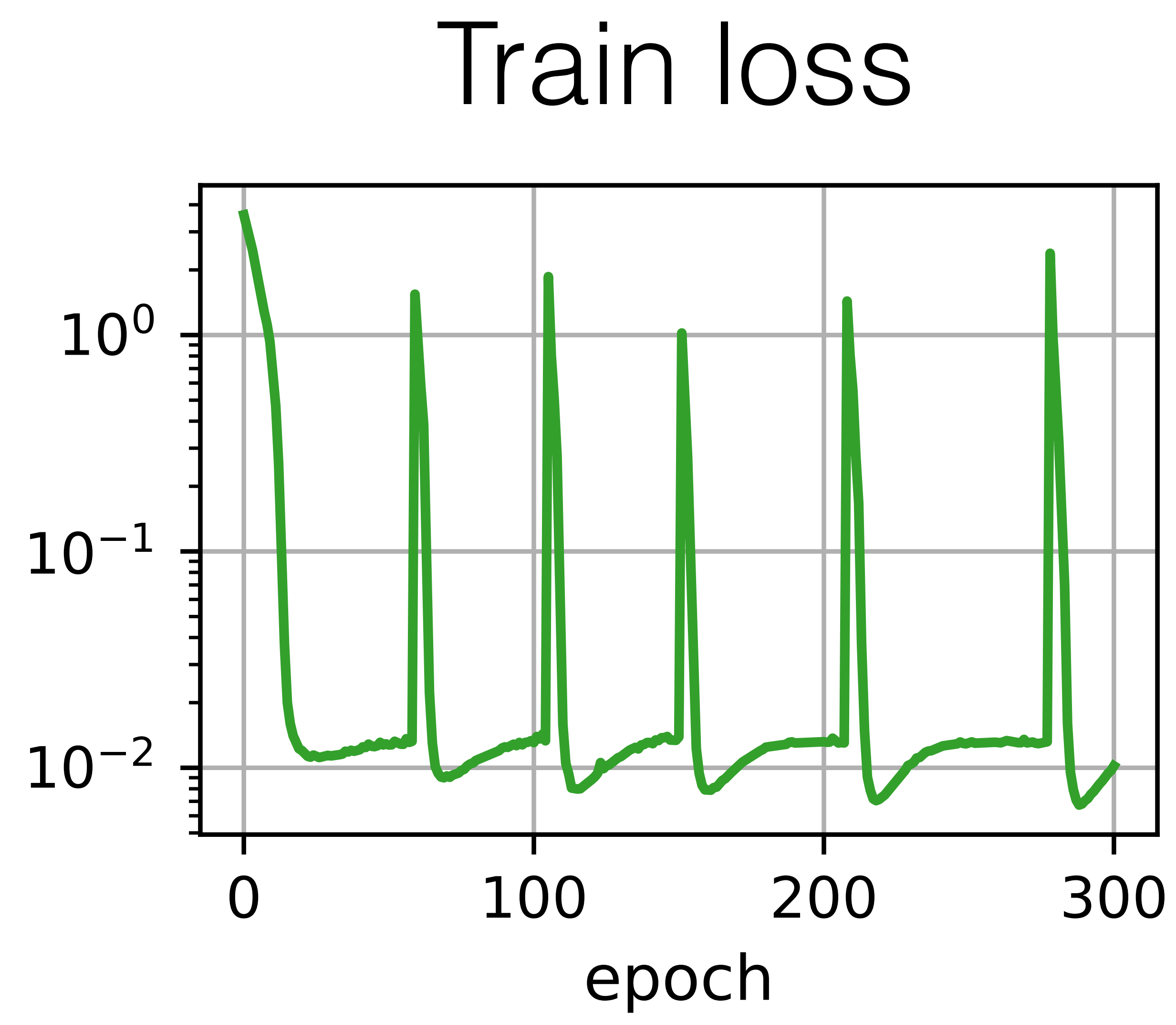
Train loss



Overview

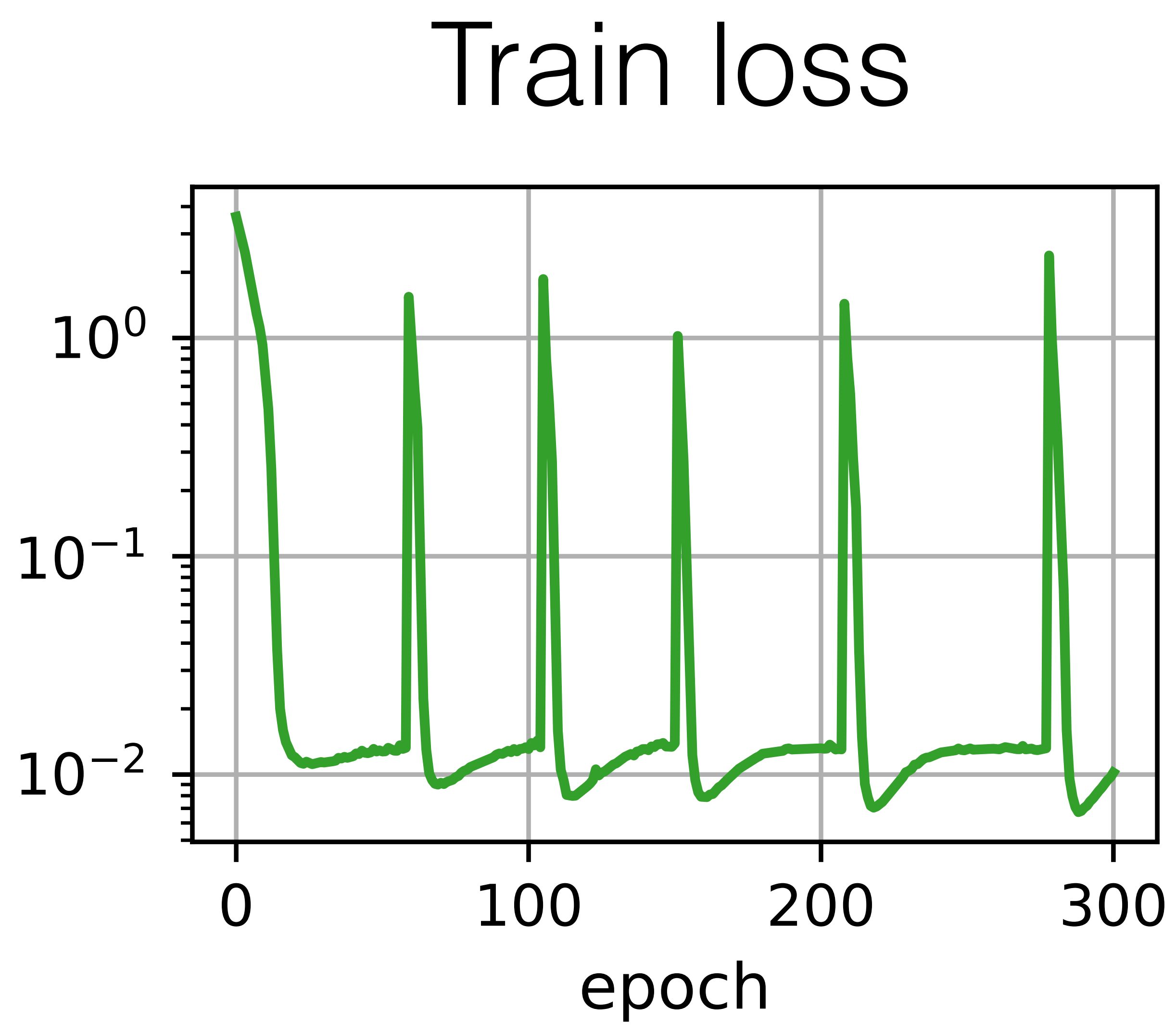
Goal 1. Find the reasons - empirical and theoretical justification

BatchNorm + Weight Decay



Overview

Goal 1. Find the reasons - empirical and theoretical justification



BatchNorm + Weight Decay

instabilities in low weight norm region

low weight norm

Overview

Goal 1. Find the reasons - empirical and theoretical justification

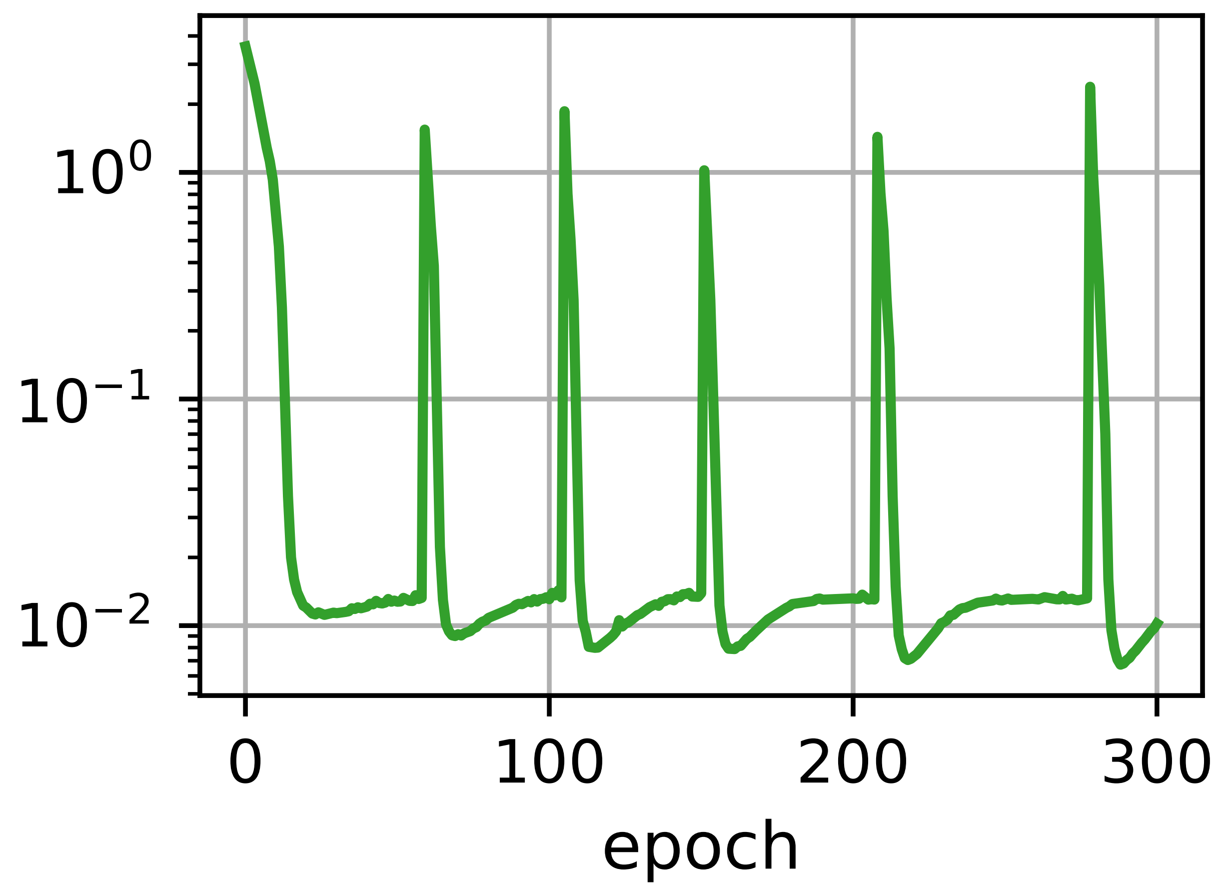
BatchNorm + Weight Decay

instabilities in low weight norm region

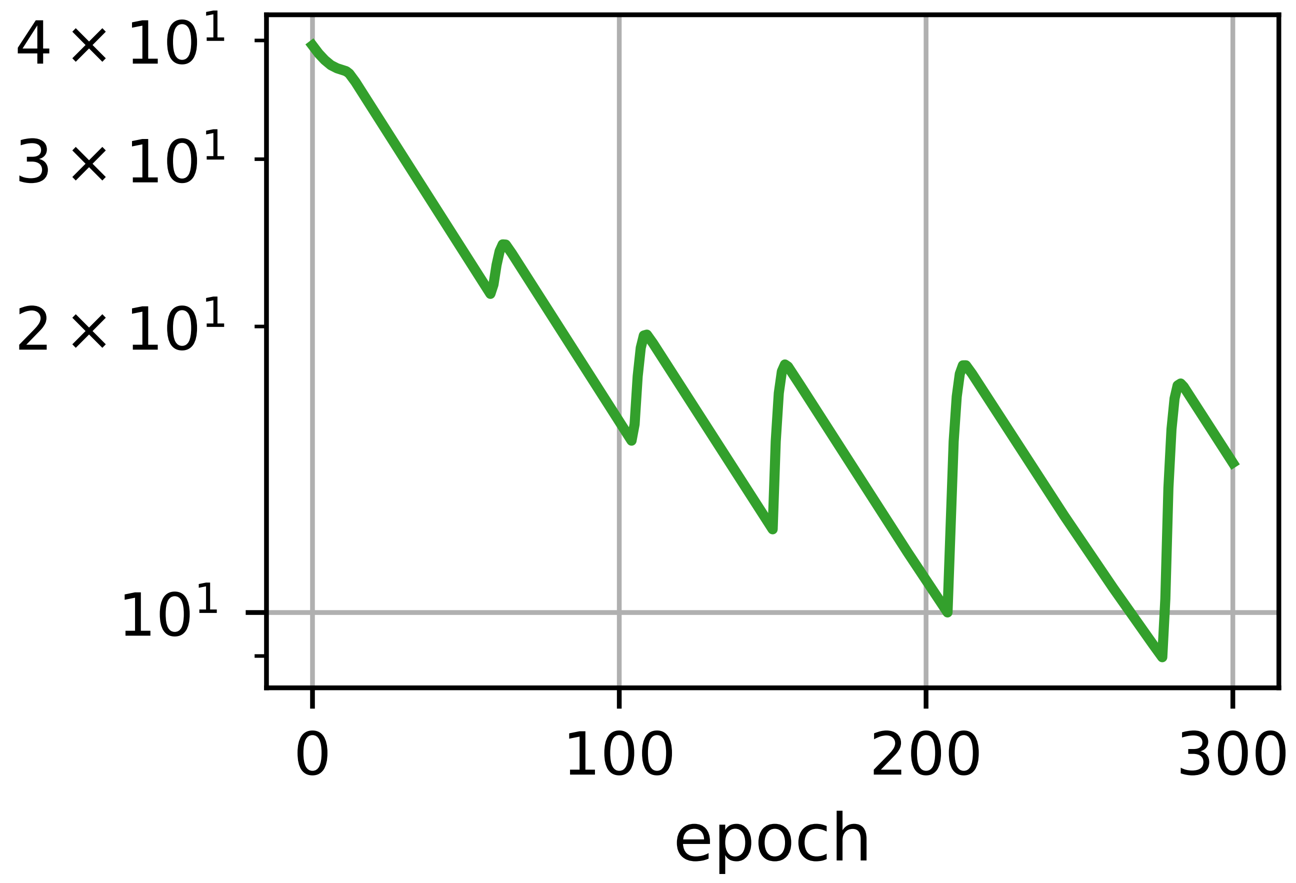
low weight norm

periodic behavior

Train loss



Weight norm



Overview

Goal 1. Find the reasons - empirical and theoretical justification

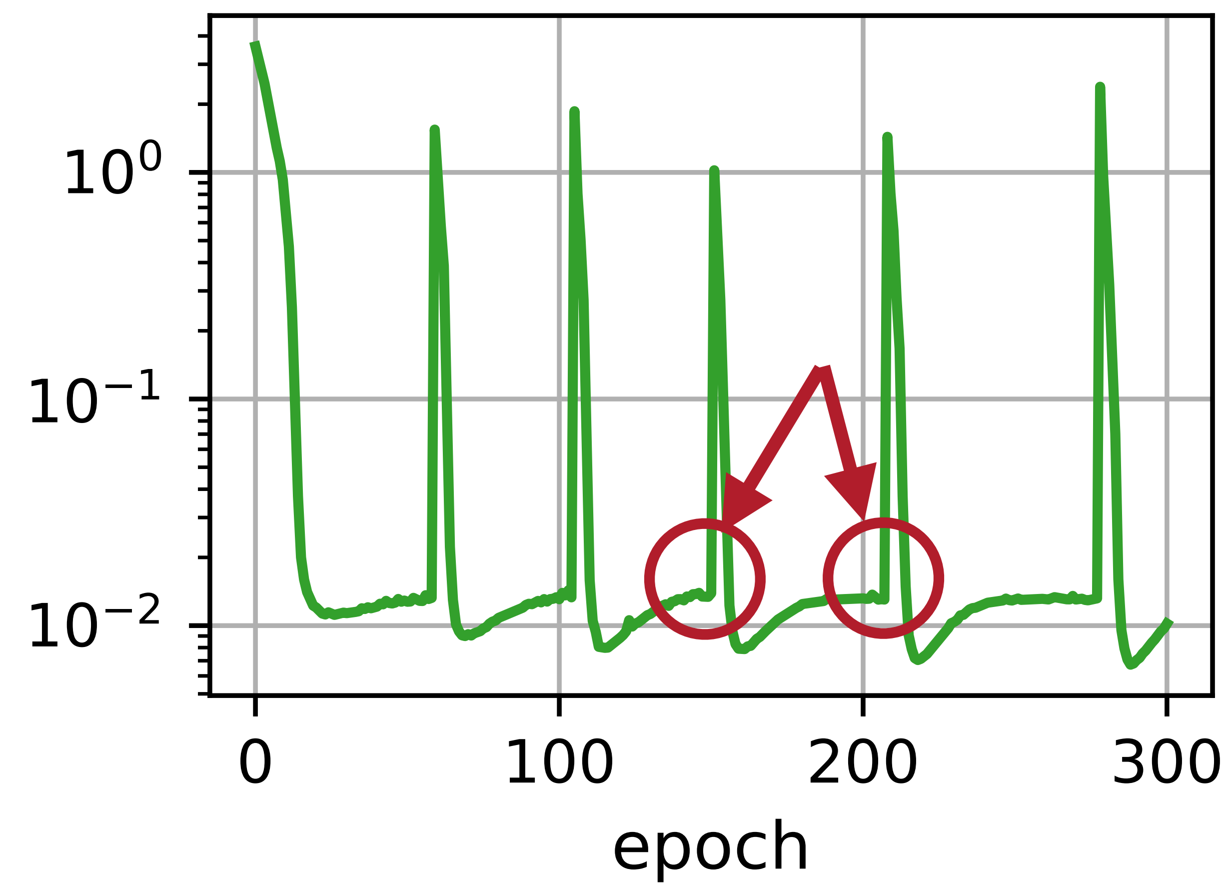
BatchNorm + Weight Decay

instabilities in low weight norm region

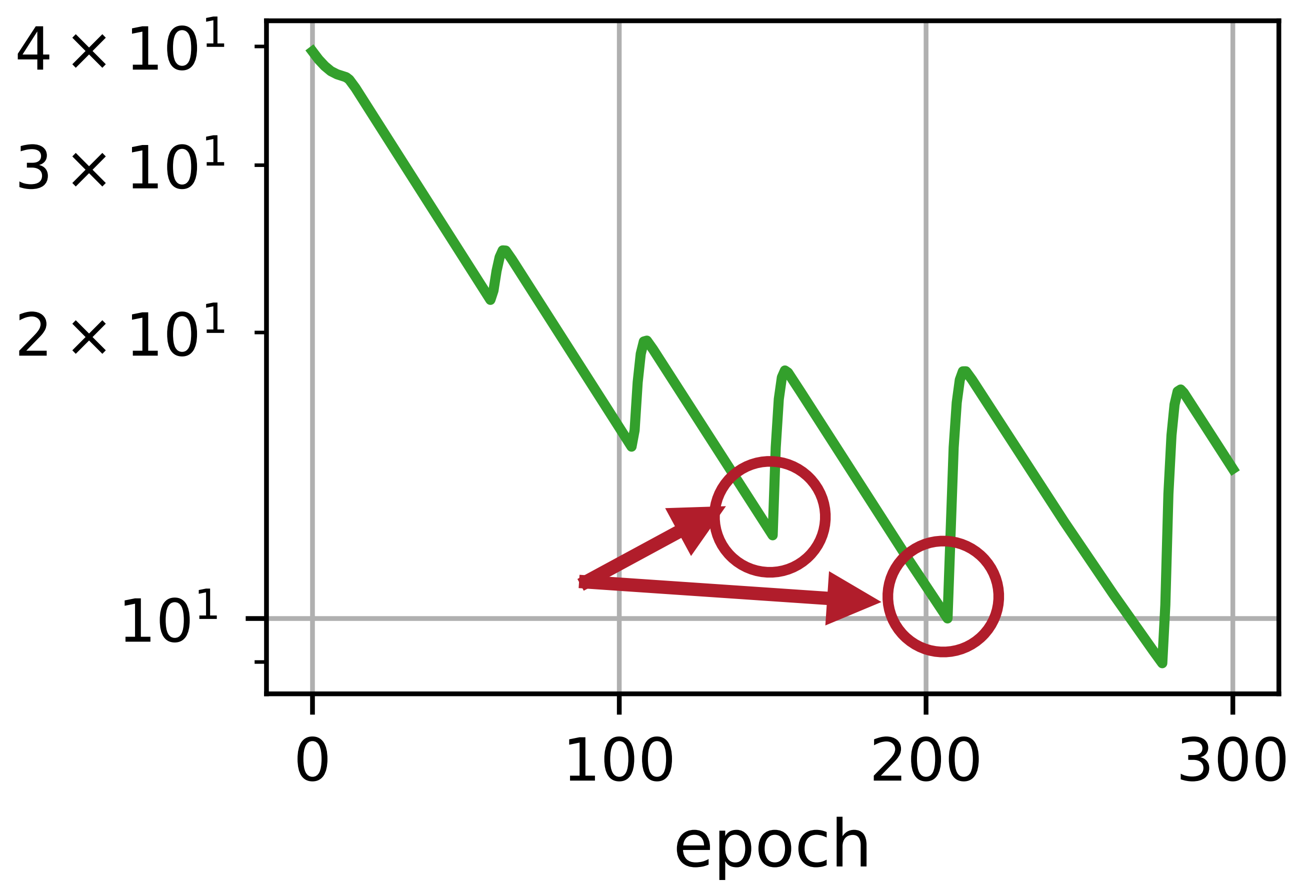
low weight norm

periodic behavior

Train loss



Weight norm



Overview

Goal 1. Find the reasons - empirical and theoretical justification

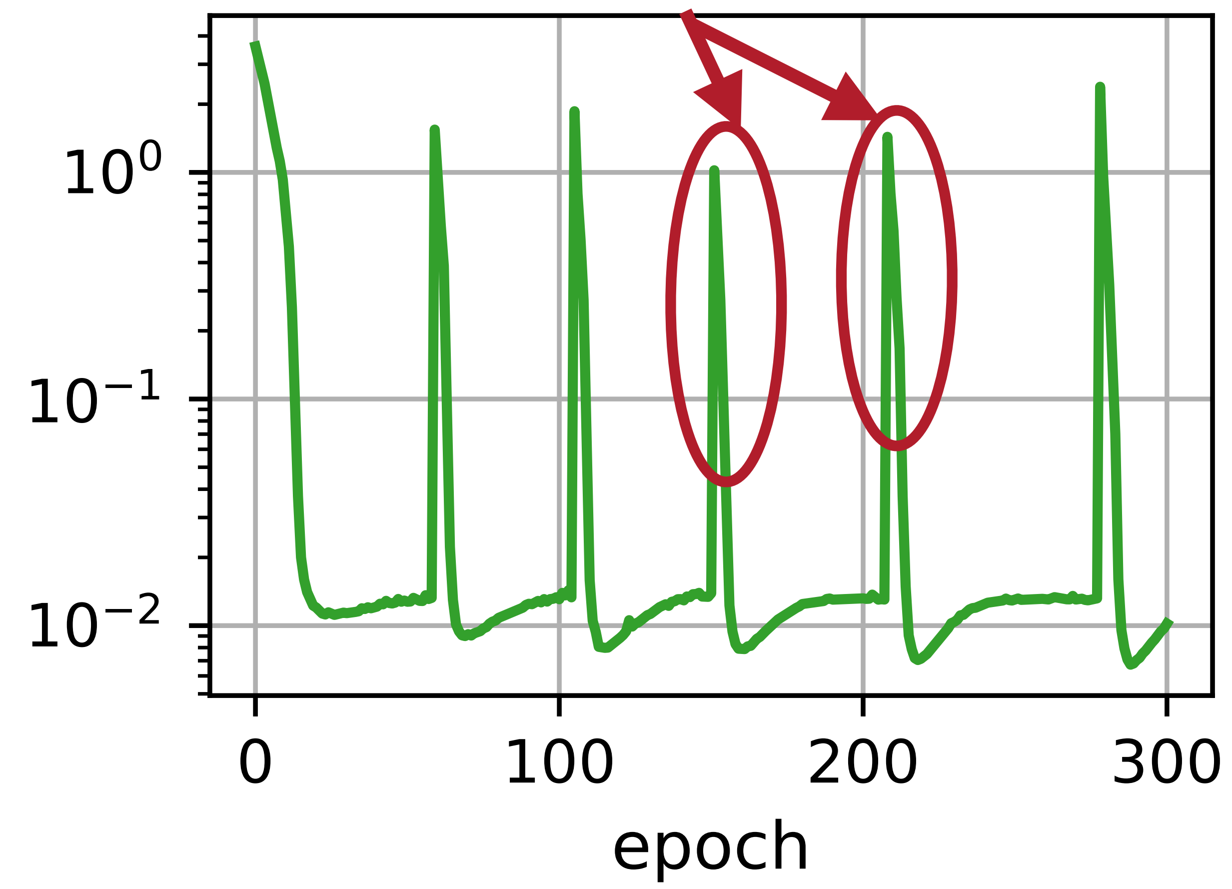
BatchNorm + Weight Decay

instabilities in low weight norm region

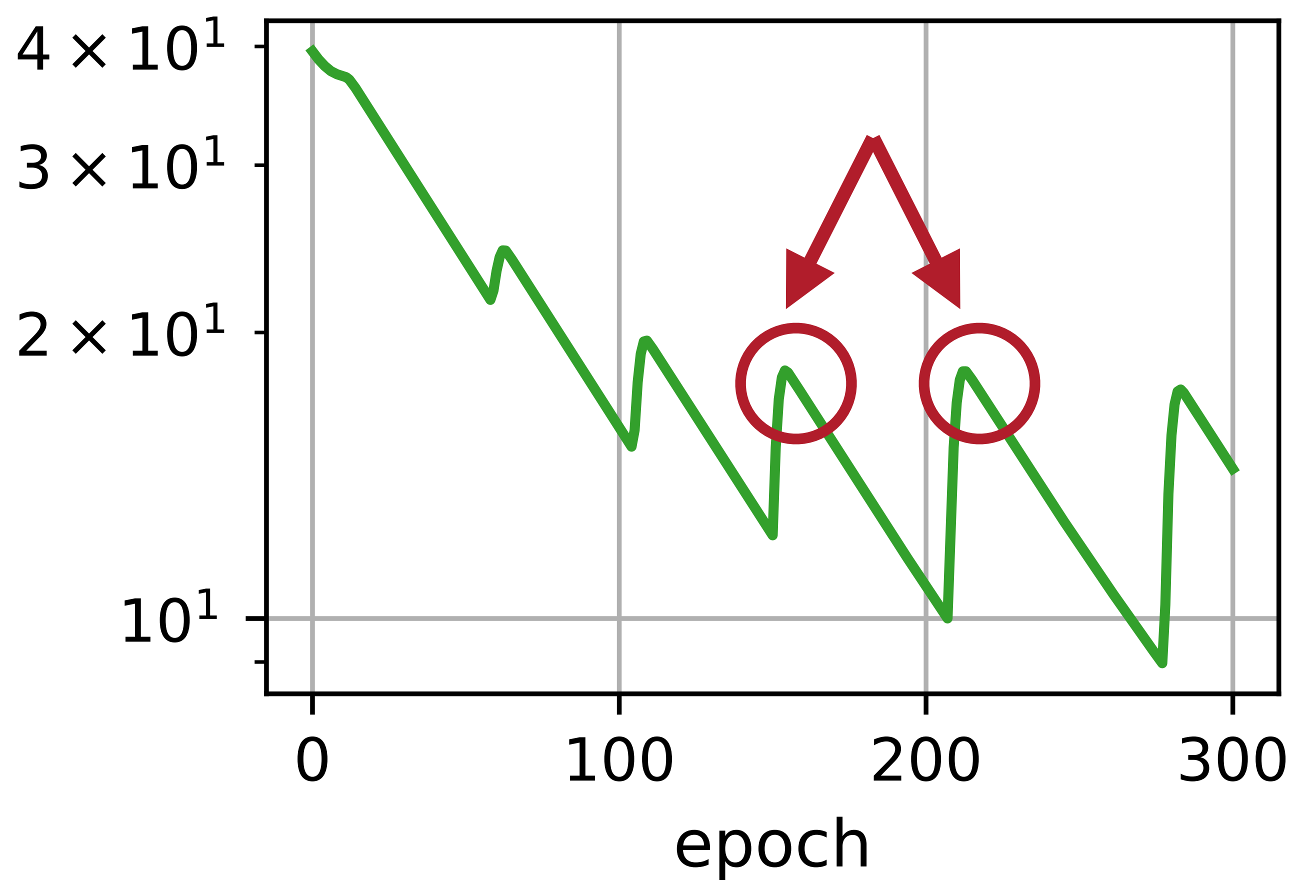
low weight norm

periodic behavior

Train loss



Weight norm



Overview

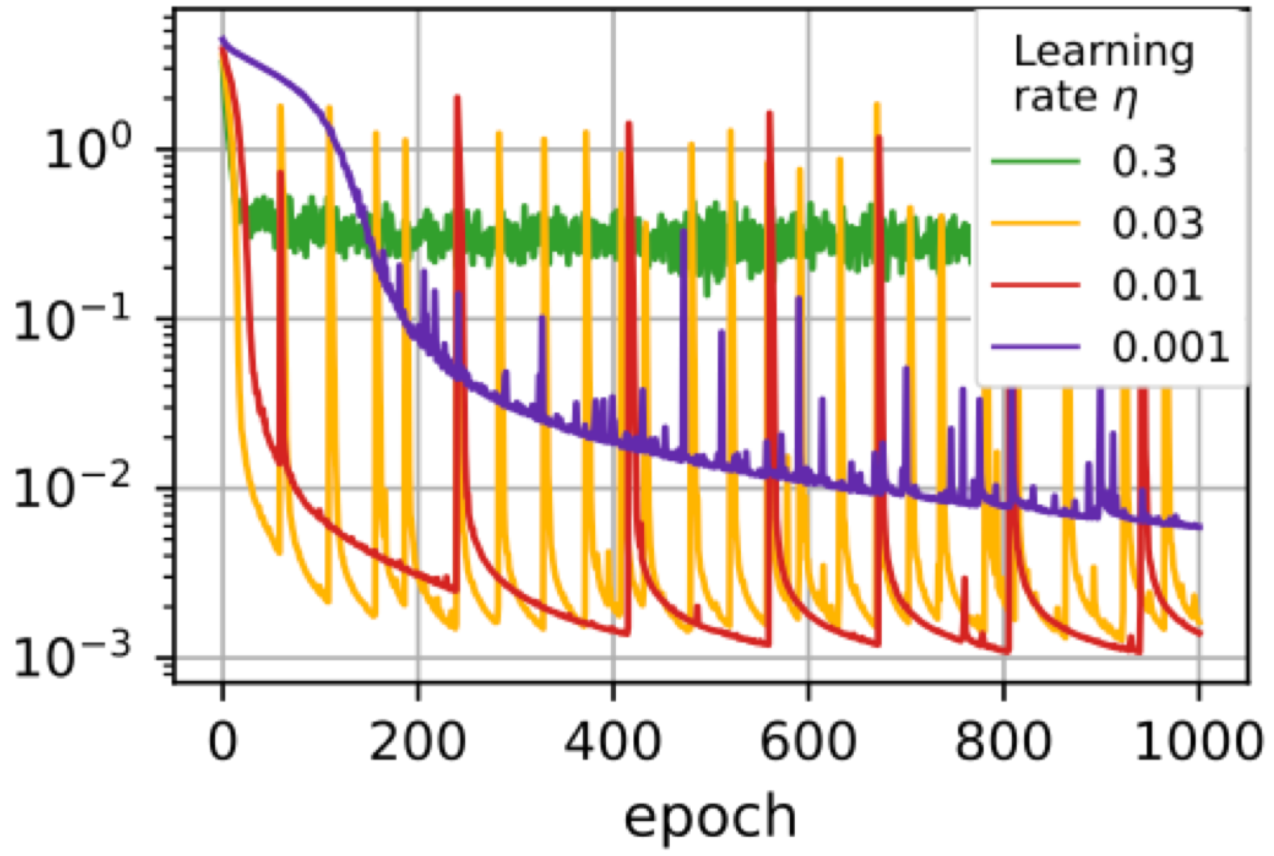
Goal 2. Empirical study:

How hyperparameters influence the behavior?

- Periodic behavior occurs for a wide range of learning rates and weight decays
- Higher learning rate or weight decay results in faster periods

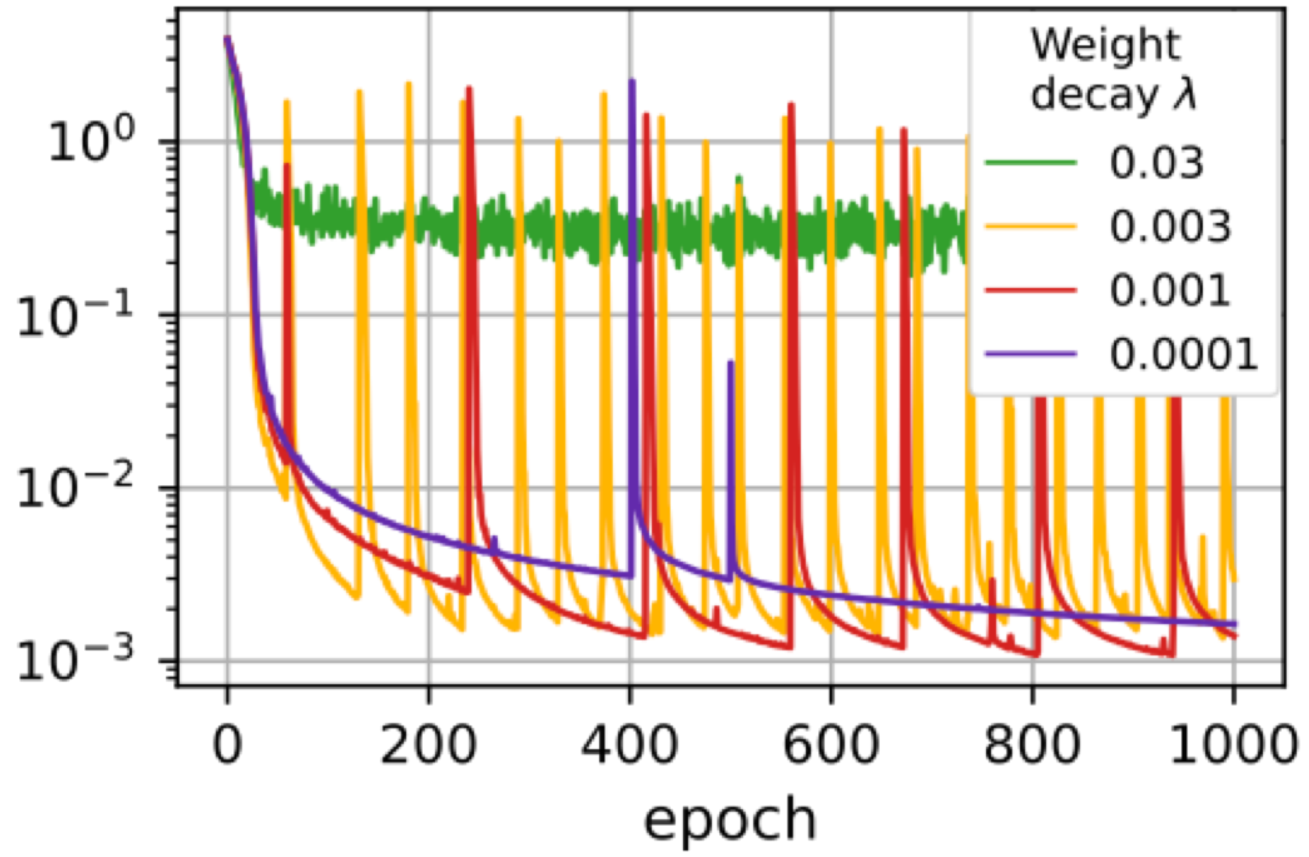
Vary learning rate

Train loss



Vary weight decay

Train loss



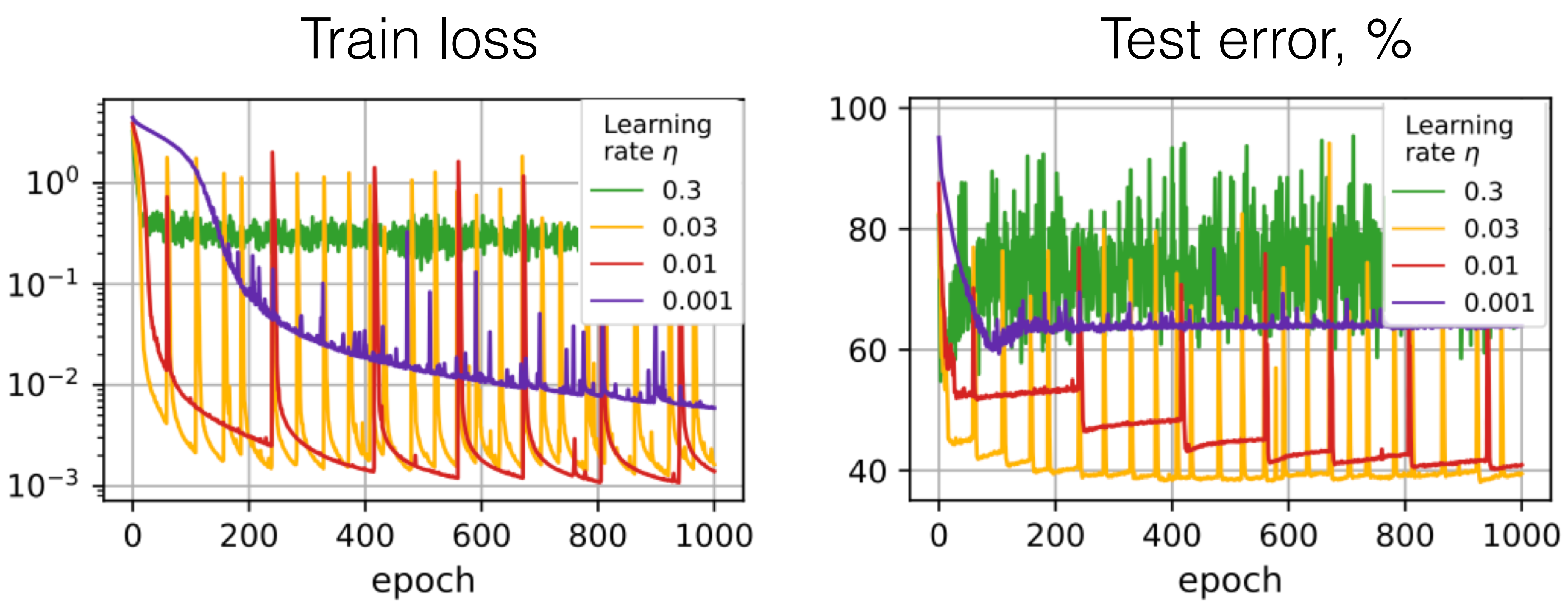
Overview

Goal 2. Empirical study:

How different are the minima at different periods?

- Minima are functionally different
- Usually minima improve with each new period at the beginning of the training

Improvement of minima



Overview

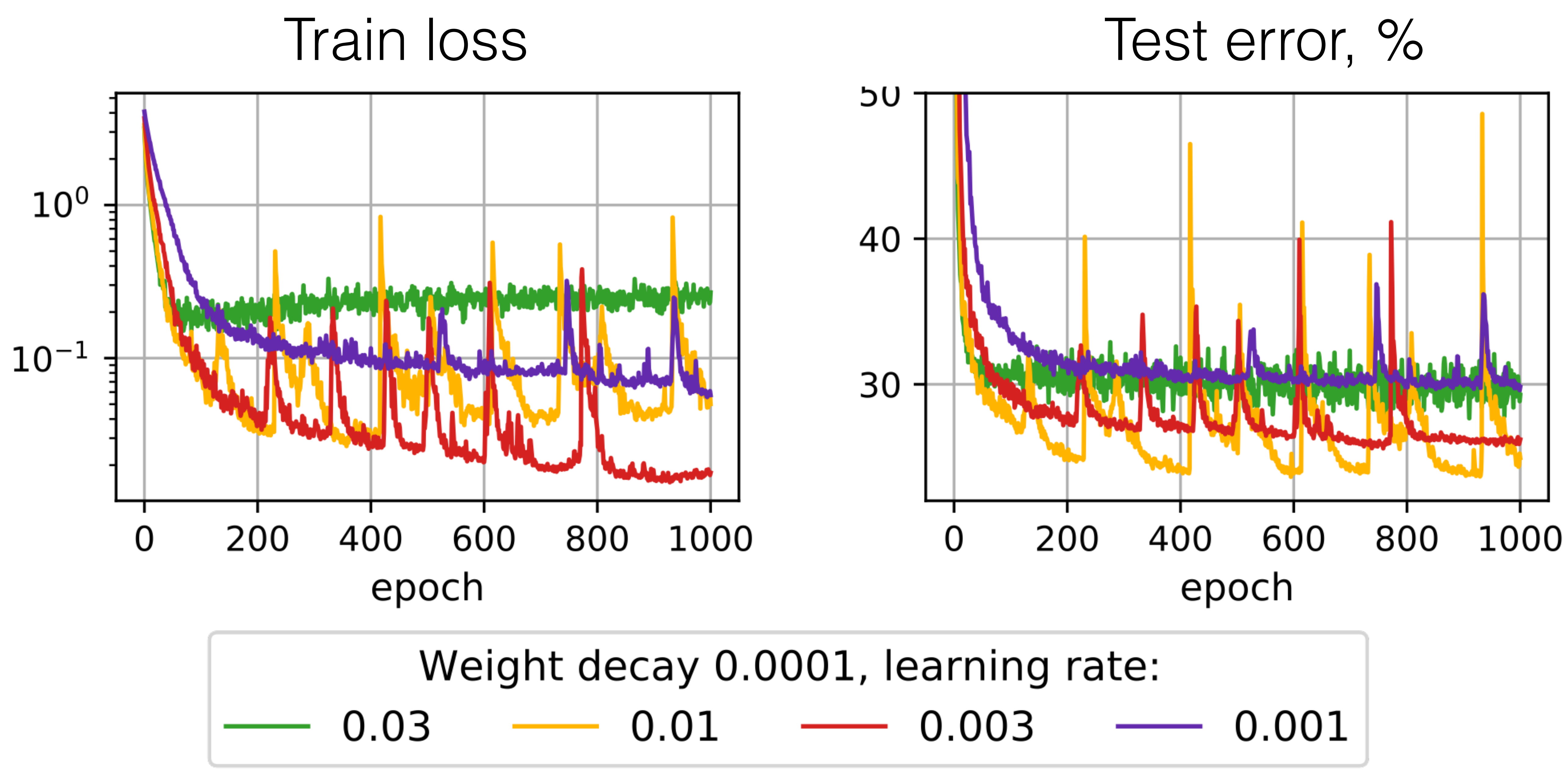
Goal 2. Empirical study:

In what practical settings the periodic behavior may occur?

Settings:

- Standard networks
- SGD with momentum
- Data augmentation
- No learning rate schedule
- Long training

Practical training of ResNet-18 on CIFAR-100



Overview

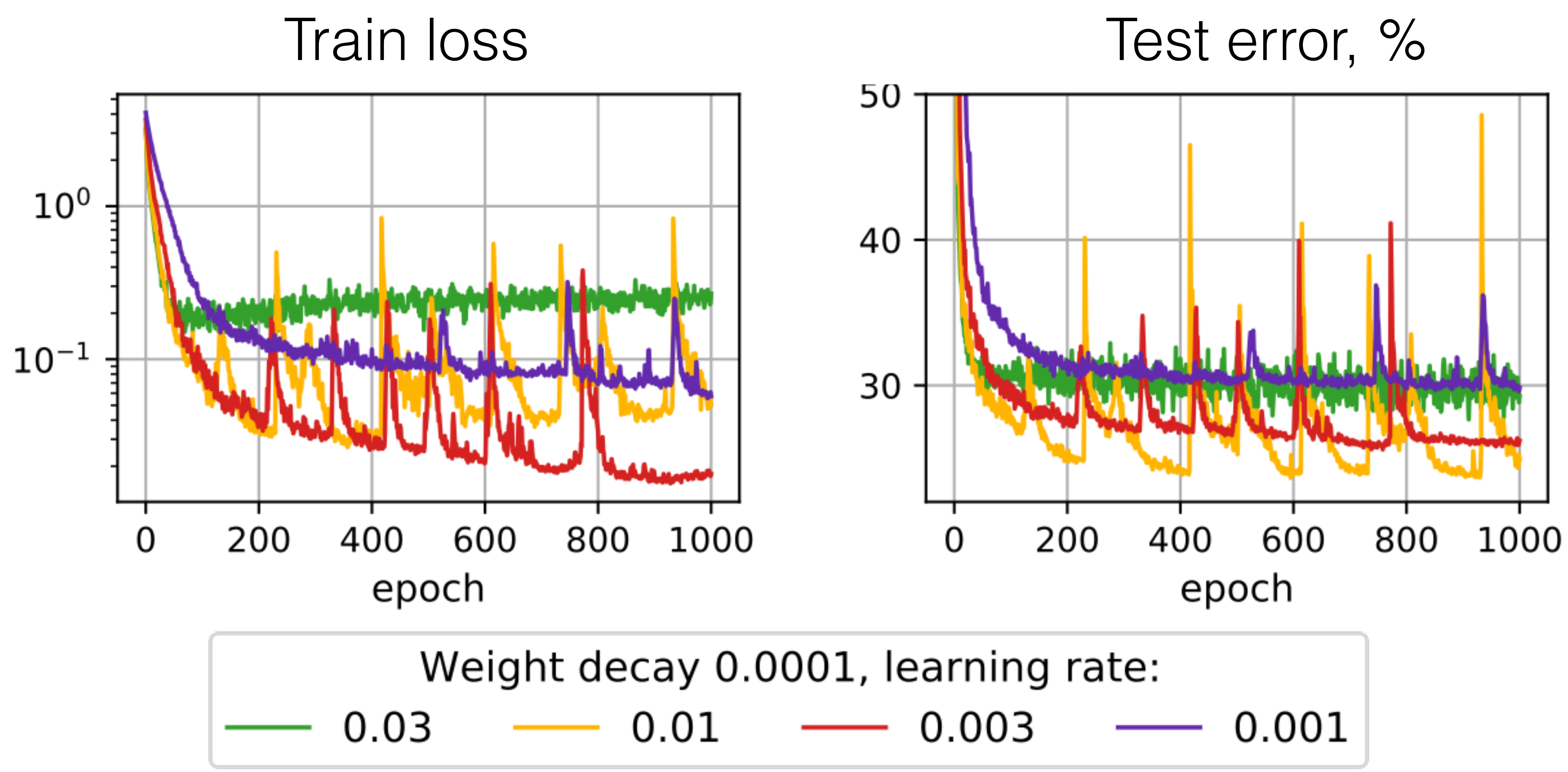
Goal 2. Empirical study:

In what practical settings the periodic behavior may occur?

Settings:

- ✔ Standard networks
- ✔ SGD with momentum
- ✔ Data augmentation
- ▬ No learning rate schedule
- ▬ Long training

Practical training of ResNet-18 on CIFAR-100



Related work

BatchNorm + Weight Decay = ?

Related work

BatchNorm + Weight Decay = ?

→ Equilibrium

- Li et al., 2020. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate.
- Wan et al., 2020. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd.

Related work

BatchNorm + Weight Decay = ?

Equilibrium

- Li et al., 2020. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate.
- Wan et al., 2020. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd.

Instability

- Li and Arora, 2020. An exponential learning rate schedule for deep learning.
- Li et al, 2020. Understanding the disharmony between weight normal-ization family and weight decay.
- Li et al, 2020. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate.

Related work

BatchNorm + Weight Decay = ?

Equilibrium

- Li et al., 2020. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate.
- Wan et al., 2020. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd.

Instability

- Li and Arora, 2020. An exponential learning rate schedule for deep learning.
- Li et al, 2020. Understanding the disharmony between weight normal-ization family and weight decay.
- Li et al, 2020. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate.

Periodic behavior generalizes both views!

Training dynamics explained

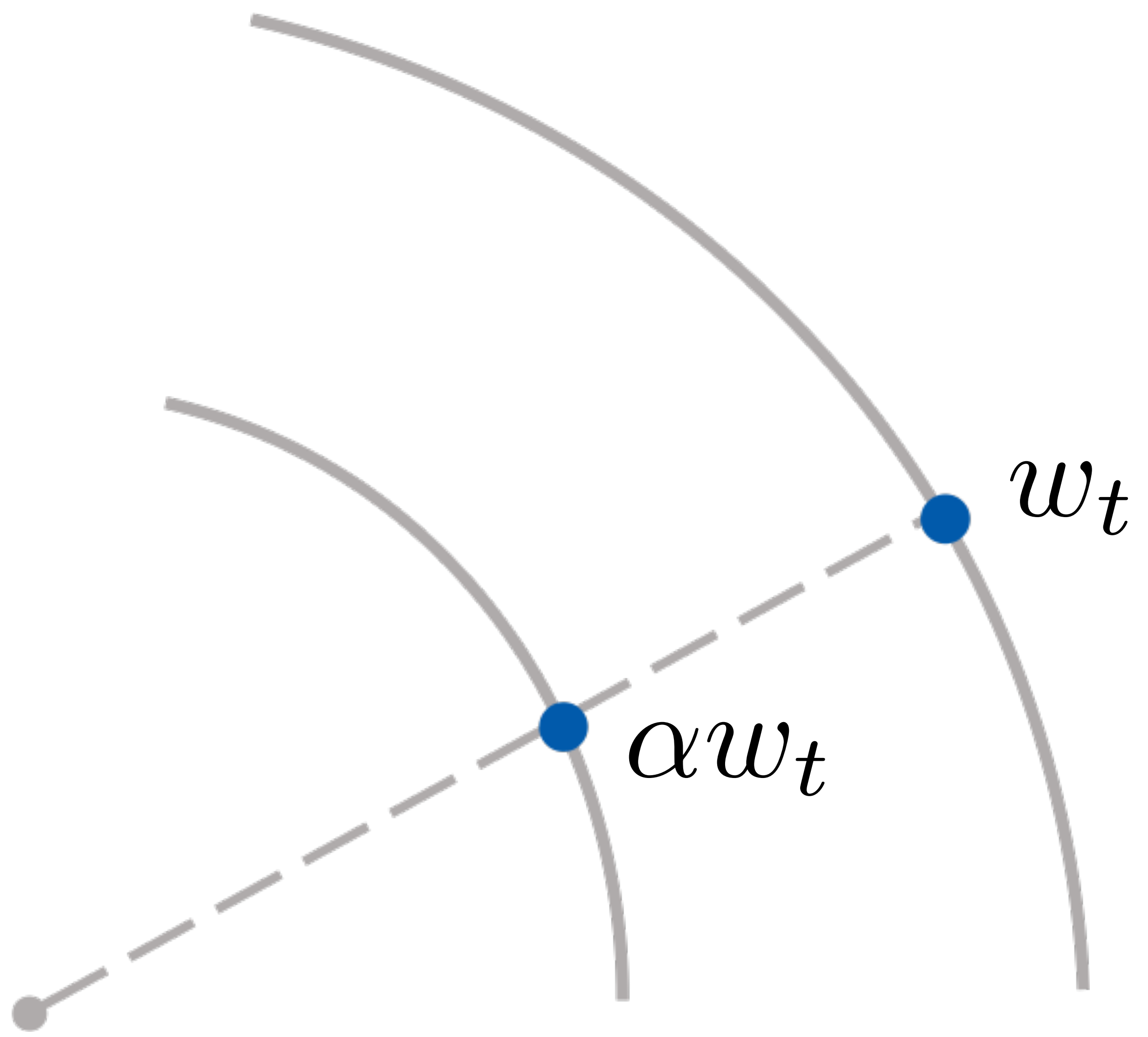
BatchNorm + Weight Decay

Training dynamics explained

BatchNorm + Weight Decay

↓

scale invariant weights



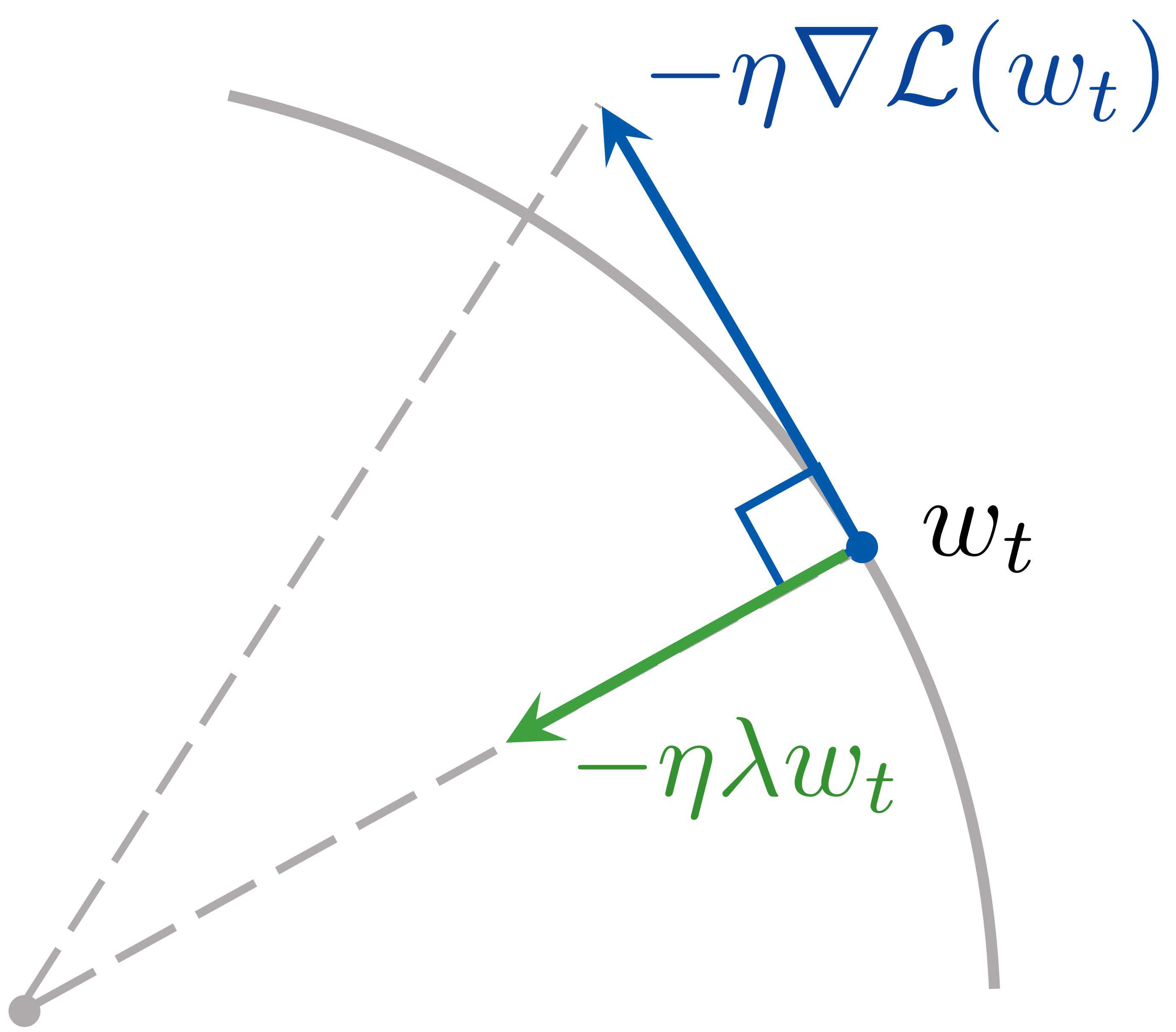
$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

Training dynamics explained

BatchNorm + Weight Decay

↓

scale invariant weights



$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

Training dynamics explained

BatchNorm

↓

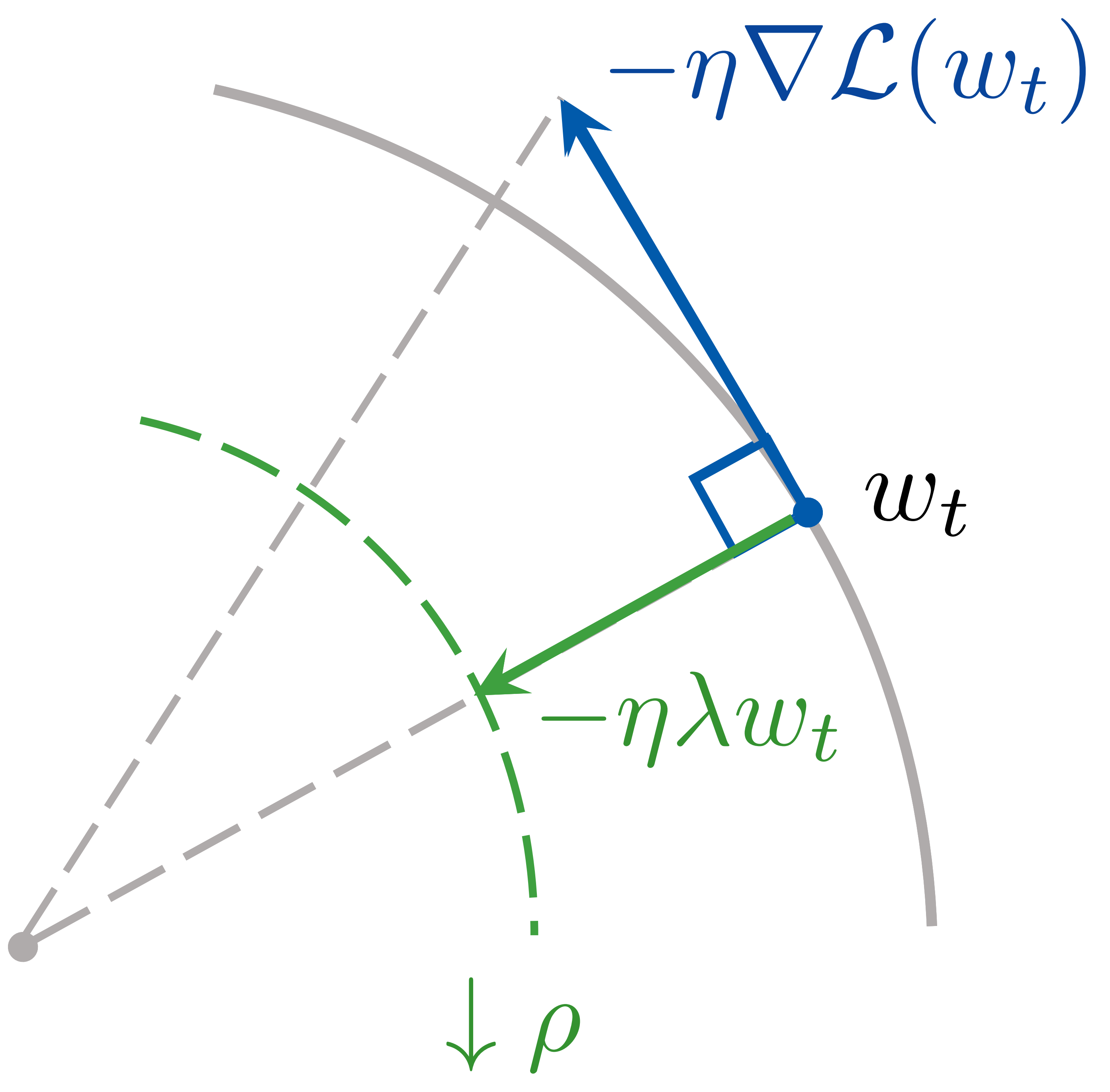
scale invariant weights

+

Weight Decay

↓

decreases weight norm ρ



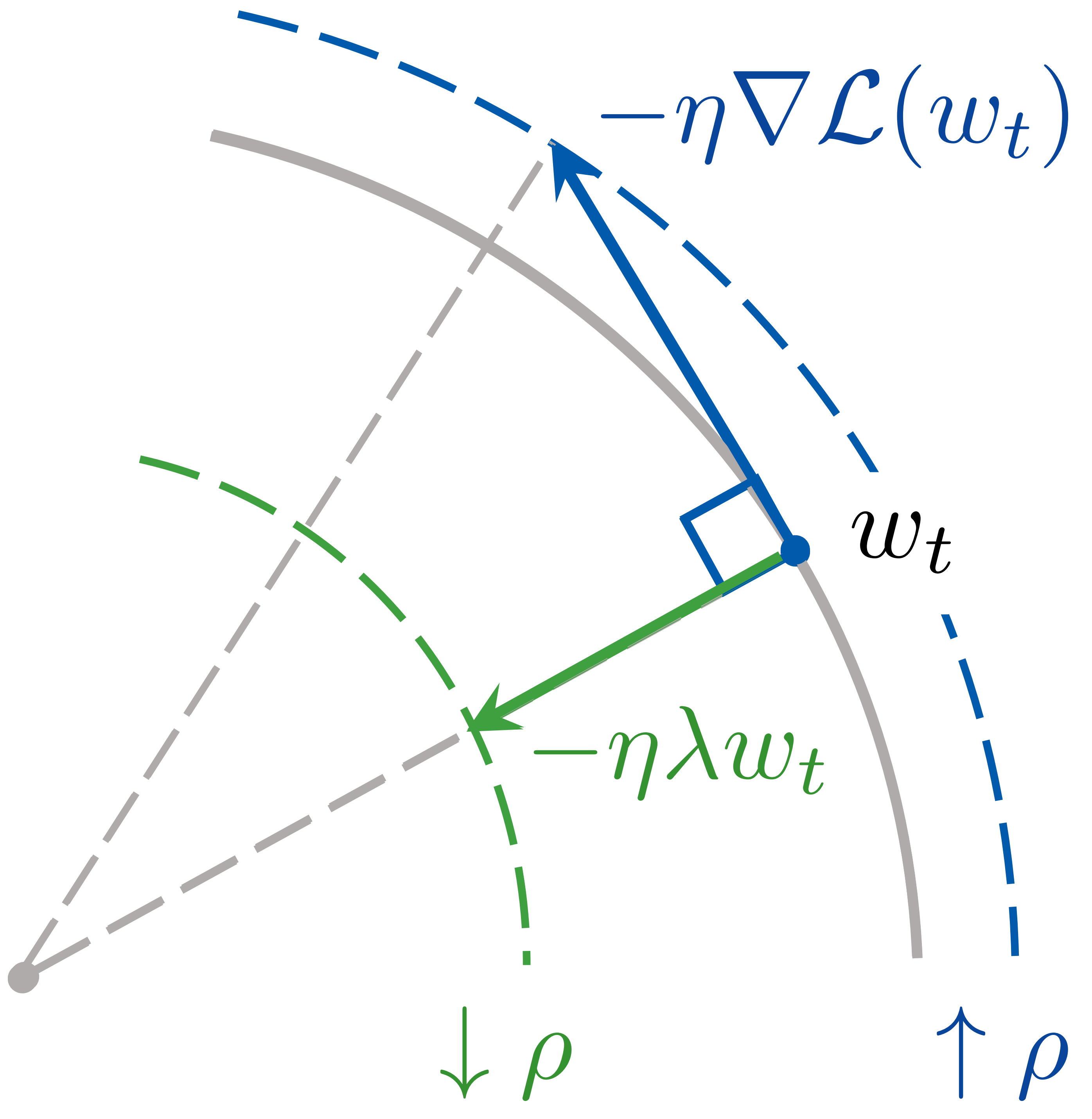
$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

Training dynamics explained

BatchNorm + Weight Decay

BatchNorm → scale invariant weights → increases weight norm ρ

Weight Decay → decreases weight norm ρ



$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

Training dynamics explained

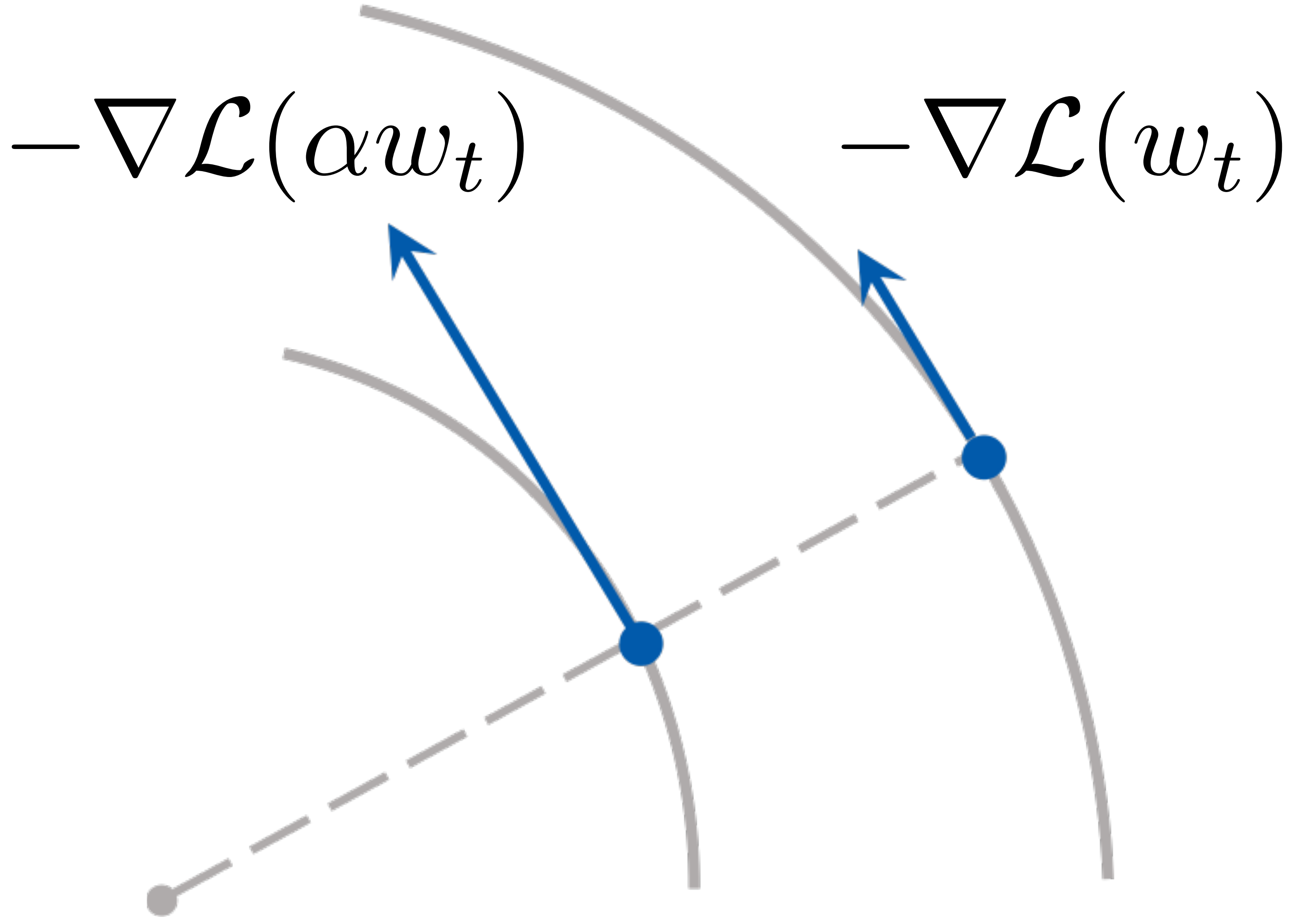
BatchNorm + Weight Decay

BatchNorm → scale invariant weights → increases weight norm ρ

Weight Decay → decreases weight norm ρ

changes optimization properties:

for lower weight norm steps are larger



$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

$$\nabla\mathcal{L}(\alpha w_t) = \frac{\nabla\mathcal{L}(w_t)}{\alpha}$$

Training dynamics explained

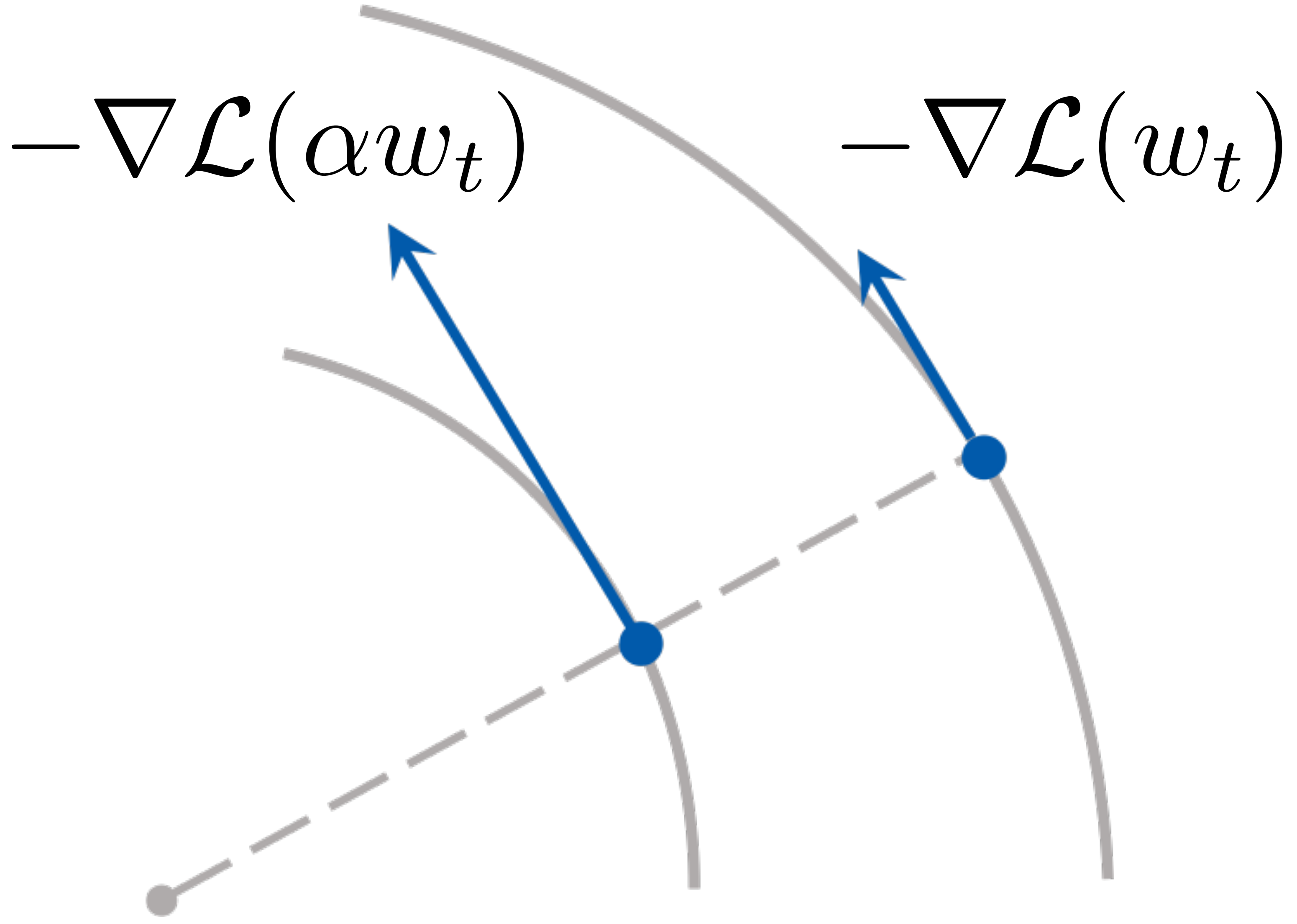
BatchNorm + Weight Decay

BatchNorm → scale invariant weights → increases weight norm ρ

Weight Decay → decreases weight norm ρ

→ changes optimization properties:
for lower weight norm steps are larger

→ optimization speed changes during training



$$\mathcal{L}(\alpha w_t) = \mathcal{L}(w_t)$$

$$\nabla\mathcal{L}(\alpha w_t) = \frac{\nabla\mathcal{L}(w_t)}{\alpha}$$

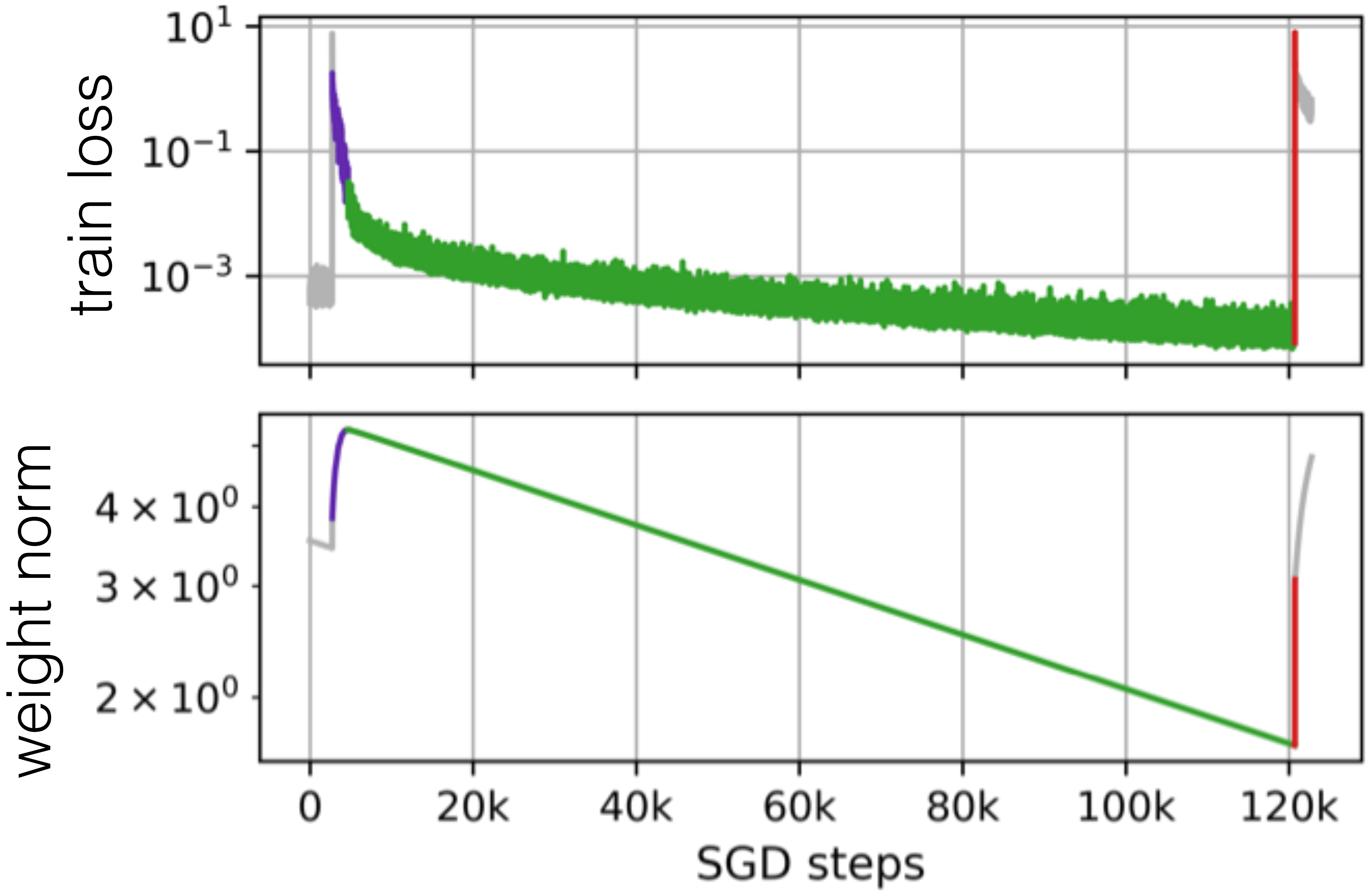
Training dynamics explained

Gradient update of the weights:

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

scale-invariant loss   weight decay

One training period



Training dynamics explained

Gradient update of the weights:

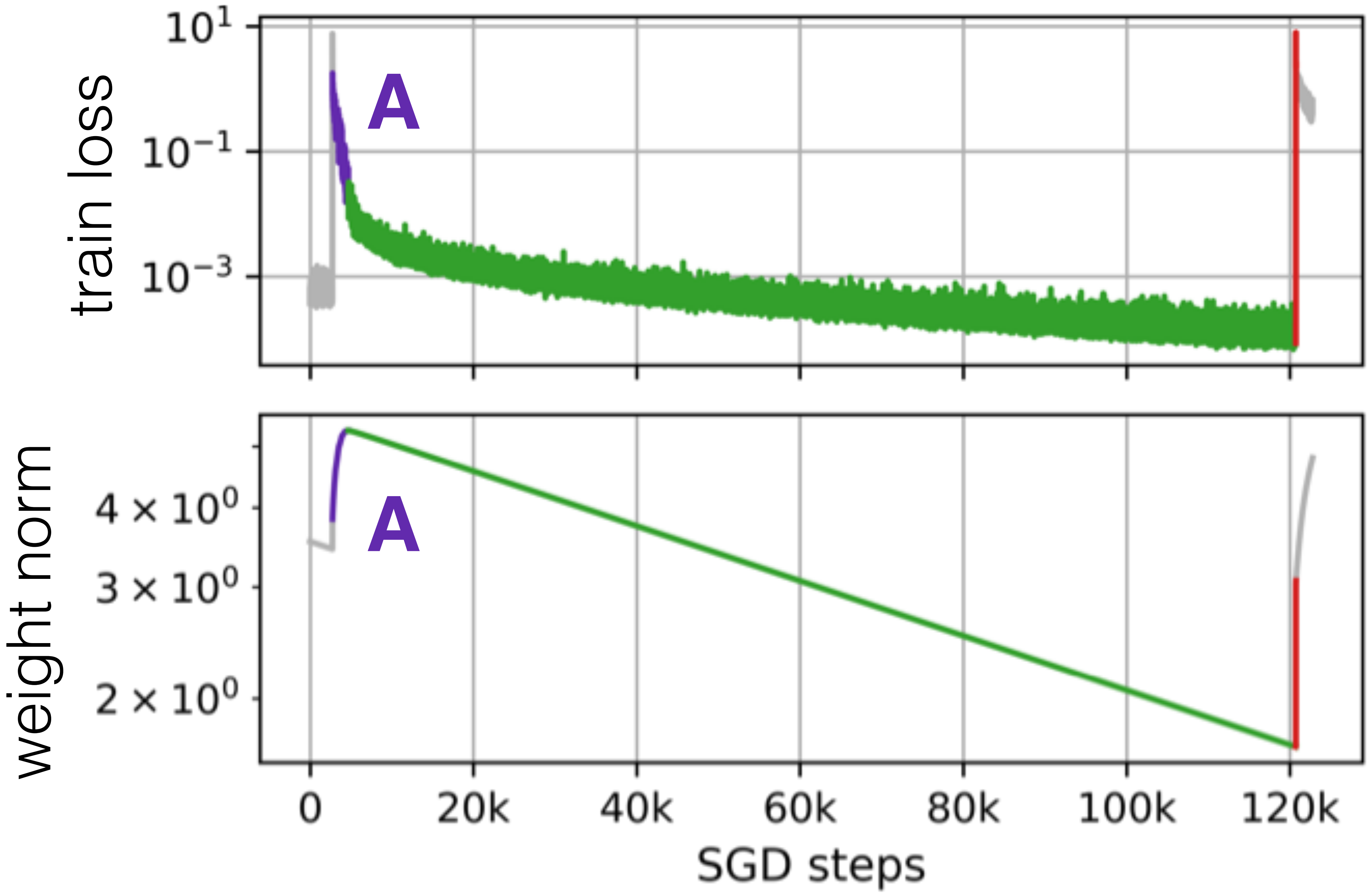
$$w_{t+1} = w_t - \eta \nabla \mathcal{L}(w_t) - \eta \lambda w_t$$

scale-invariant loss   weight decay

A: loss component is stronger

→ weight norm increase

One training period



Training dynamics explained

Gradient update of the weights:

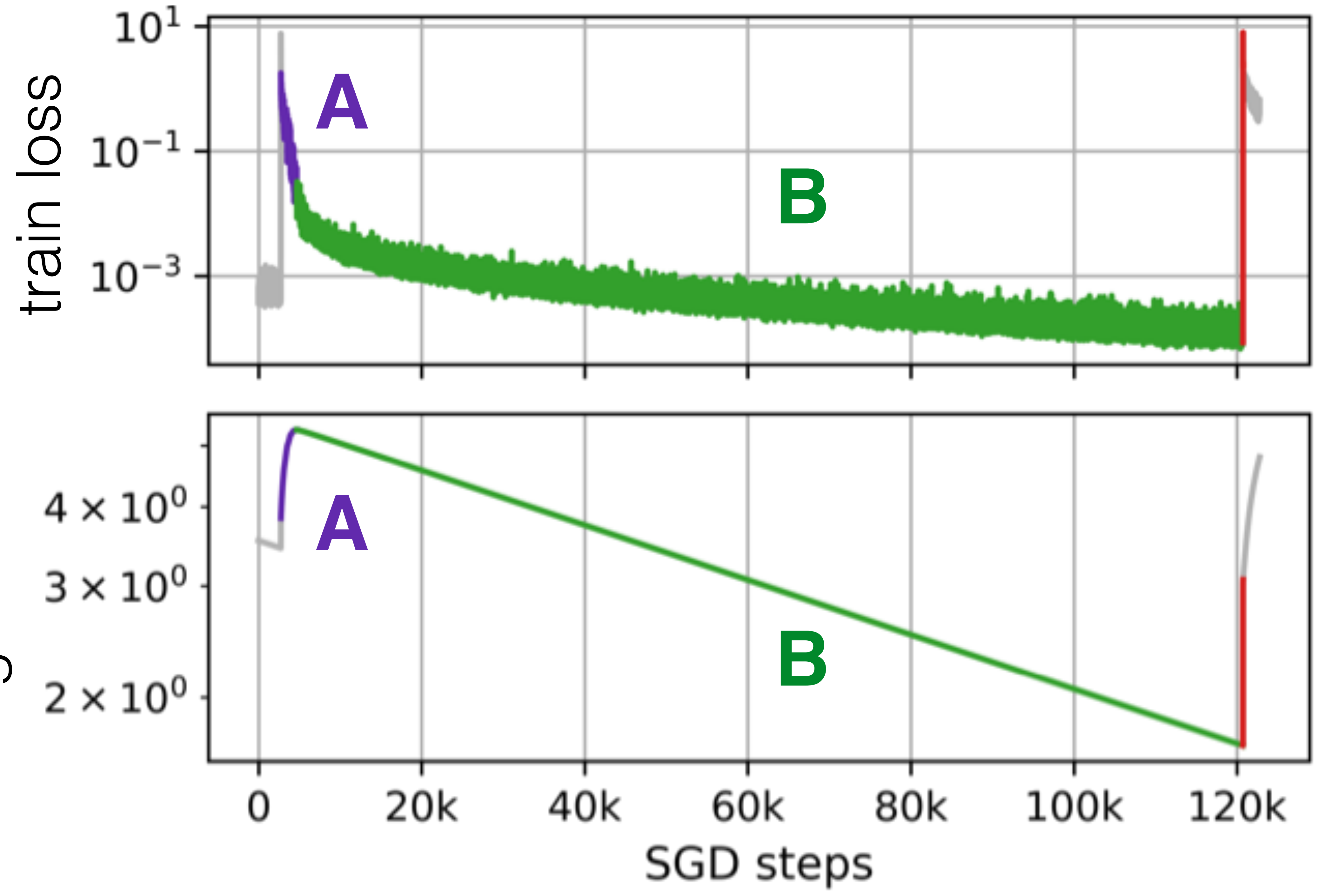
$$w_{t+1} = w_t - \eta \nabla \mathcal{L}(w_t) - \eta \lambda w_t$$

scale-invariant loss   weight decay

A: loss component is stronger
→ weight norm increase

B: weight decay component is stronger
→ weight norm decrease

One training period

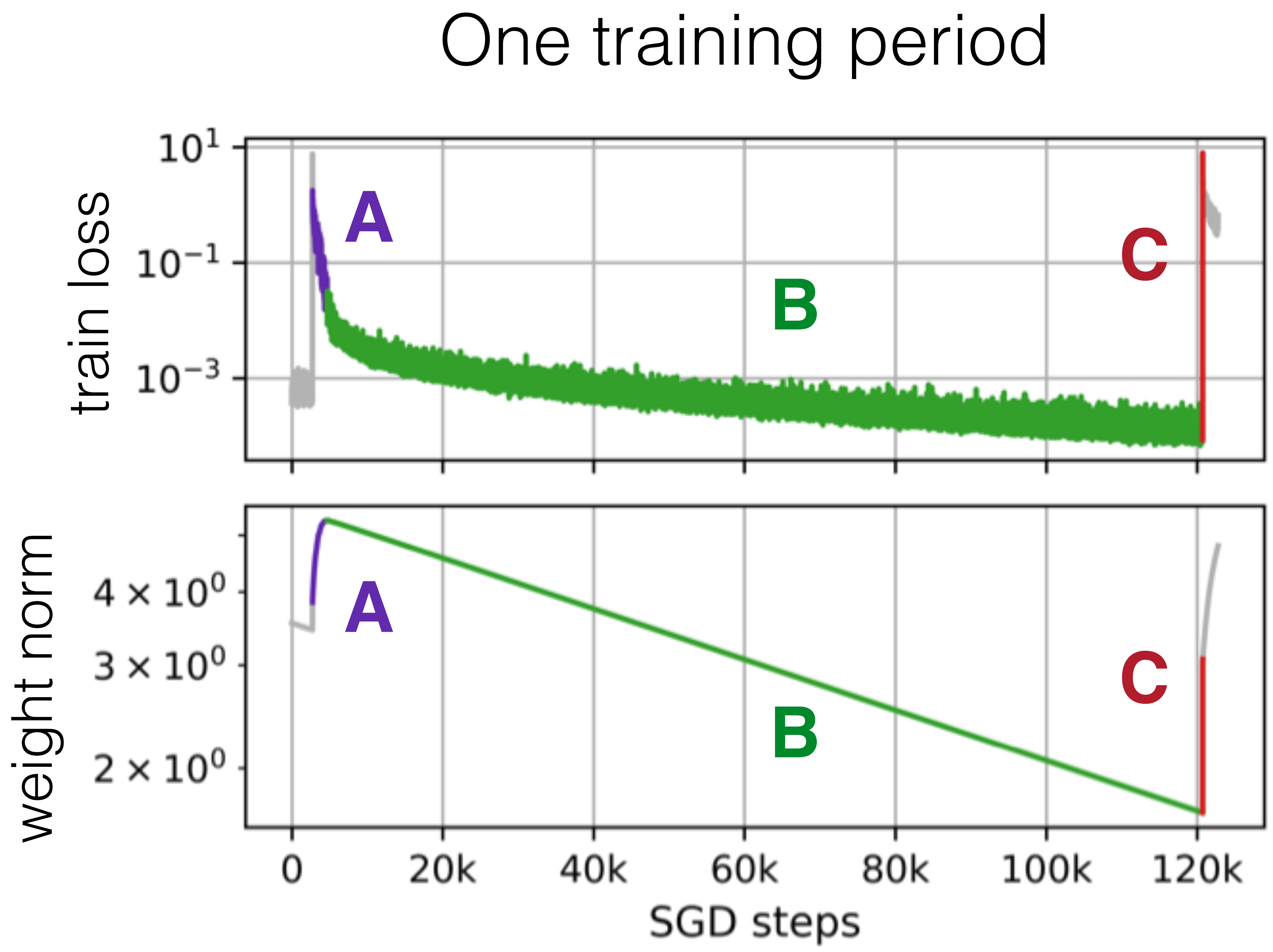


Training dynamics explained

Gradient update of the weights:

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

scale-invariant loss   weight decay



A: loss component is stronger
→ weight norm increase

B: weight decay component is stronger
→ weight norm decrease

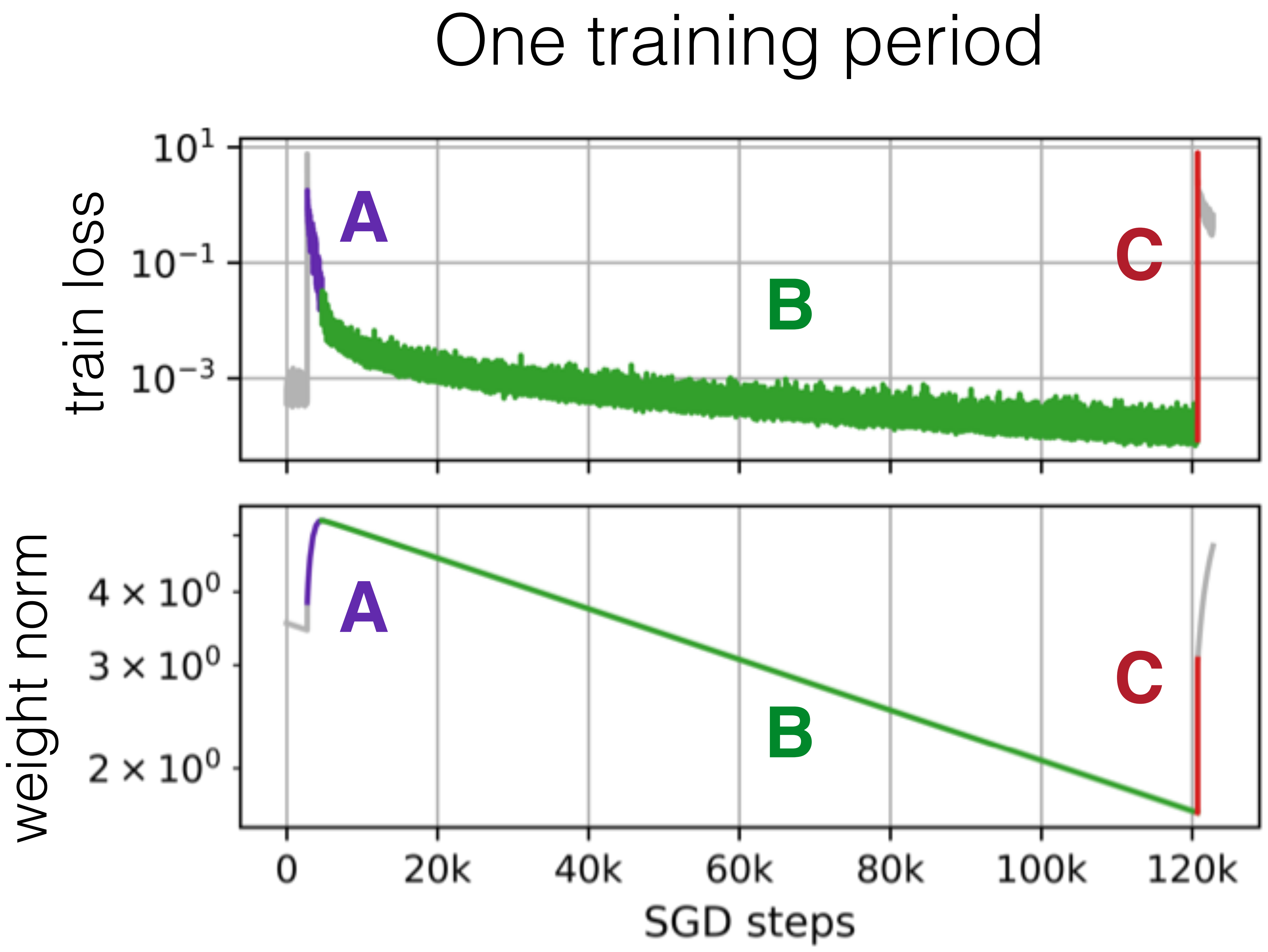
C: low weight norm → divergence

Training dynamics explained

Gradient update of the weights:

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

scale-invariant loss   weight decay



A: loss component is stronger
→ weight norm increase

B: weight decay component is stronger
→ weight norm decrease

C: low weight norm → divergence → high weight norm → new period

Empirical justification

We want to verify:

BatchNorm and Weight Decay influence on the weight norm causes periodic behavior

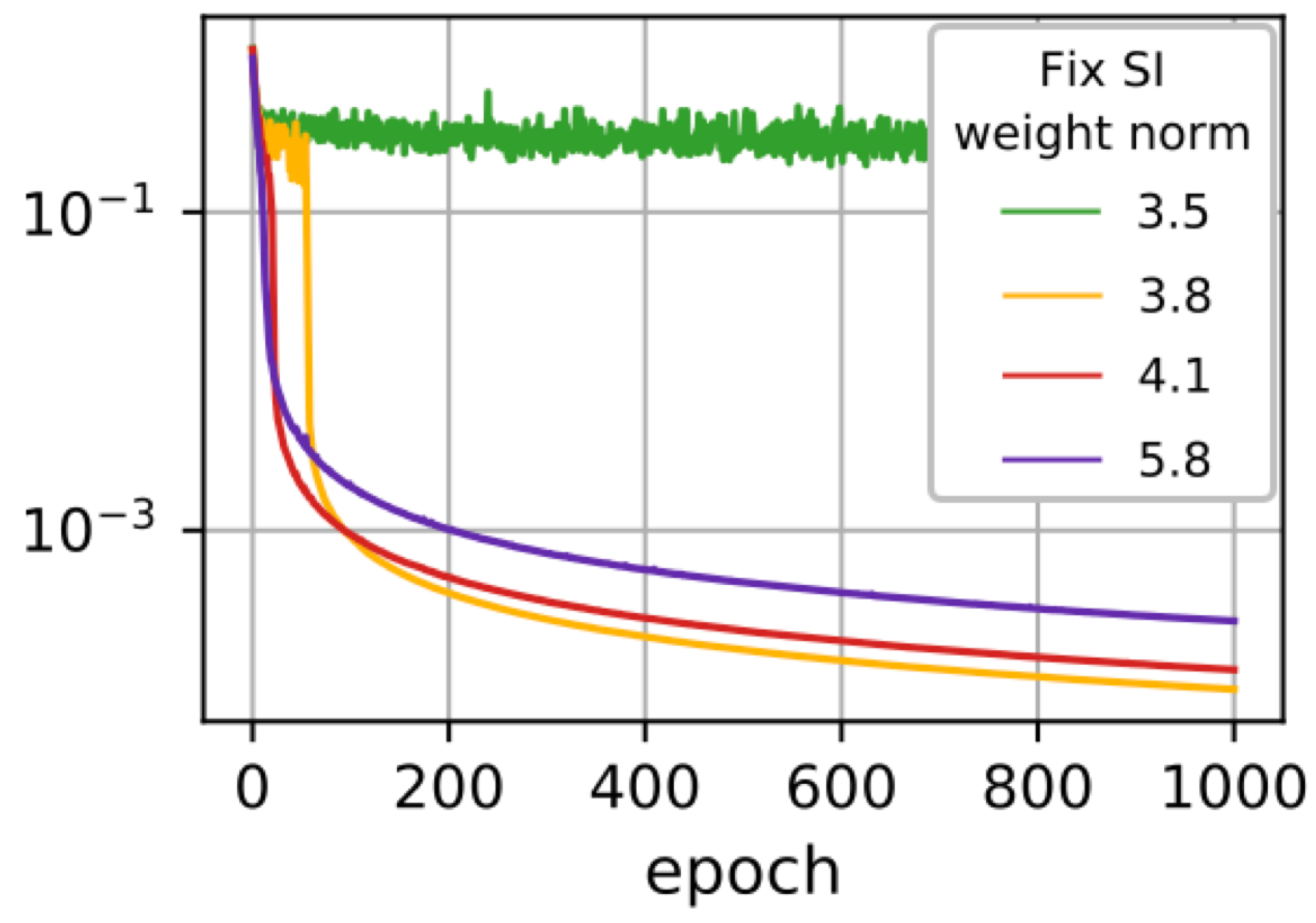
Experiment setting:

To prohibit this influence we fix the weight norm during training

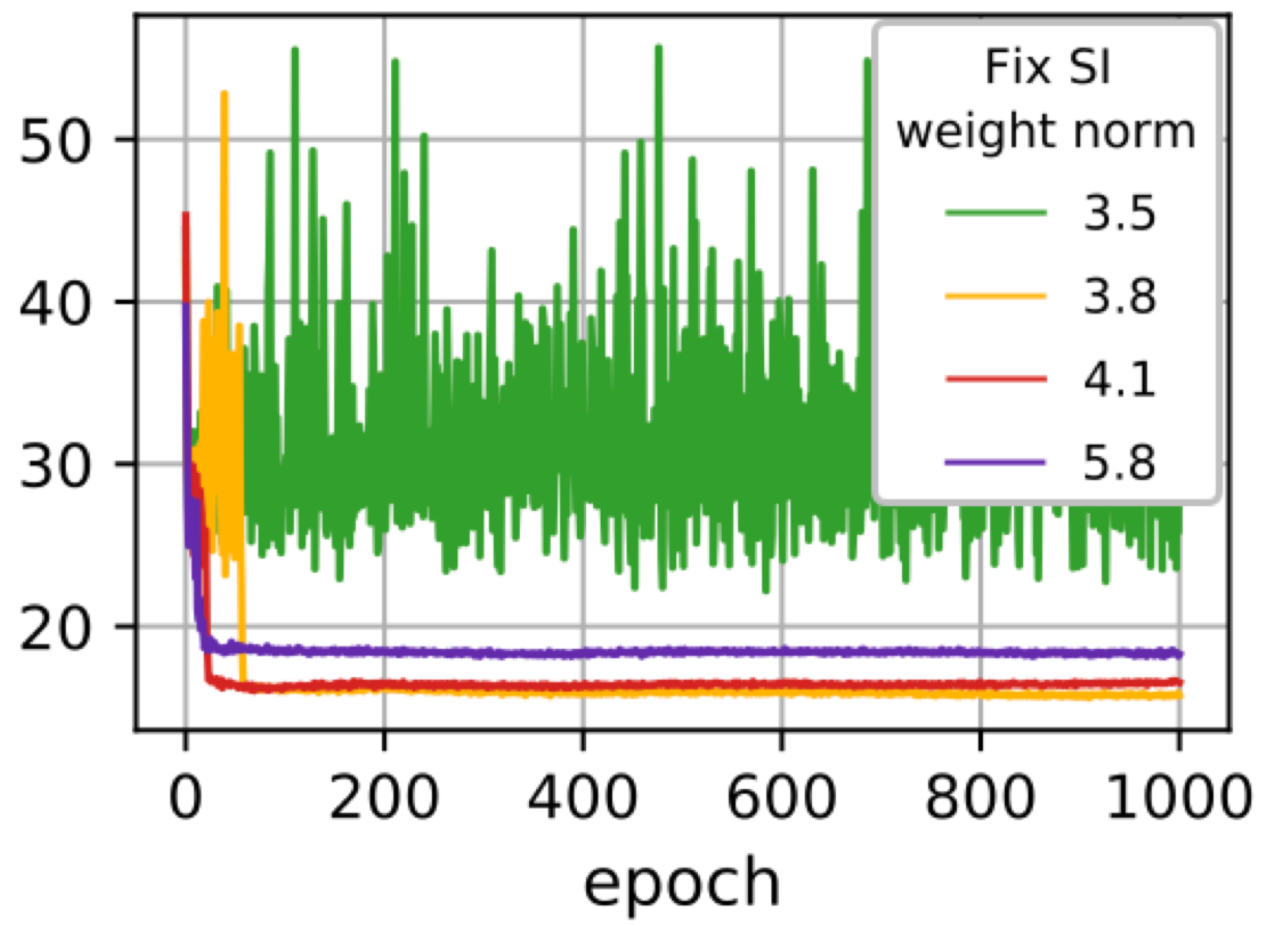
Result:

No periodic behavior ⟶ the weight norm change is the key!

Train loss



Test error, %



Theoretical justification

Conditions for destabilization:

At what weight norm it is possible / guaranteed

Theoretical justification

Conditions for destabilization:

At what weight norm it is possible / guaranteed

Periods frequency dependency on the hyperparameters:

Periods frequency $\propto$ learning rate $\times$ weight decay

Theoretical justification

Conditions for destabilization:

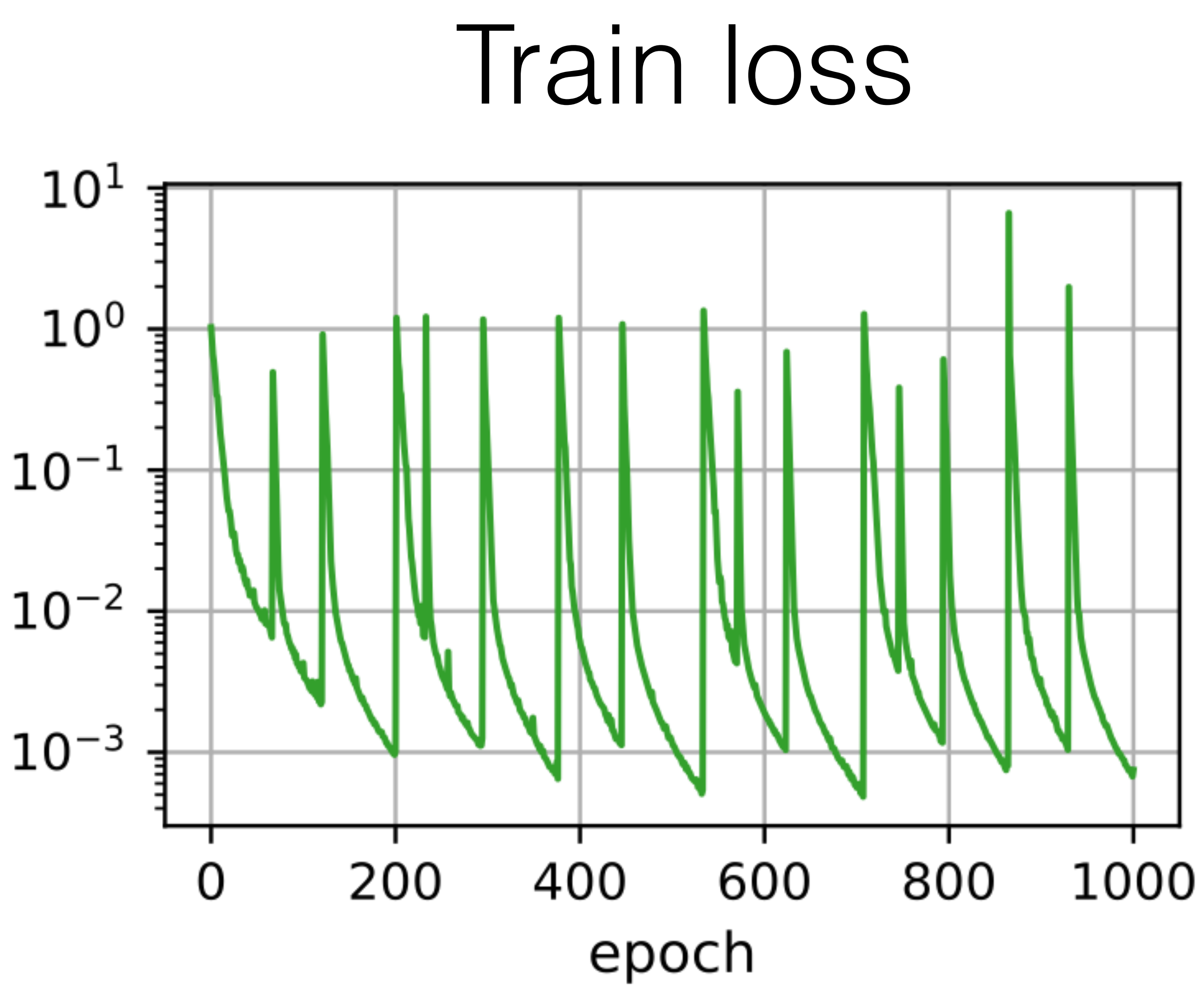
At what weight norm it is possible / guaranteed

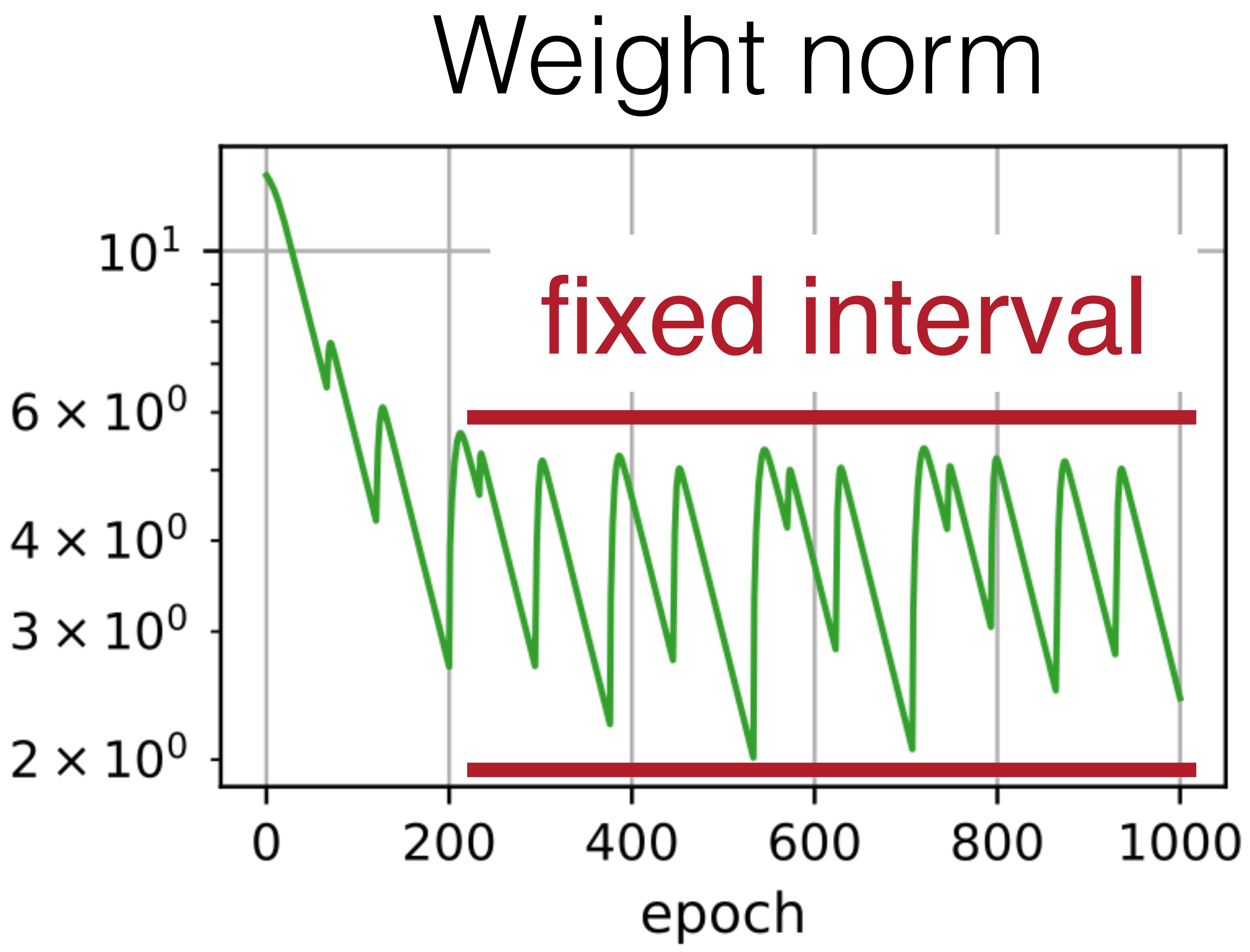
Periods frequency dependency on the hyperparameters:

Periods frequency $\propto$ learning rate $\times$ weight decay

Generalization of the equilibrium:

Training dynamics converge to a stable periodic behavior





Empirical study

Architectures:

3-layer ConvNet, ResNet-18

Datasets:

CIFAR-10, CIFAR-100

Later on the slides:

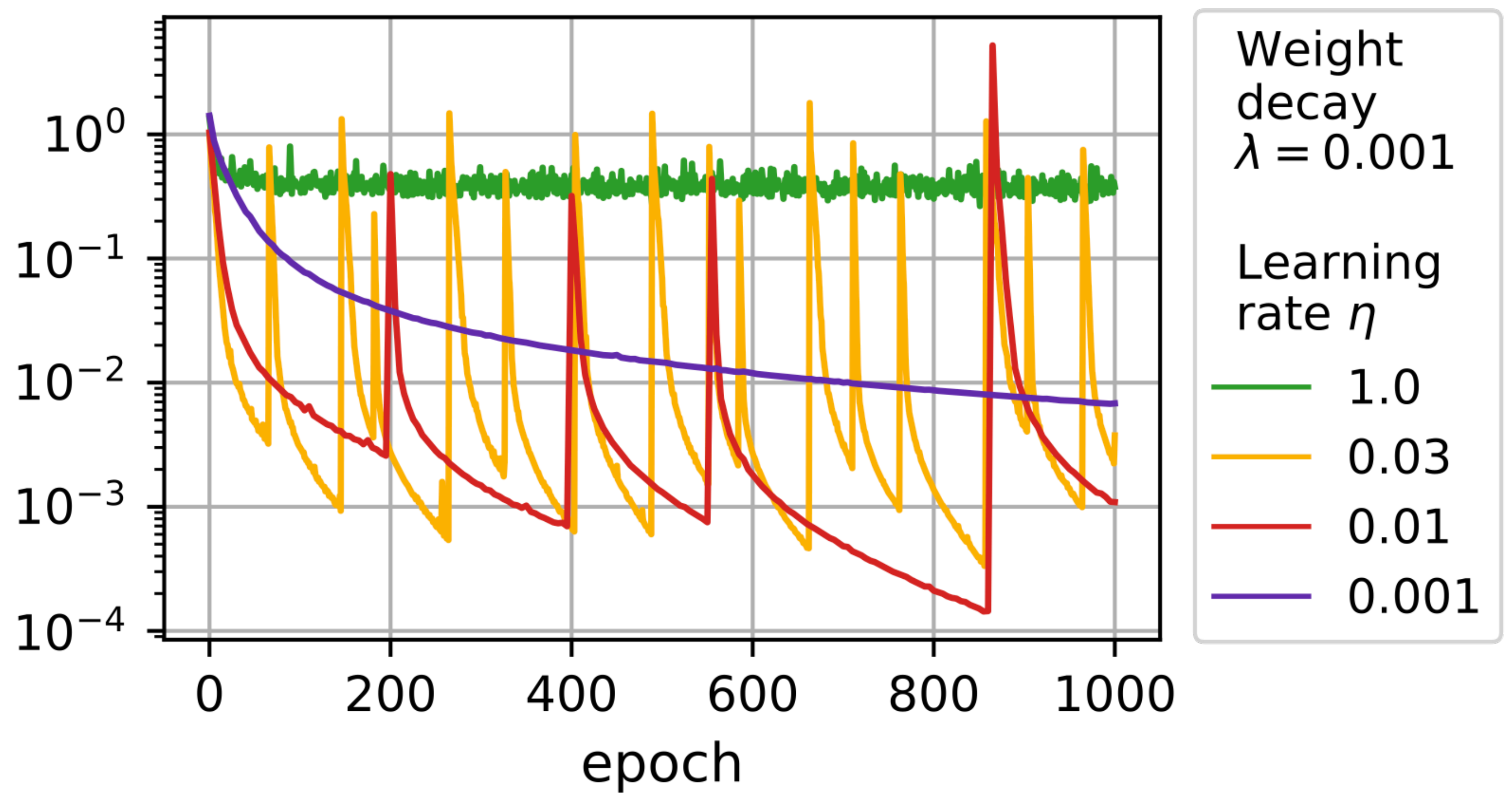
ConvNet on CIFAR-10

Empirical study - hyperparameters

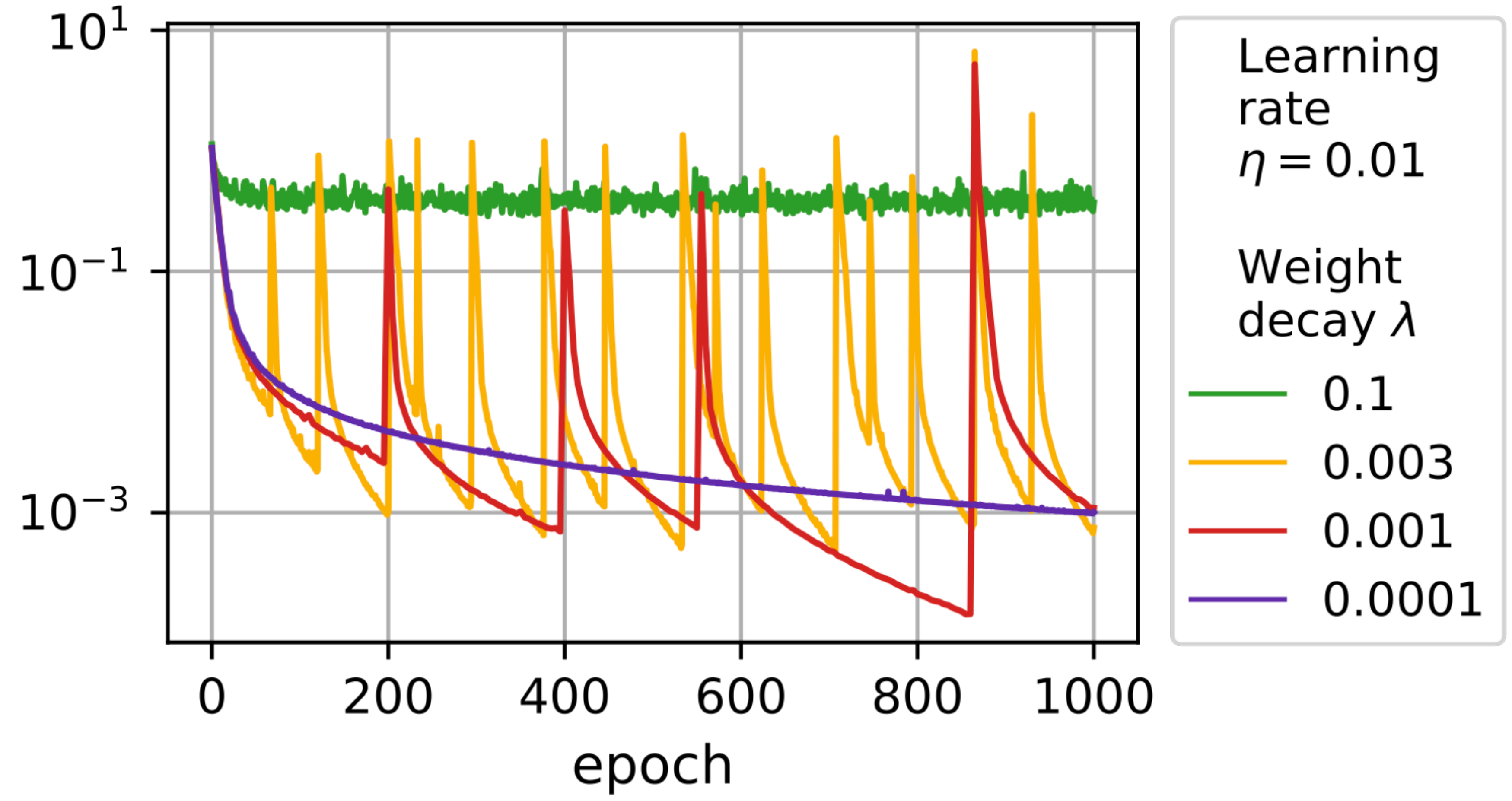
Simplified setting:

- Fully scale-invariant networks
- SGD
- No learning rate schedule
- No data augmentation

Vary learning rate



Vary weight decay



Empirical study - hyperparameters

Simplified setting:

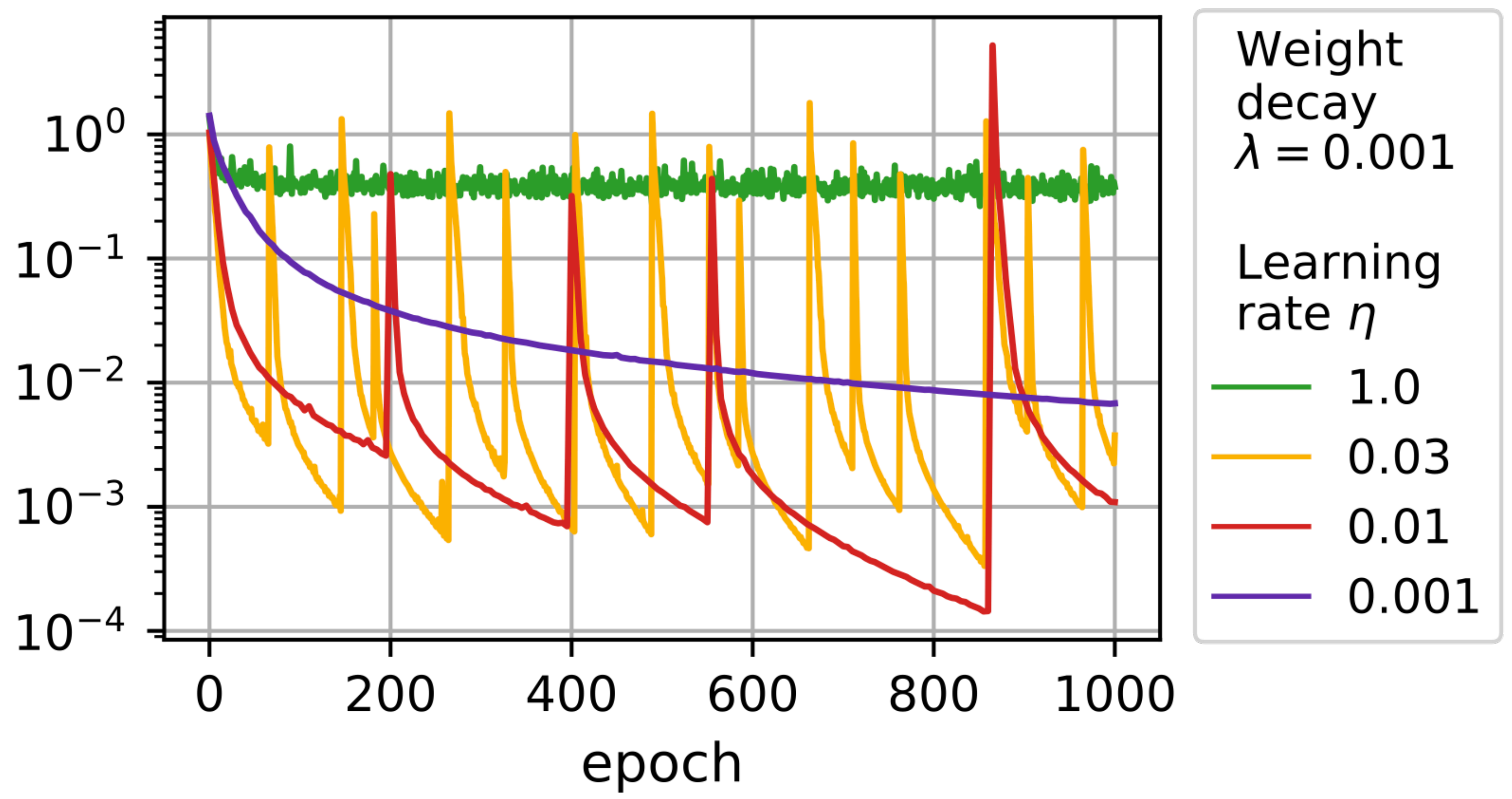
- Fully scale-invariant networks
- SGD
- No learning rate schedule
- No data augmentation

Periods for a wide range of hyperparameters

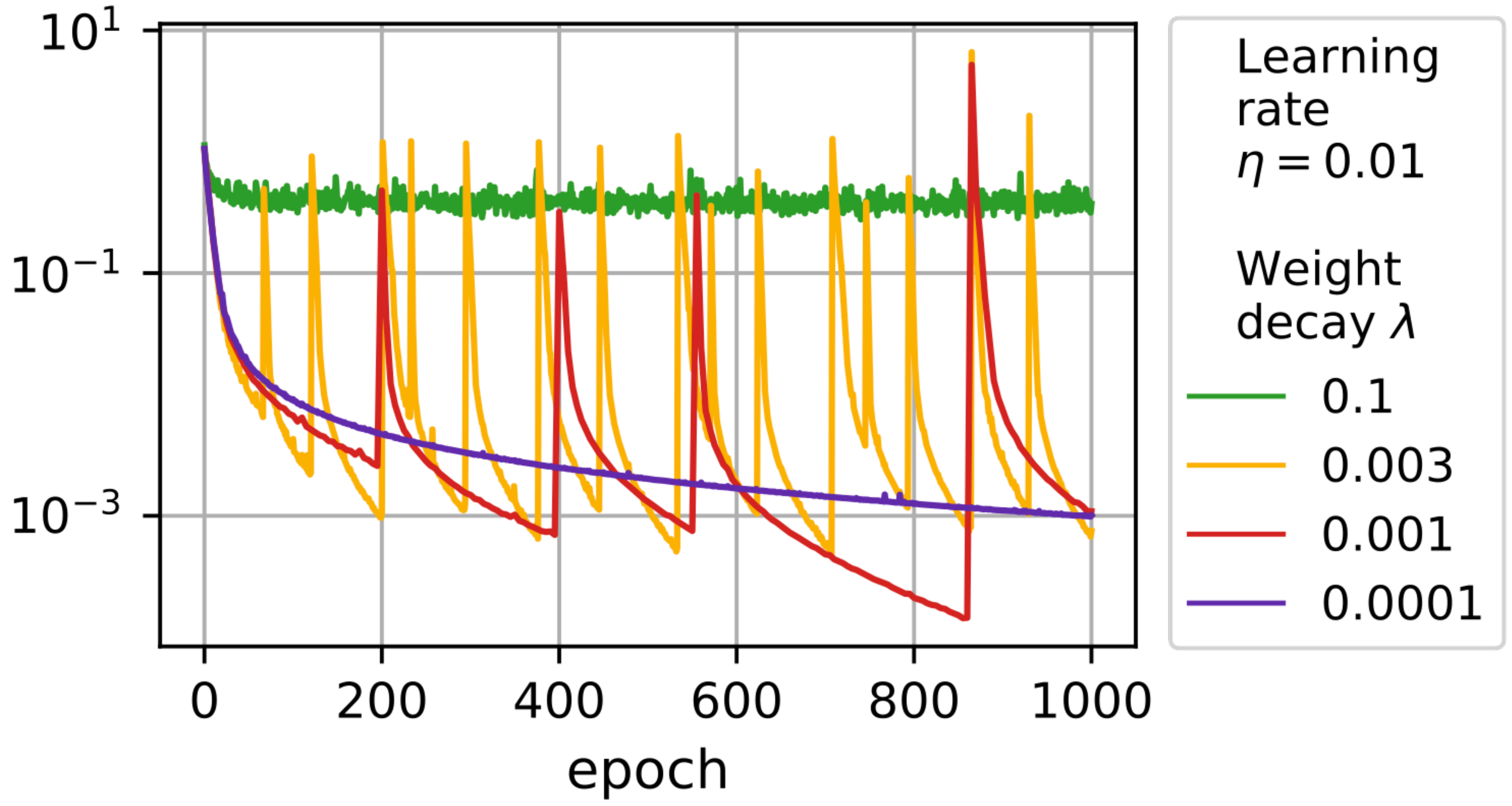
Low values ⟶ too slow training

High values ⟶ unstable training

Vary learning rate



Vary weight decay



Empirical study - hyperparameters

Simplified setting:

- Fully scale-invariant networks
- SGD
- No learning rate schedule
- No data augmentation

Periods for a wide range of hyperparameters

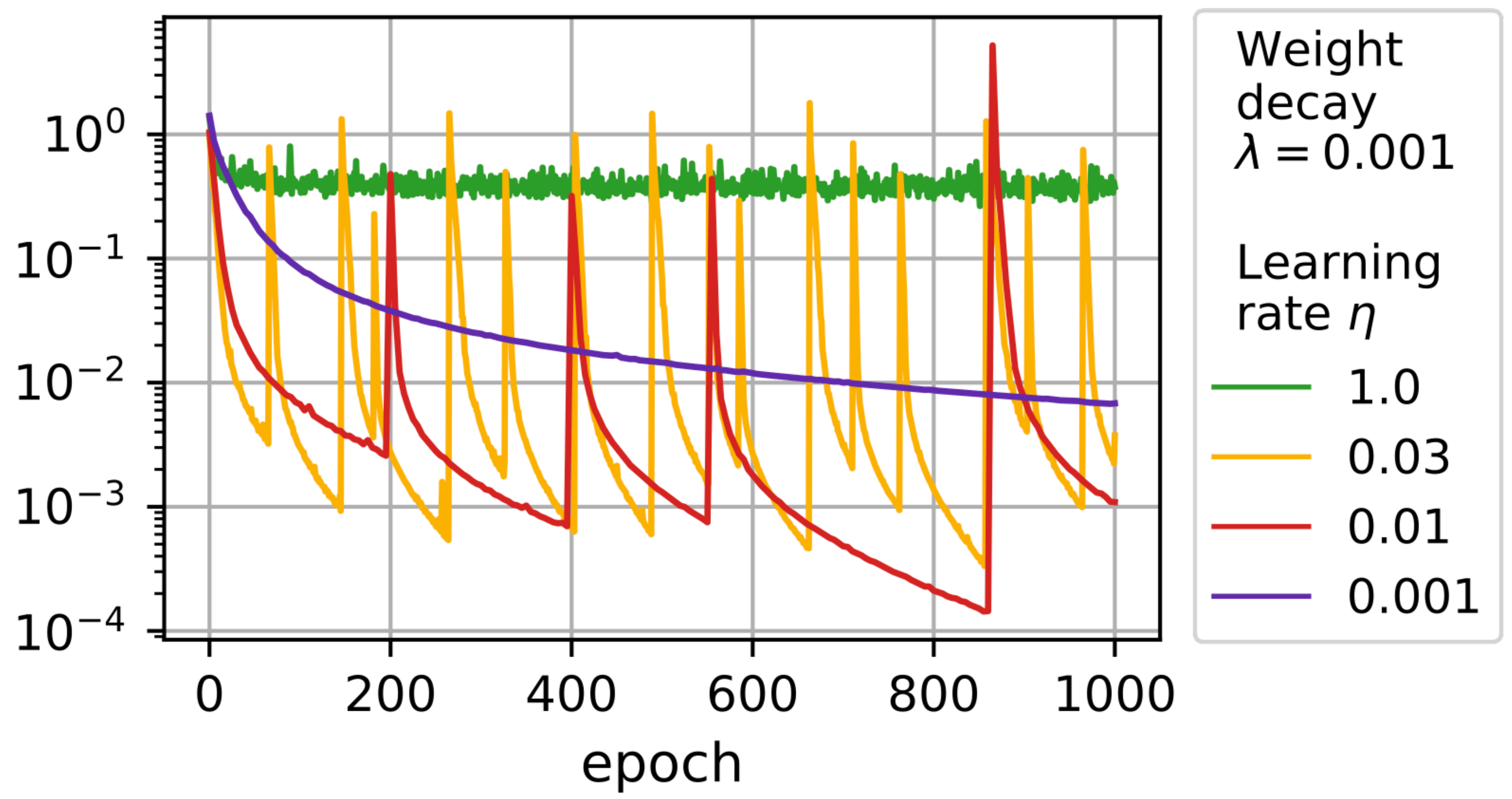
Low values ⟶ too slow training

High values ⟶ unstable training

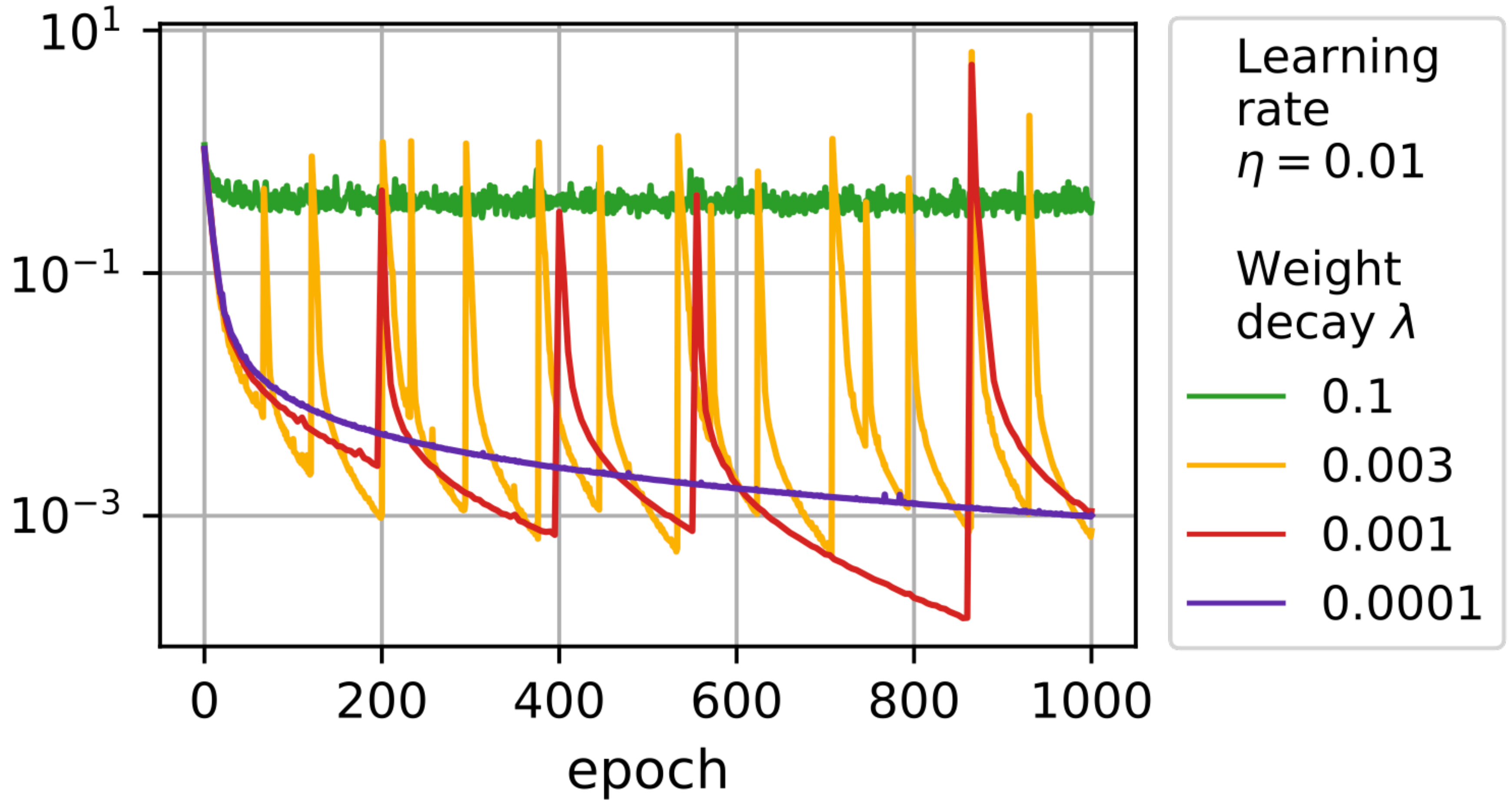
Empirical results agree with theoretical expectations:

Periods frequency $\propto$ learning rate $\times$ weight decay

Vary learning rate



Vary weight decay

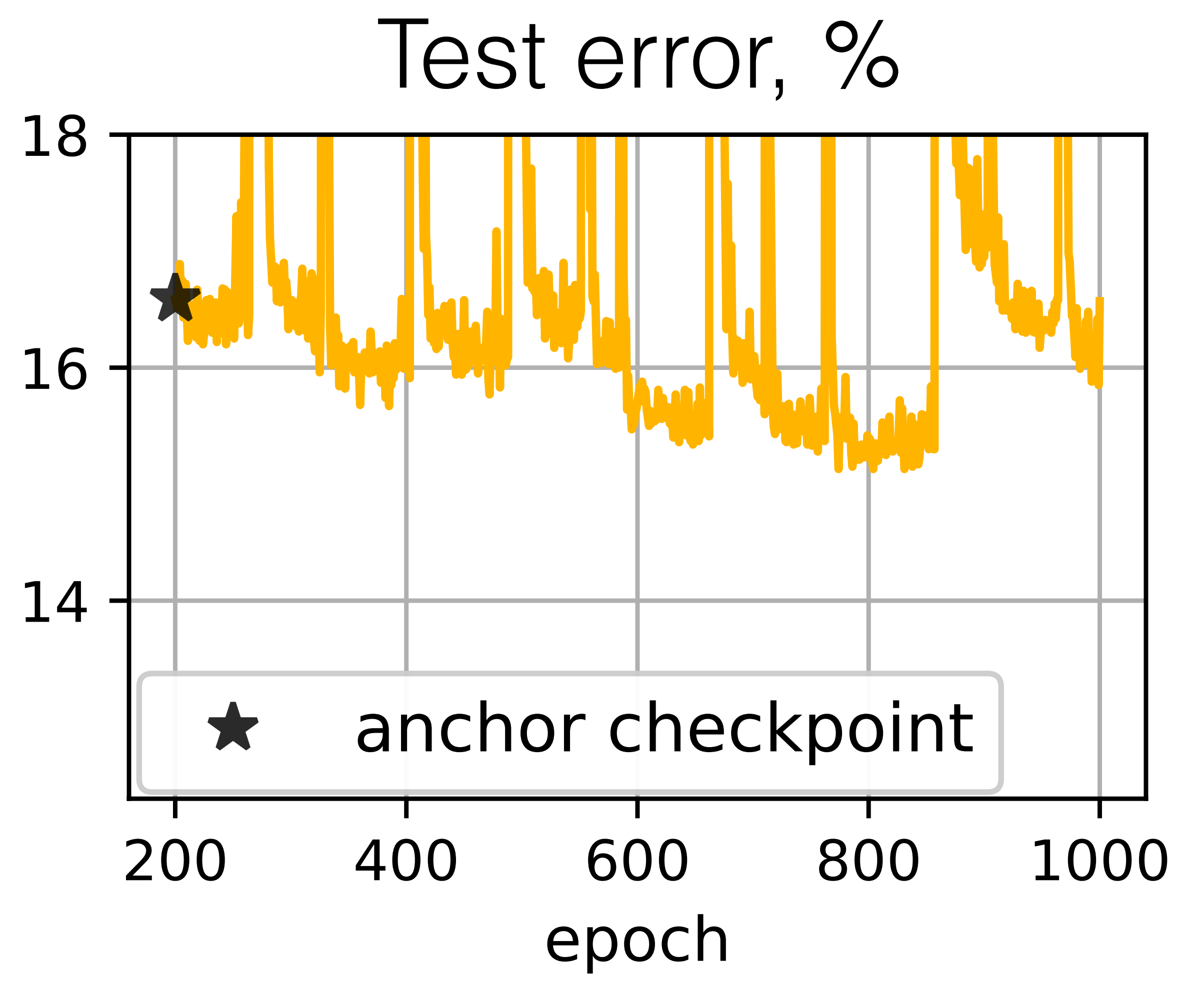


Empirical study - diverse minima

one experiment

anchor checkpoint

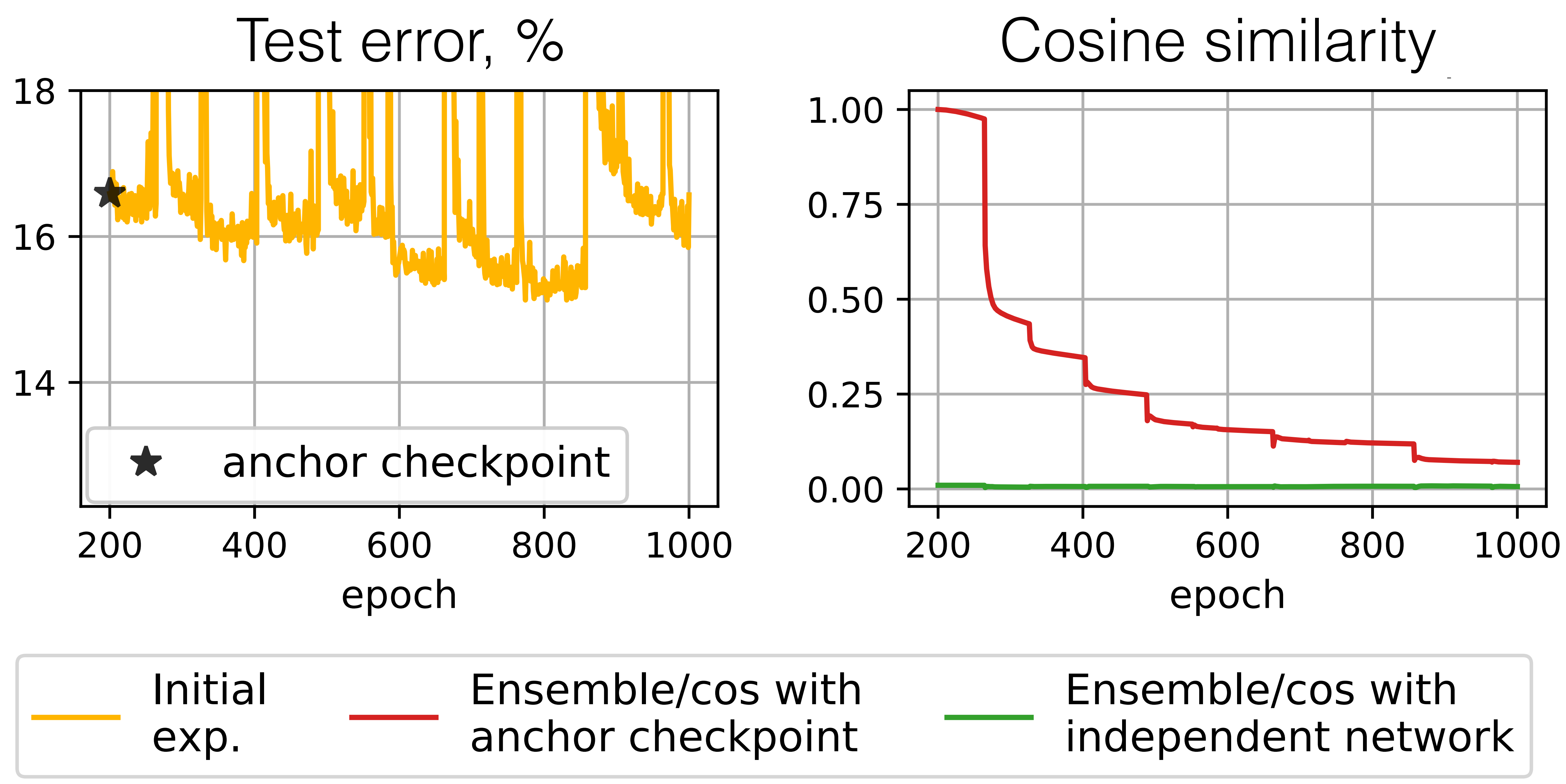
subsequent checkpoints



Empirical study - diverse minima

one experiment

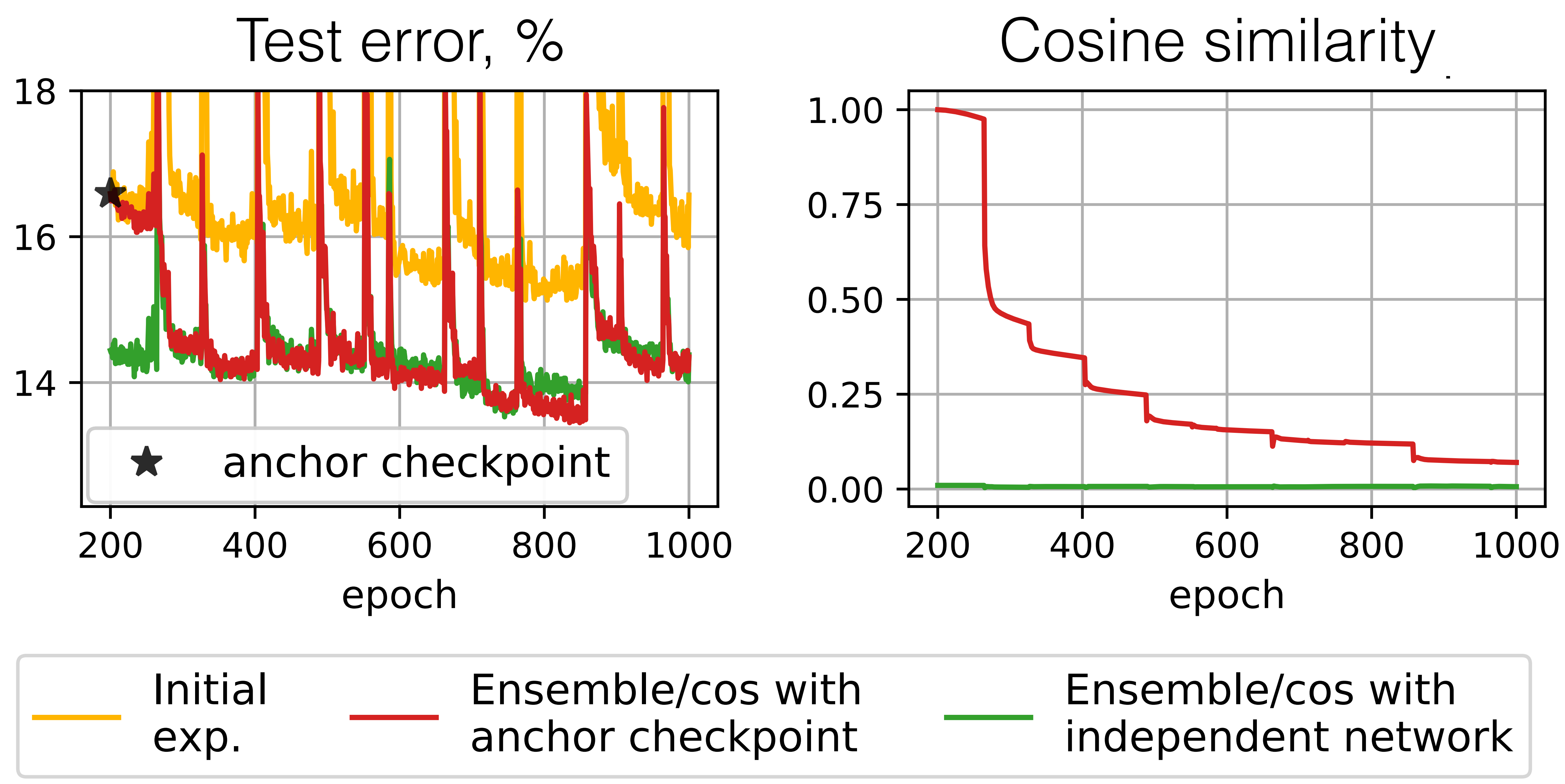
anchor checkpoint ⟷ subsequent checkpoints ⟷ independently trained network



Empirical study - diverse minima

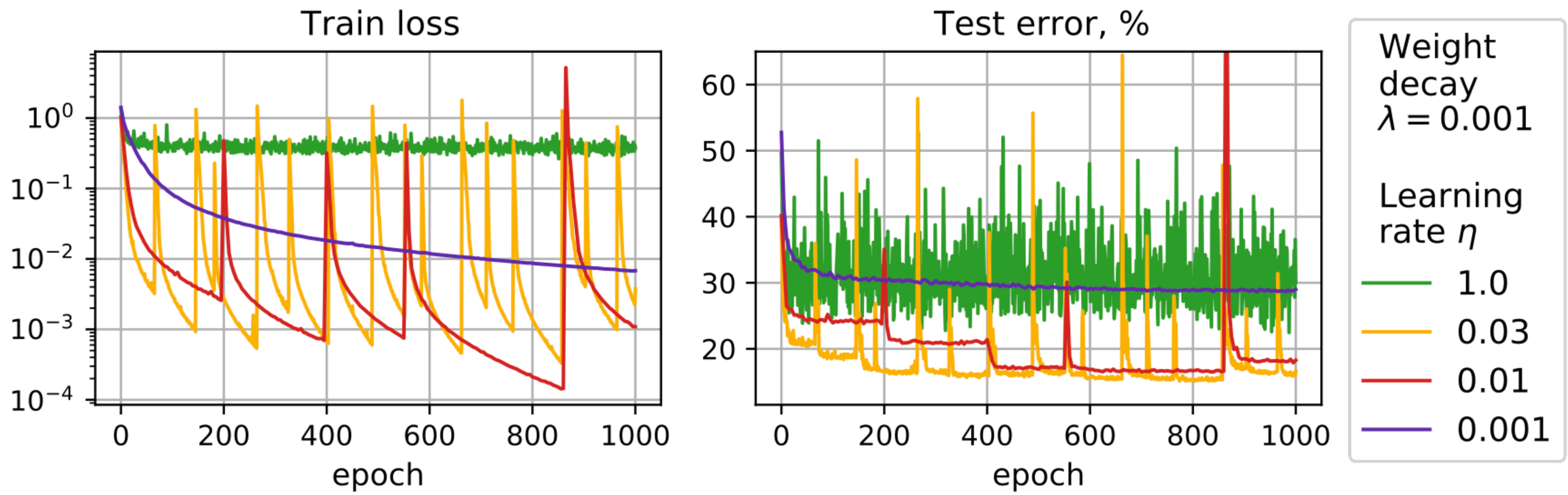
one experiment

anchor checkpoint ⟷ subsequent checkpoints ⟷ independently trained network



Empirical study - diverse minima

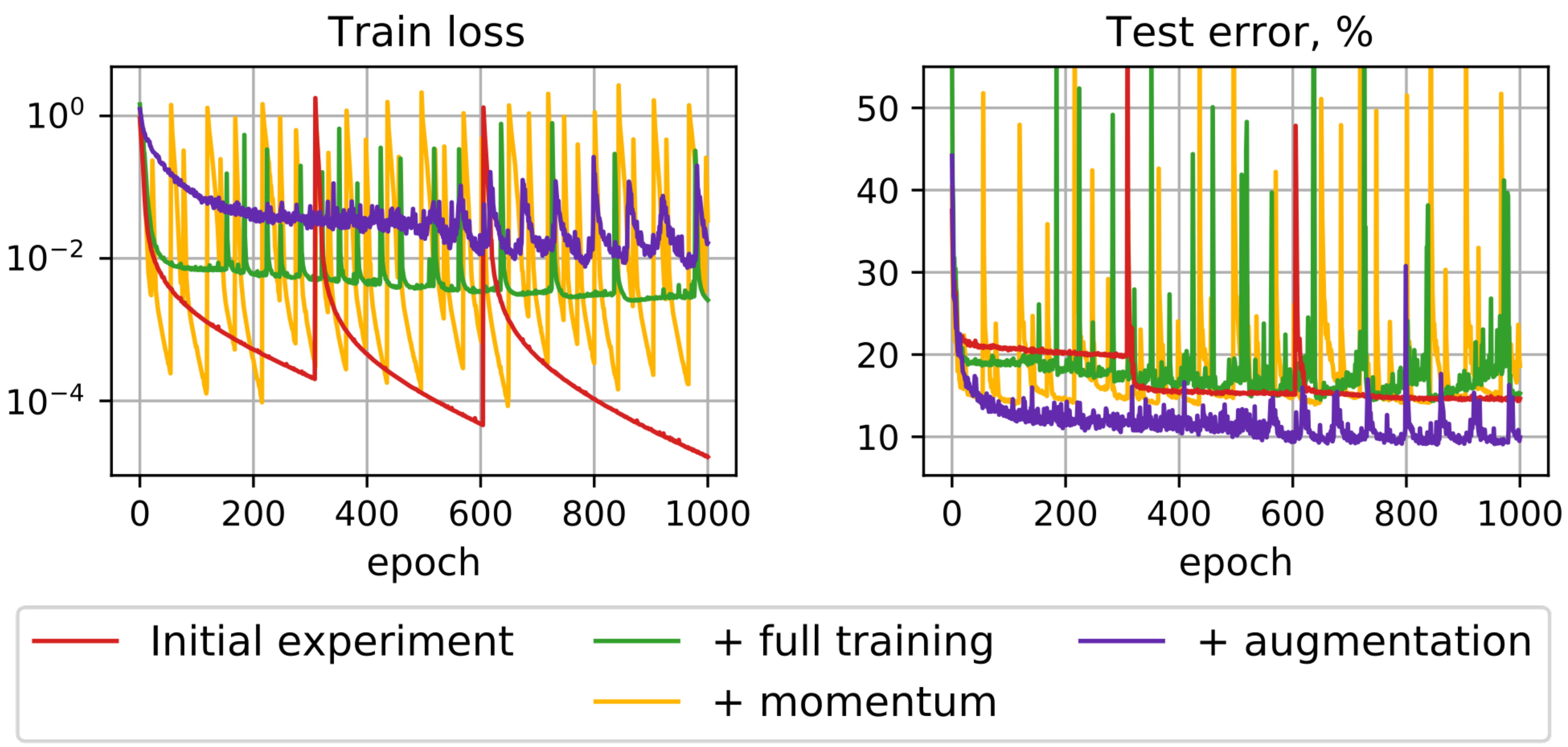
At the beginning of training, minima usually improve with each new period:



Empirical study - practical setting

Simplified setting:

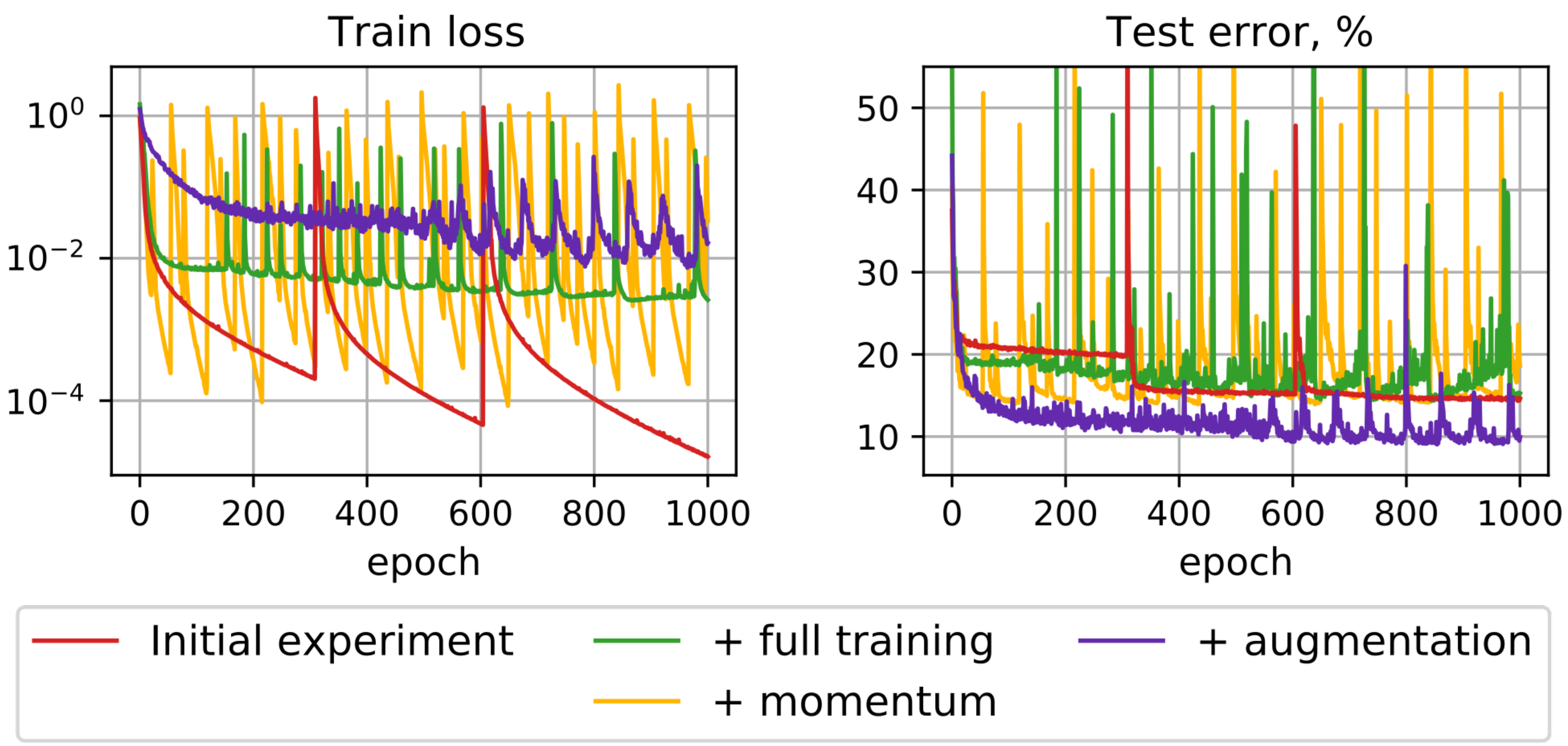
- Fully scale-invariant networks
- SGD
- No data augmentation



Empirical study - practical setting

Simplified setting:

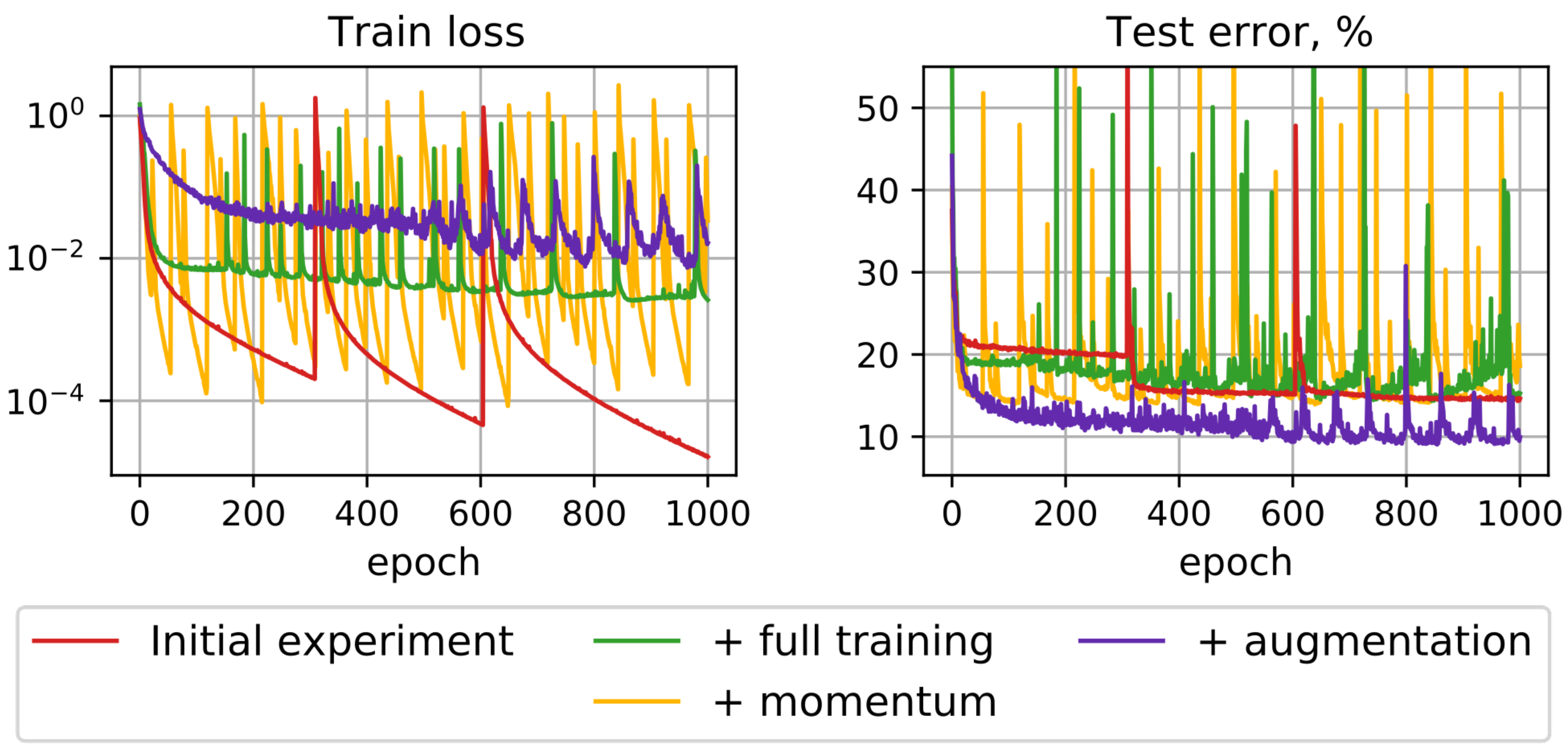
- ~~Fully scale-invariant networks~~ → Standard networks
- SGD
- No data augmentation



Empirical study - practical setting

Simplified setting:

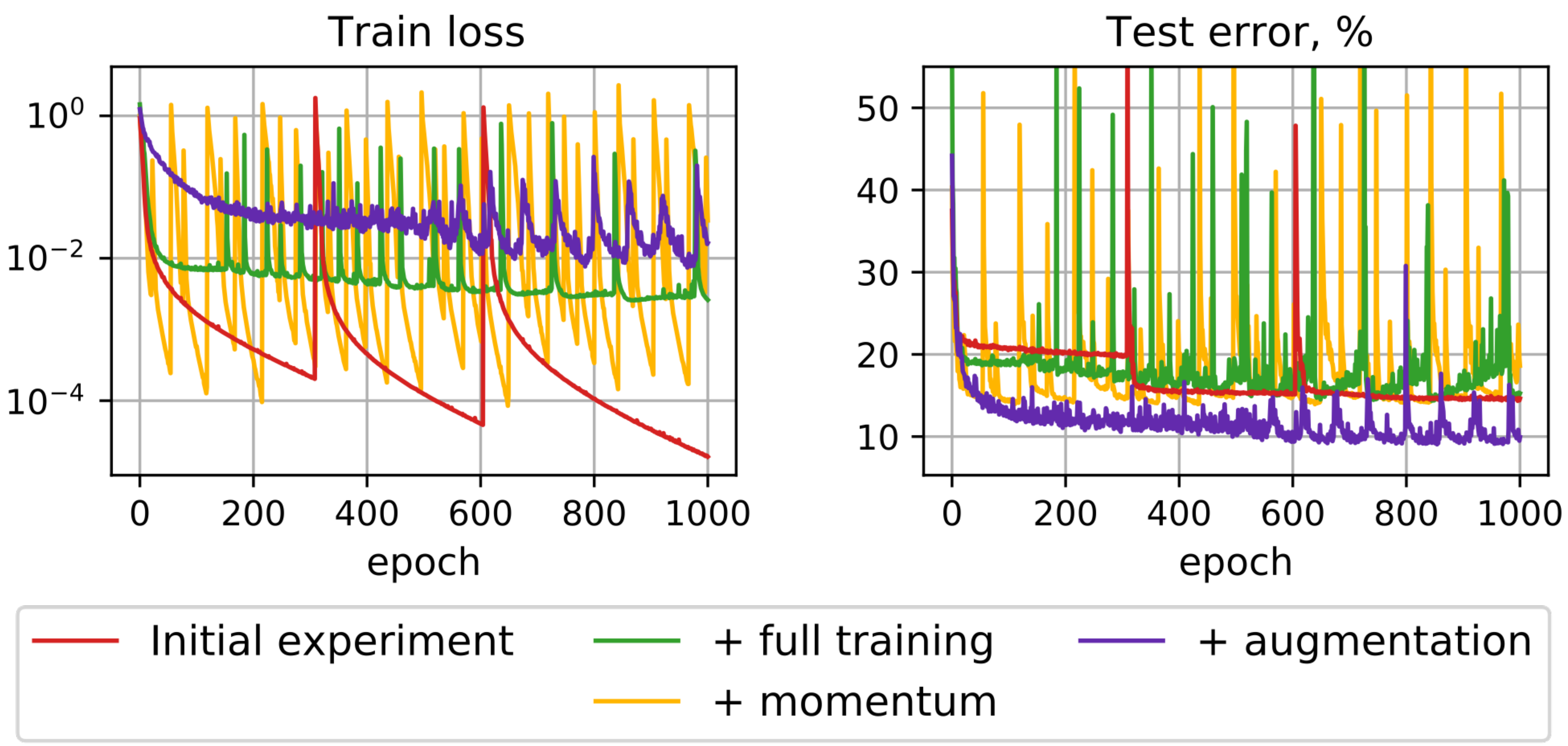
- Fully scale-invariant networks
- ~~SGD~~ → SDG + momentum
- No data augmentation



Empirical study - practical setting

Simplified setting:

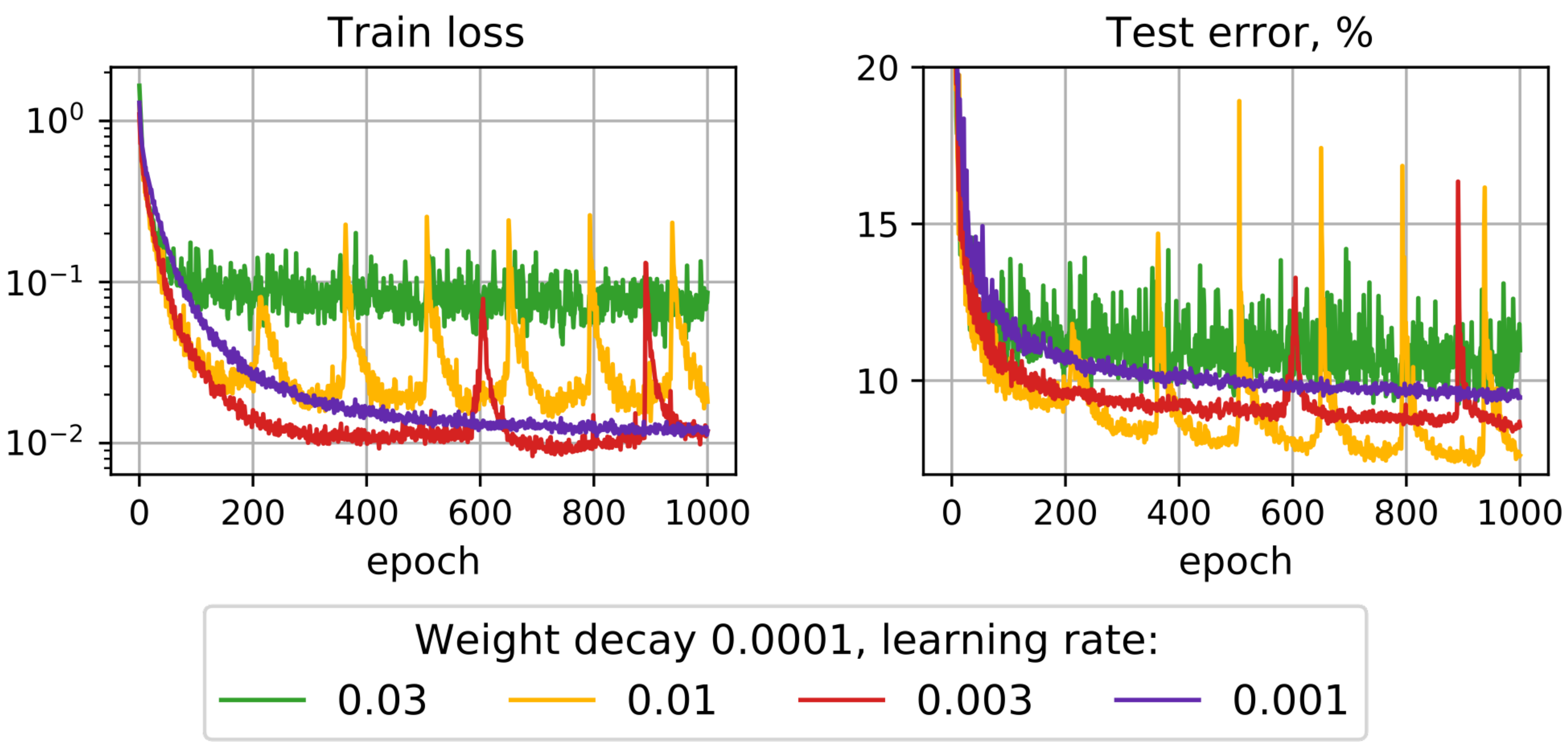
- Fully scale-invariant networks
- SGD
- ~~No data augmentation~~ → With data augmentation



Empirical study - practical setting

Practical setting:

- Standard networks
- SGD + momentum
- With data augmentation



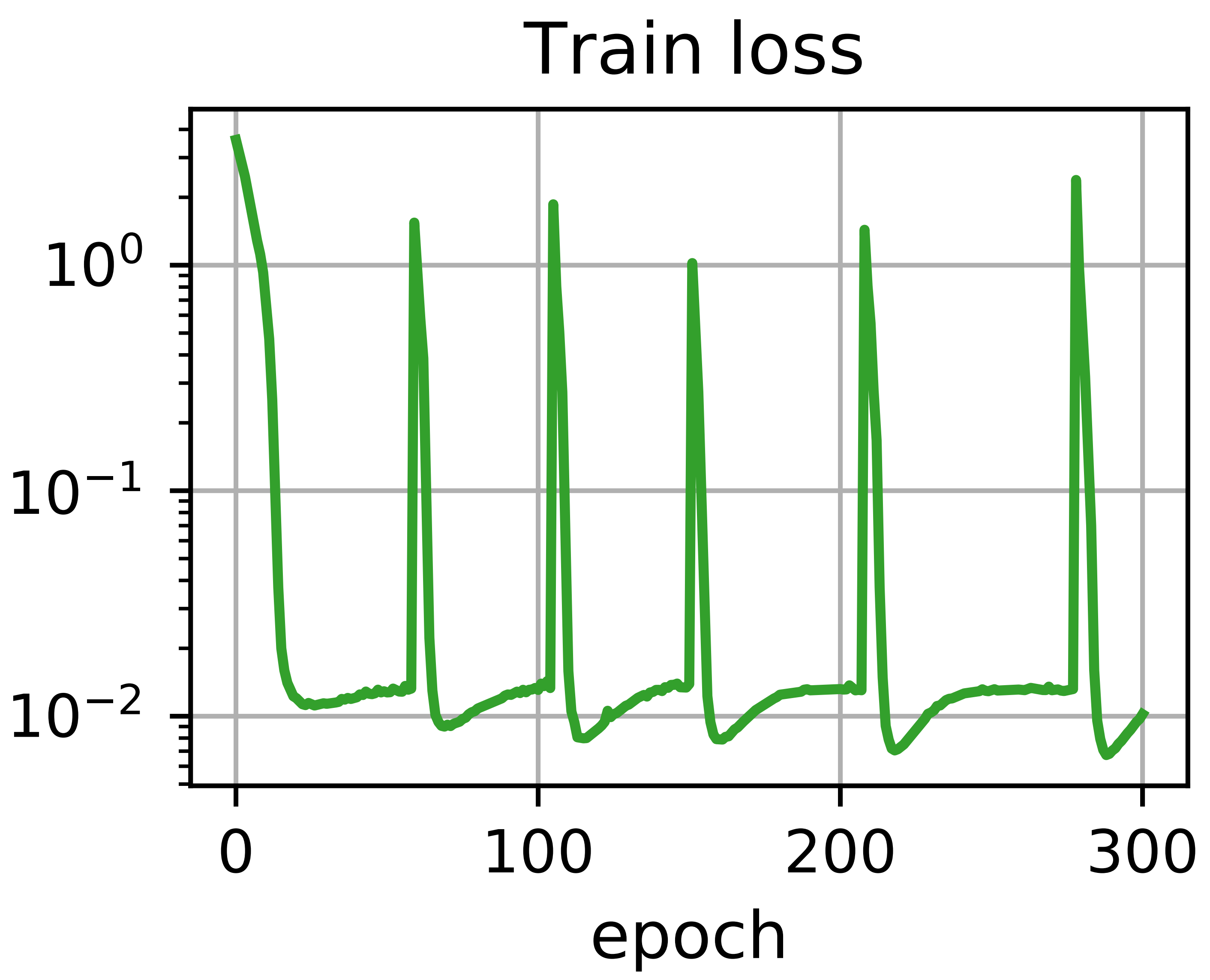
Conclusion

Periodic training behavior

Reason: BatchNorm + Weight Decay

Empirical study:

- Influence of hyperparameters
- Minima diversity
- Practical setting



Paper: https://arxiv.org/abs/2106.15739

Code: https://github.com/tipt0p/periodic_behavior_bn_wd

p(B|A)yesgroup.ru