Where Do Large Learning Rates Lead Us? A Feature Learning Perspective

Ildus Sadrtdinov, Maxim Kodryan, Eduard Pokonechny, Ekaterina Lobacheva*, Dmitry Vetrov*

CONSTRUCTOR UNIVERSITY

HSE UNIVERSITY

Overview and main idea

Existing empirical and theoretical research: for optimal results, network training should start with a large initial learning rate (LR).

What **features** are learned by neural networks when trained with different initial LRs?

We study feature learning in the **controlled synthetic example** and **image classification setup** and discover that**:**

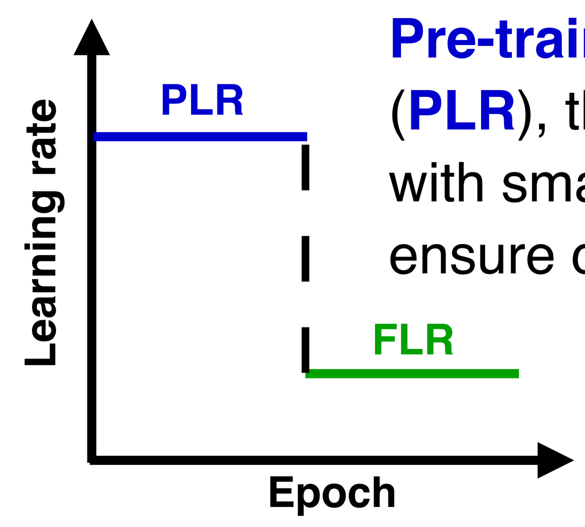
- ✓ optimal initial LRs lead to learing a sparse set of the most useful features
- ✗ smaller initial LRs try to capture all relevant features without specialization
- ✗ larger initial LRs fail to extract useful features from data and thus hurt quality

Setup

Controlled setup (for accurate experiments with fixed LRs) [1]:

- fully scale-invariant networks
- training on the unit sphere

In this setup, training happens in one of three regimes depending on LR



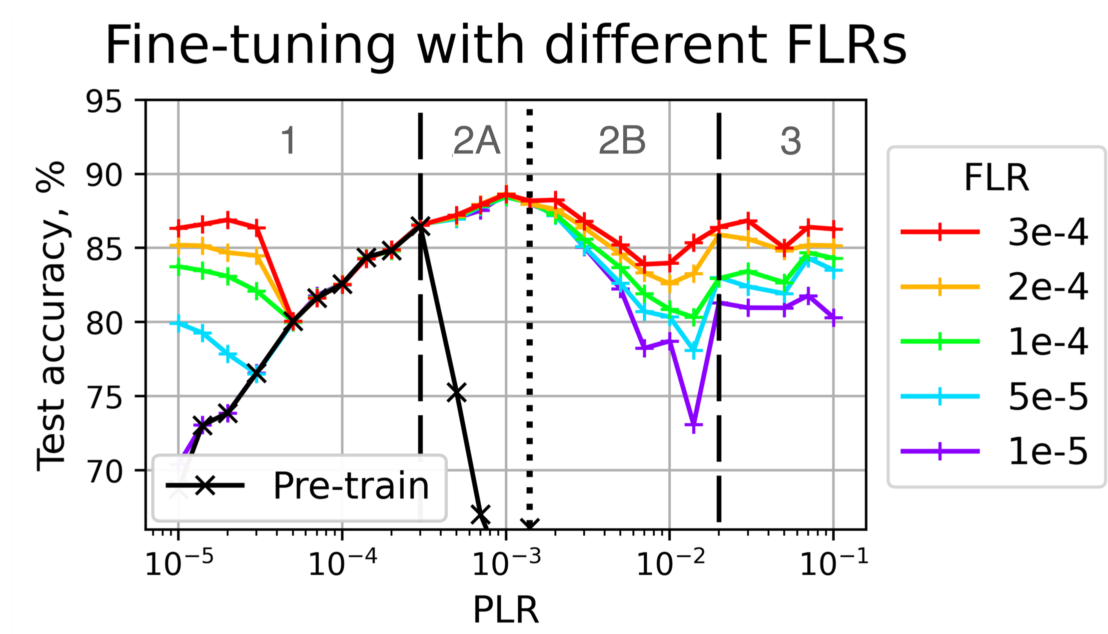
Pre-train with one LR (**PLR**), then **fine-tune** with small **FLR** to ensure convergence [2]

[1]. M. Kodryan et al., Training scale-invariant neural networks on the sphere can happen in three regimes, NeurIPS 2022

[2] E. Lobacheva et al., Large Learning Rates Improve Generalization: But How Large Are We Talking About?, NeurIPS 2023 Workshop M3L

Fine-tuning 3 regimes

Scale-invariant ResNet-18 on CIFAR-10



Regime 1: pre-training converges

- FLR < PLR: no changes
- FLR > PLR: jump to better optimum

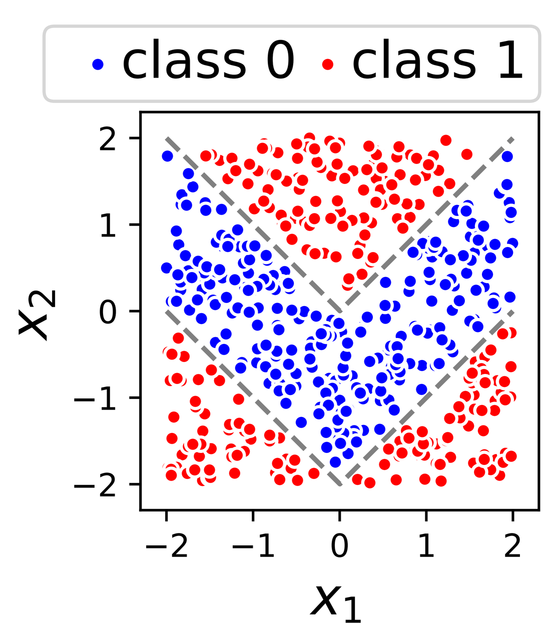
Regime 2: pre-training noisily stabilizes

- 2A: the same optimal quality for all FLRs
- 2B: different suboptimal quality when varying FLRs

Regime 3: pre-training diverges

- similar to training from scratch

Synthetic example



Experimental setup:

- binary classification
- 3-layer scale-invariant MLP
- 16 identically distributed "tick" features

Measuring feature importance

Create 16 single-feature test datasets with only one feature present

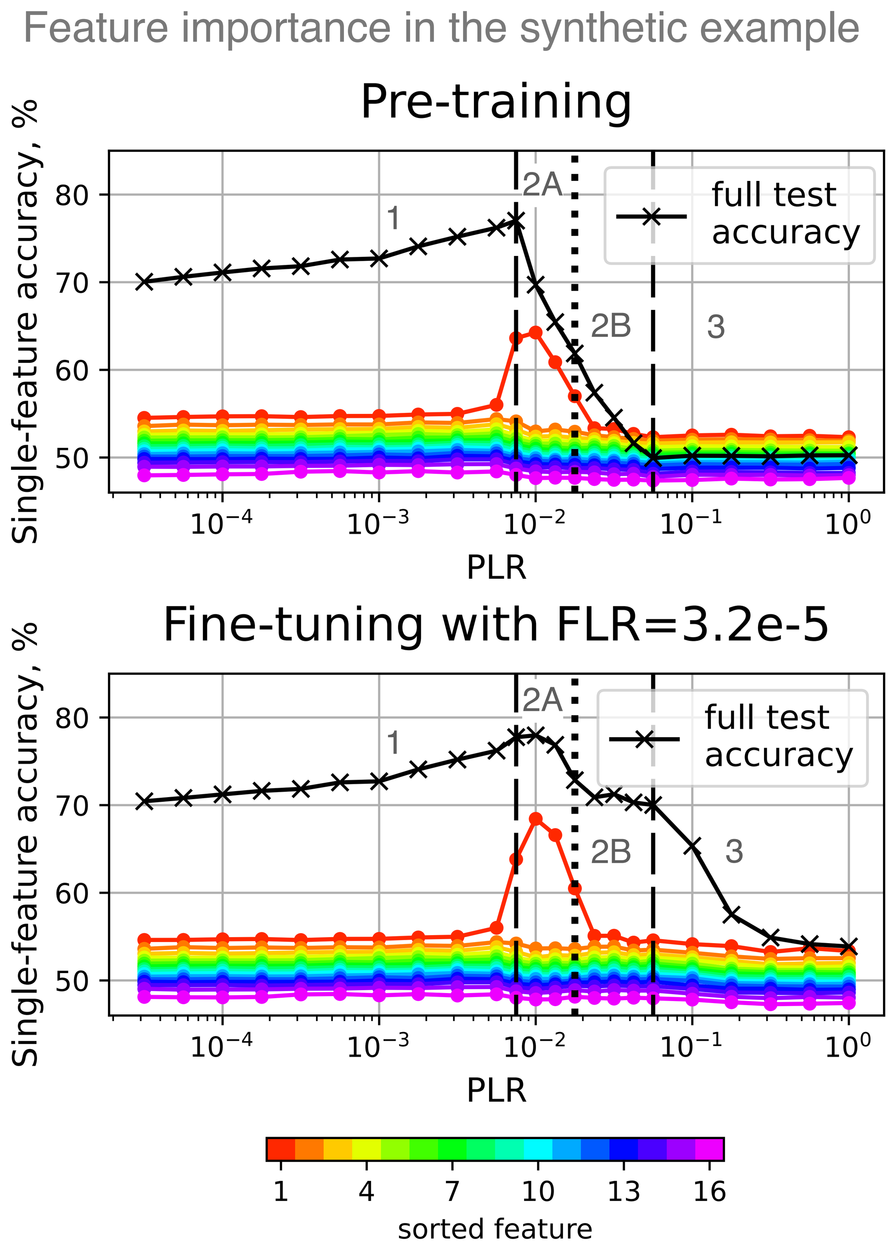
↓

Calculate accuracy on these samples

↓

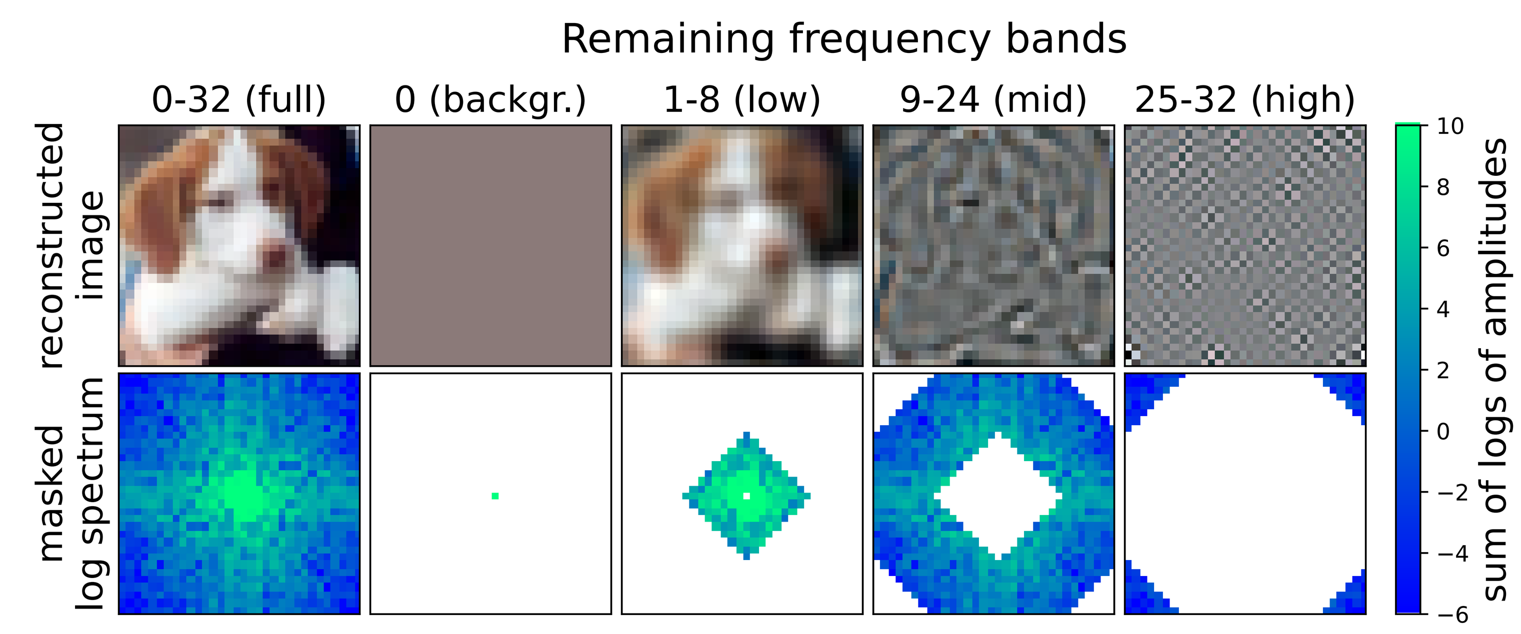
Sort values over features for each individual run

Feature importance in the synthetic example



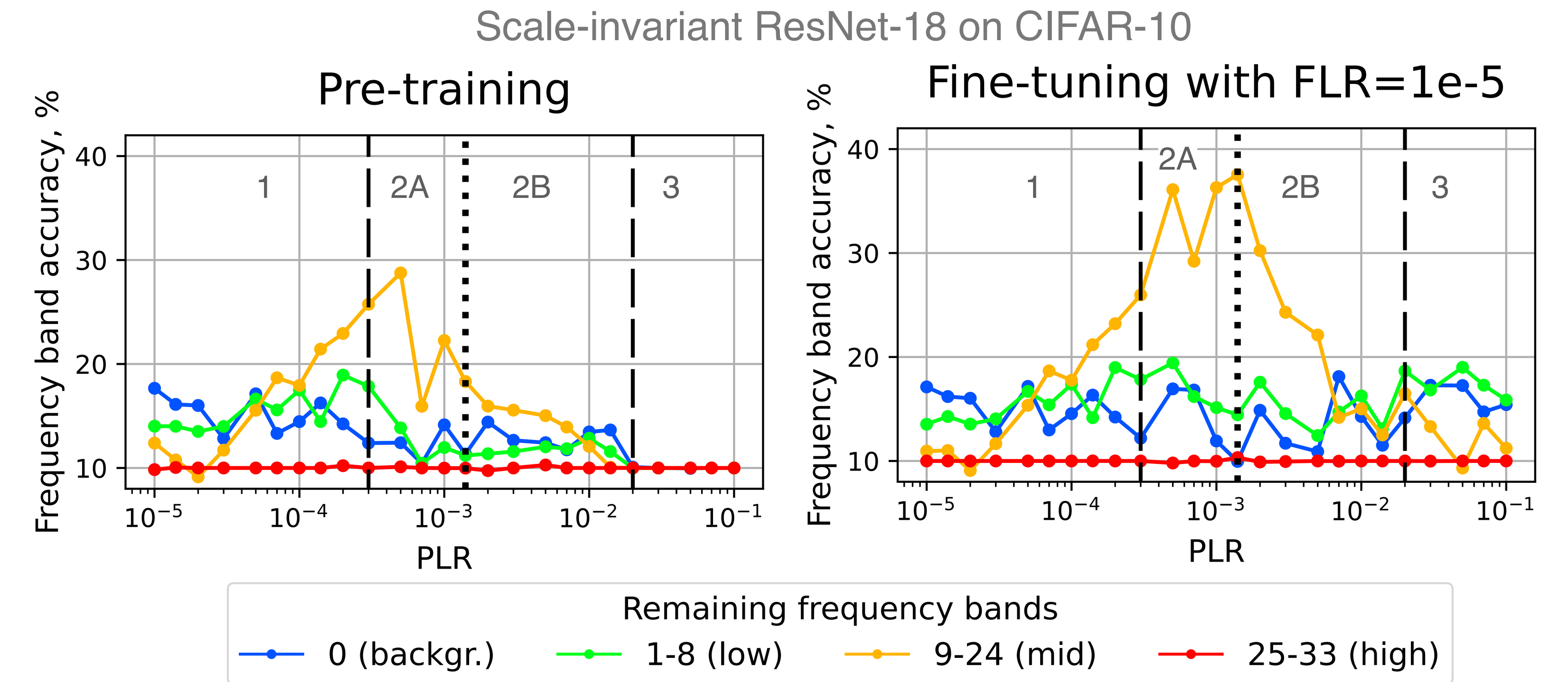
- Pre-training in reg. 1 gives roughly the same importance to all features
- Although **all features are equally useful**, pre-training in reg. 2A selects **only one feature** leading to **sparsity**
- When pre-training in reg. 2B and 3, feature learning ability is decreased, leading to lower quality and no sparsity

Fourier frequency bands as features



Apply 2D-DFT to test images → Zero out all but one group of frequency bands → Reconstruct images with inverse 2D-DFT and evaluate models

Scale-invariant ResNet-18 on CIFAR-10



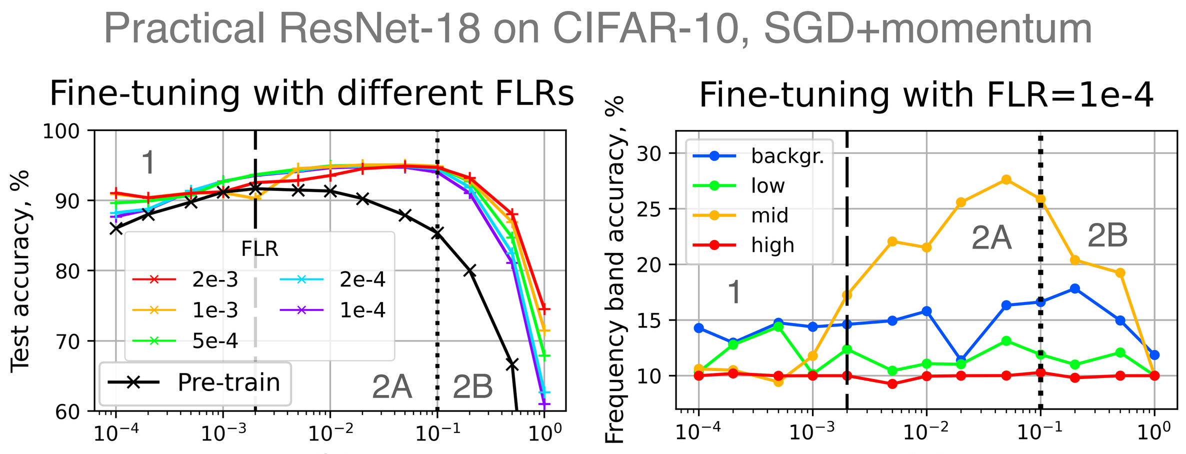
- Pre-training in reg. 2A shows feature sparsity with a focus on **mid frequencies**, persisting after fine-tuning
- Small PLRs of reg. 1 slightly favour **background** and **low-frequency** features
- Increasing PLR to reg. 2B and 3 removes sparsity

Practical setting

The same feature learning analysis for practical setting:

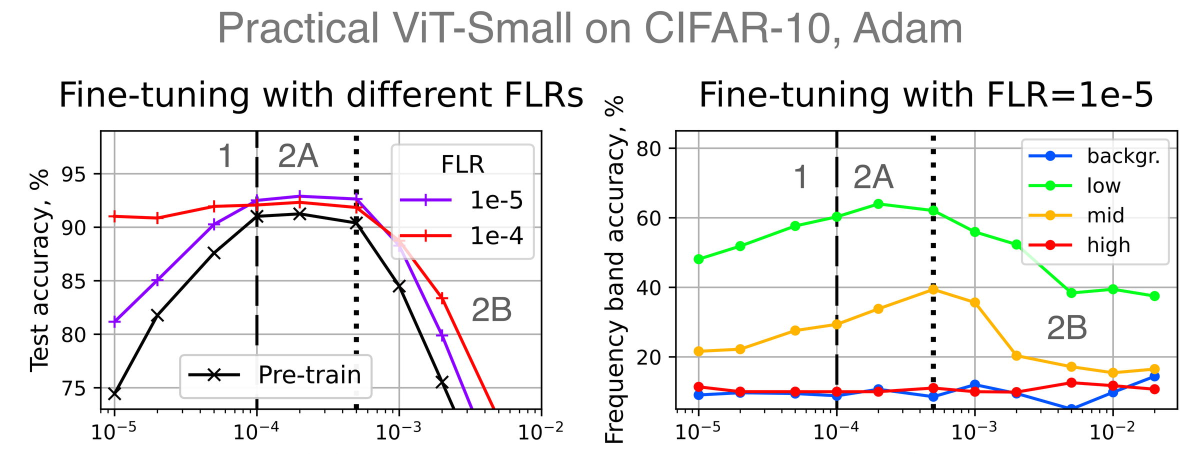
- regular (not fully scale-invariant) models
- image augmentations
- weight decay

Practical ResNet-18 on CIFAR-10, SGD+momentum



- Similarly to the scale-invariant setup, the importance of **mid-frequency** features for practical ResNet peaks in reg. 2A

Practical ViT-Small on CIFAR-10, Adam



- In contrast, ViT focuses on both **low-frequency** and **mid-frequency** features, preferring the former component

More results about other setups and SWA are here:

Paper

You may also like:

Large LRs **3 regimes**