Where Do Large Learning Rates Lead Us?

Ildus Sadrtdinov*

Maxim Kodryan*

Eduard Pokonechny*

Ekaterina Lobacheva†

Dmitry Vetrov†

CONSTRUCTOR UNIVERSITY

HSE University

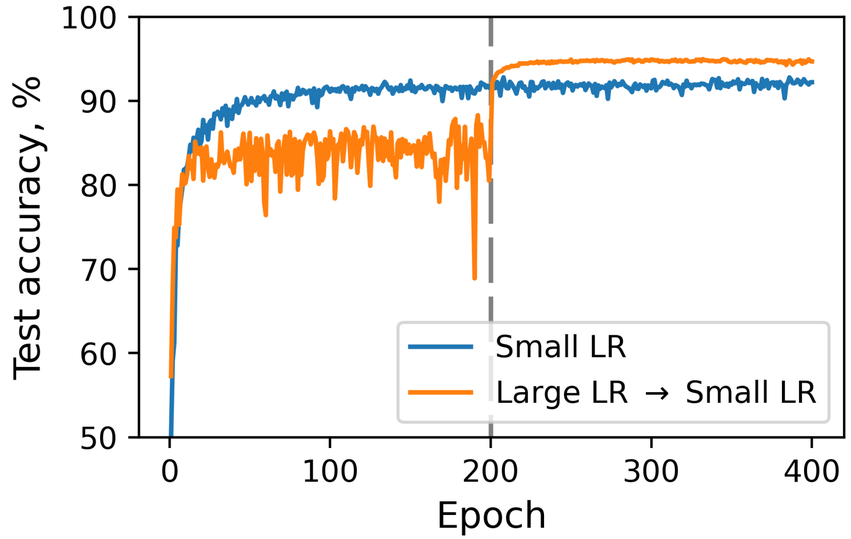
p(B|A)yesgroup

* equal contribution

† shared senior authorship

Benefits of large Learning Rates (LRs)

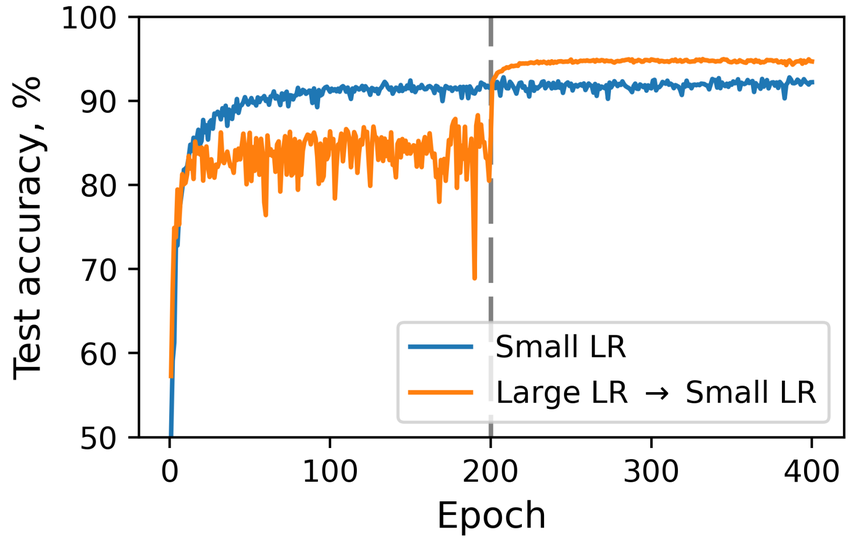
Start training of NN with
large LR and then anneal it



Training with **small LR** vs. **large initial LR + annealing**
ResNet, CIFAR-10

Benefits of large Learning Rates (LRs)

Start training of NN with
large LR and then anneal it



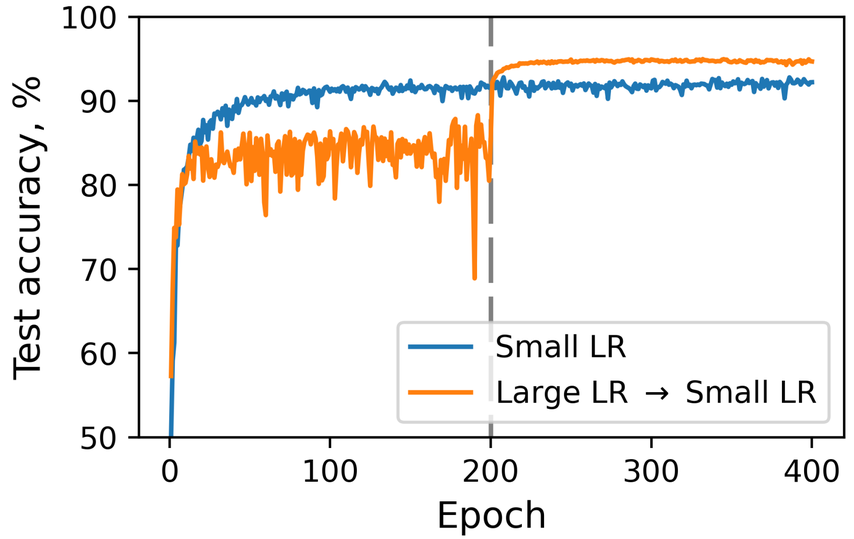
Training with **small LR** vs. **large initial LR + annealing**
ResNet, CIFAR-10

Large initial LRs:

- faster convergence
- minima with favorable properties (generalization, flatter landscape, …)

Benefits of large Learning Rates (LRs)

Start training of NN with **large LR** and then anneal it



Training with **small LR** vs. **large initial LR + annealing**
ResNet, CIFAR-10

Large initial LRs:

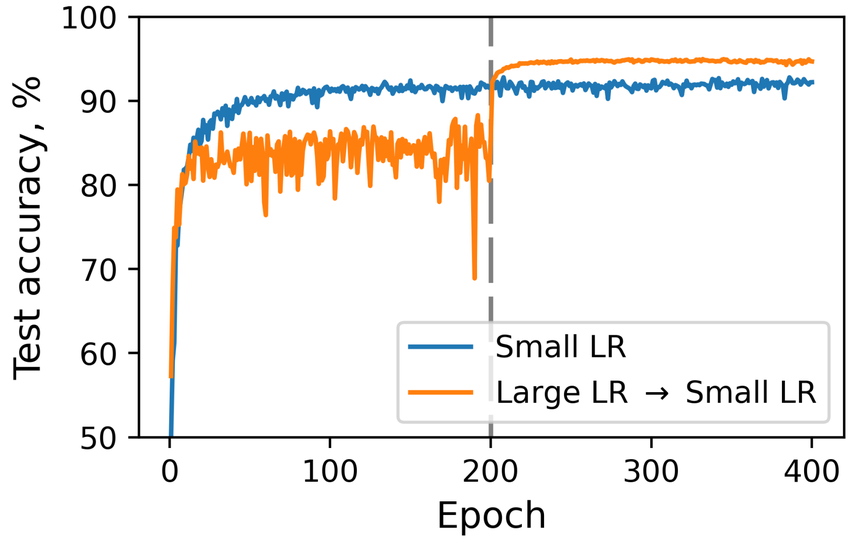
- faster convergence
- minima with favorable properties (generalization, flatter landscape, …)

Our focus:

1. Which initial LRs are **optimal** for test performance?

Benefits of large Learning Rates (LRs)

Start training of NN with **large LR** and then anneal it



Training with **small LR** vs. **large initial LR + annealing**
ResNet, CIFAR-10

Large initial LRs:

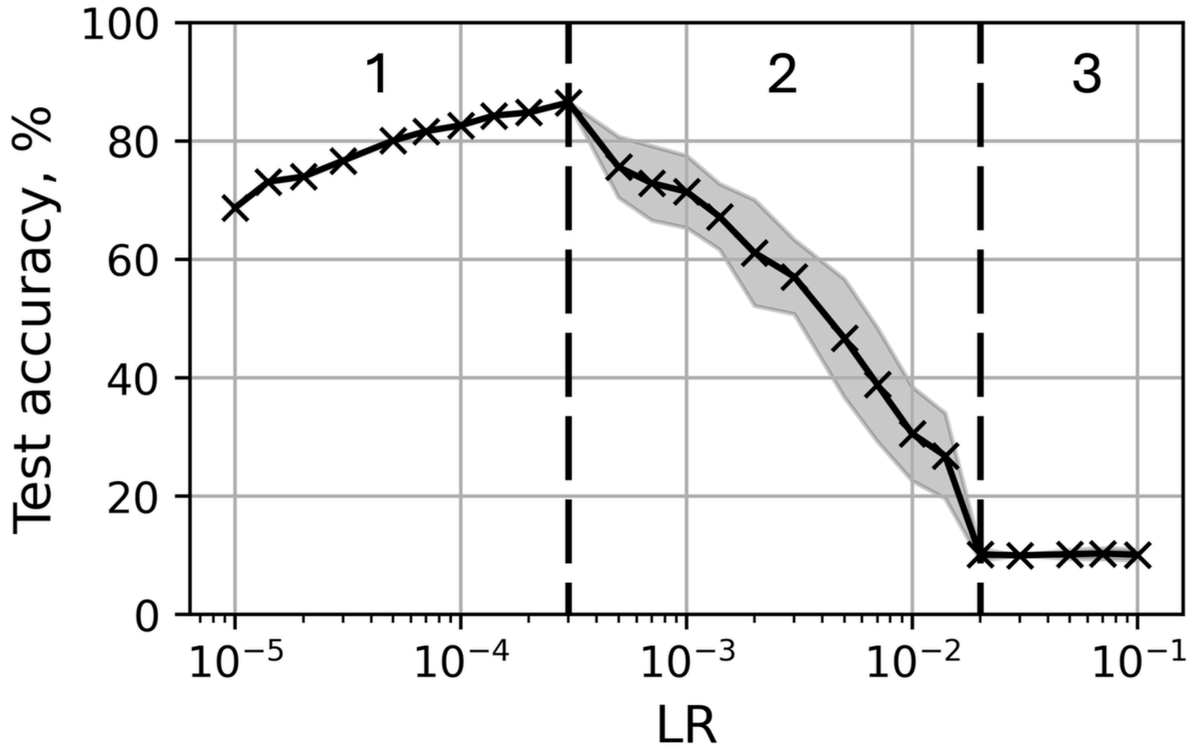
- faster convergence
- minima with favorable properties (generalization, flatter landscape, …)

Our focus:

1. Which initial LRs are **optimal** for test performance?

2. What are the key characteristics of models trained with **different LRs**?

Setup and three regimes of training



SI ResNet-18, CIFAR-10

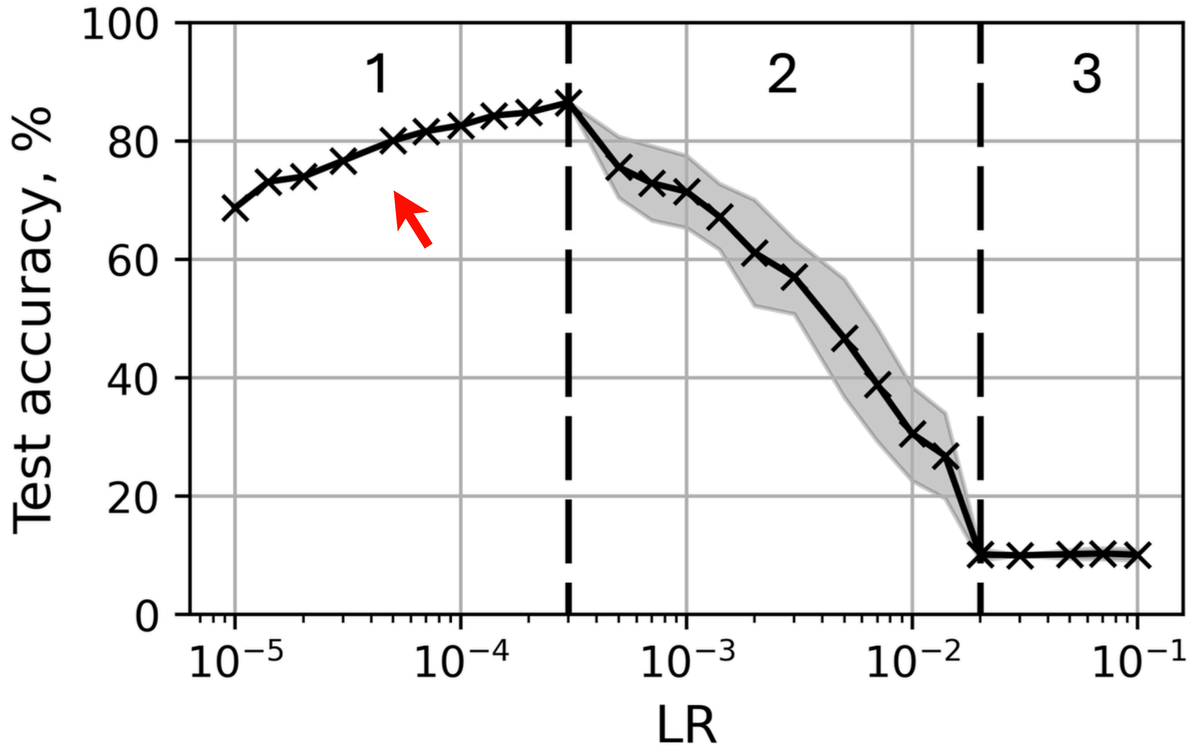
Controlled scale-invariant (SI) setup from [Kodryan et al., 2022]

(1) *convergence*

(2) *chaotic equilibrium*

(3) *divergence*

Setup and three regimes of training



SI ResNet-18, CIFAR-10

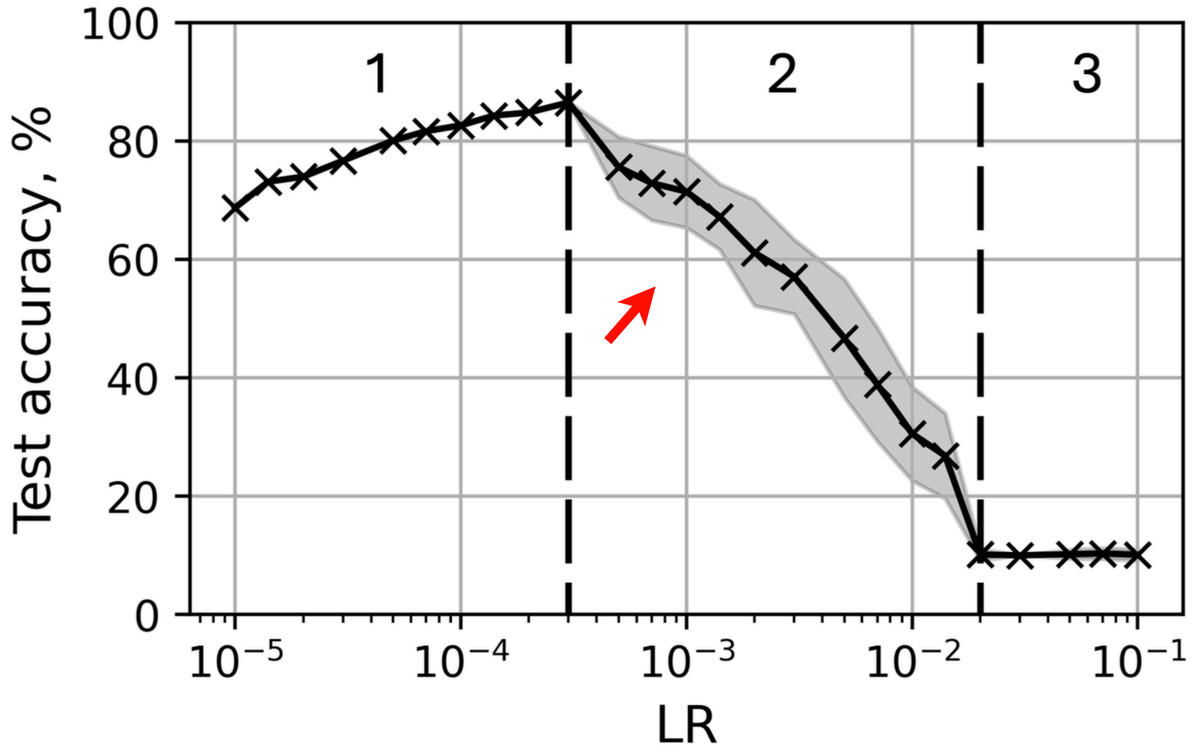
Controlled scale-invariant (SI) setup from [Kodryan et al., 2022]

(1) *convergence*

(2) *chaotic equilibrium*

(3) *divergence*

Setup and three regimes of training



SI ResNet-18, CIFAR-10

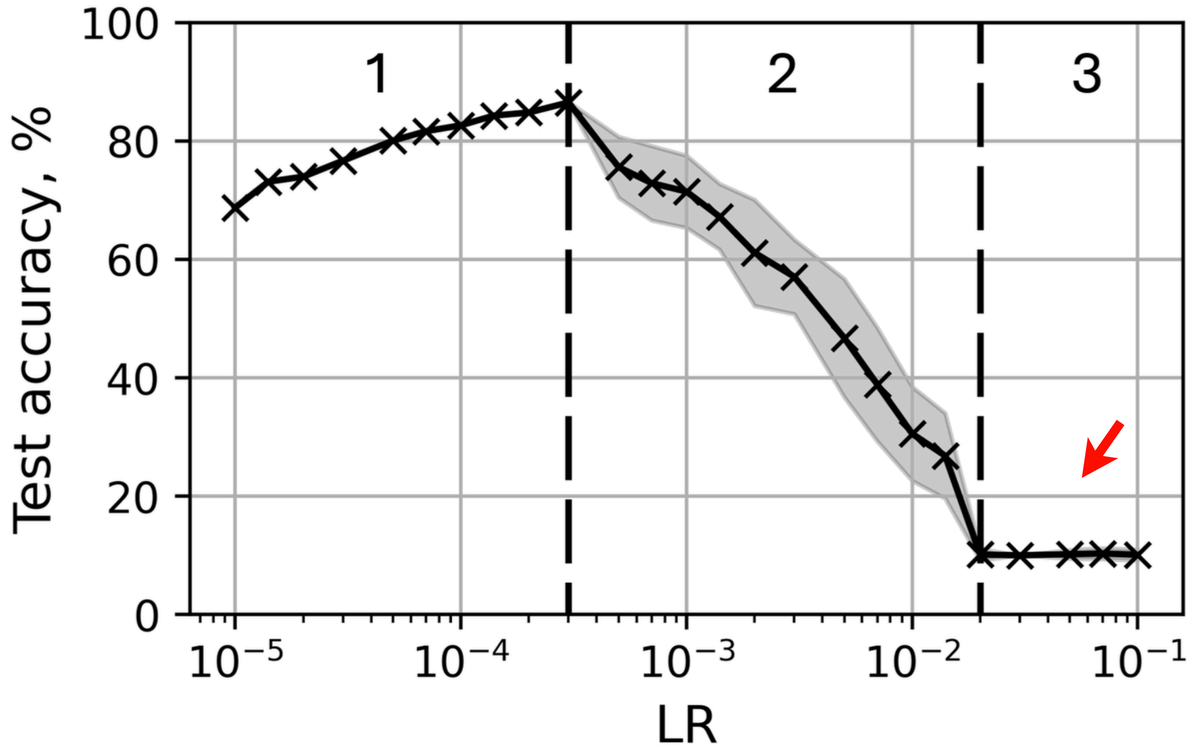
Controlled scale-invariant (SI) setup from [Kodryan et al., 2022]

(1) *convergence*

(2) *chaotic equilibrium*

(3) *divergence*

Setup and three regimes of training



SI ResNet-18, CIFAR-10

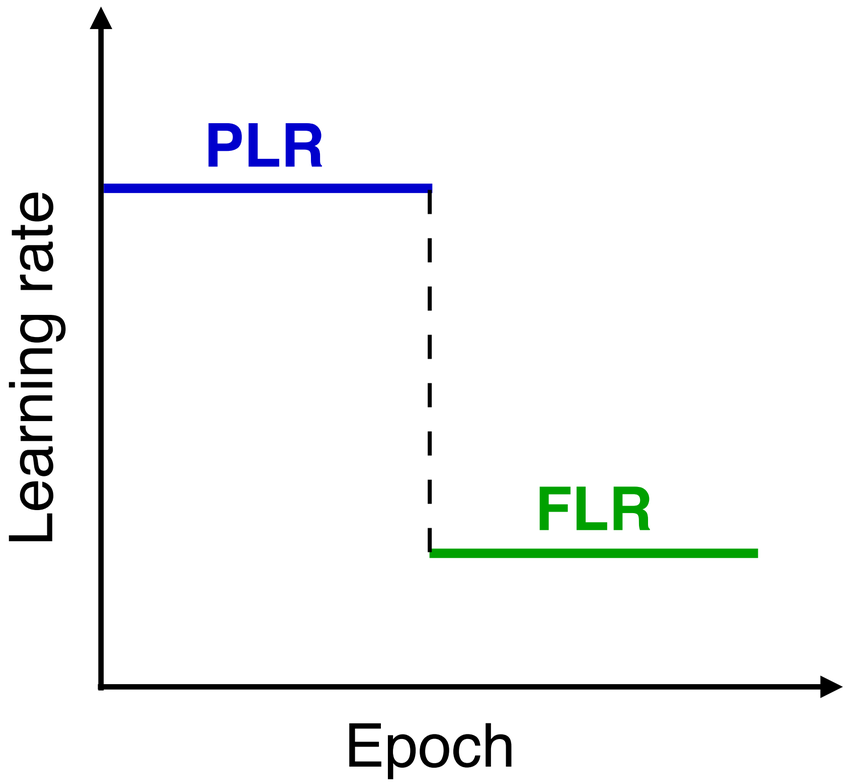
Controlled scale-invariant (SI) setup from [Kodryan et al., 2022]

(1) *convergence*

(2) *chaotic equilibrium*

(3) *divergence*

Methodology



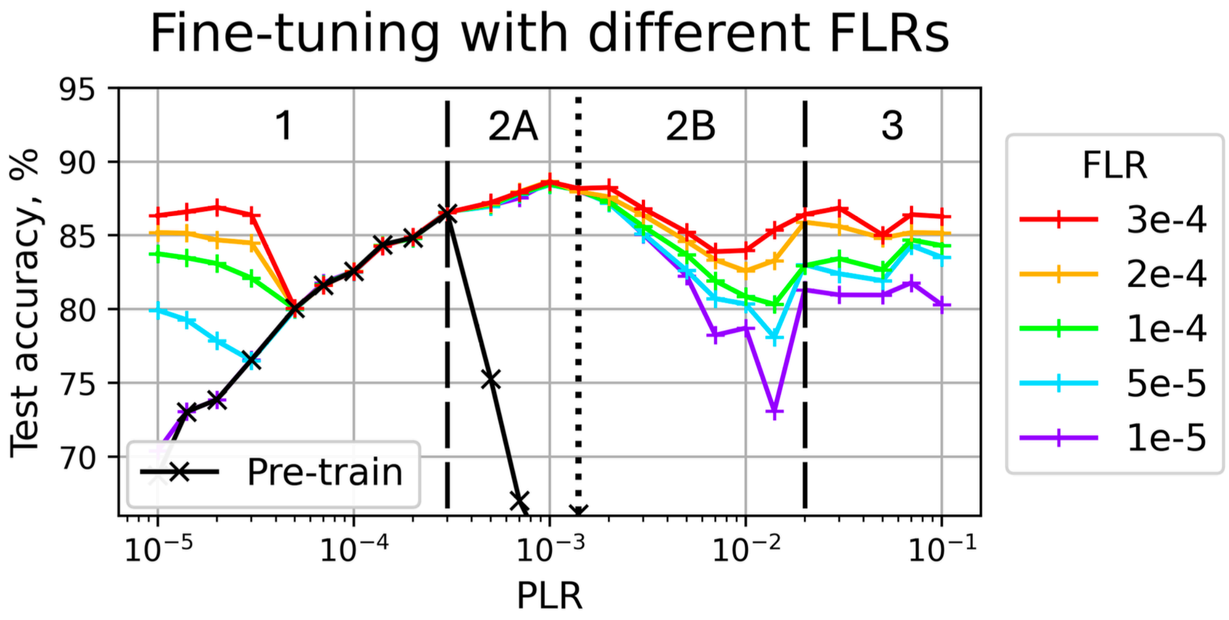
Start training

- **Pre-train** with a fixed LR (**PLR**)
- **PLR** is taken from different regimes

Obtain final solution

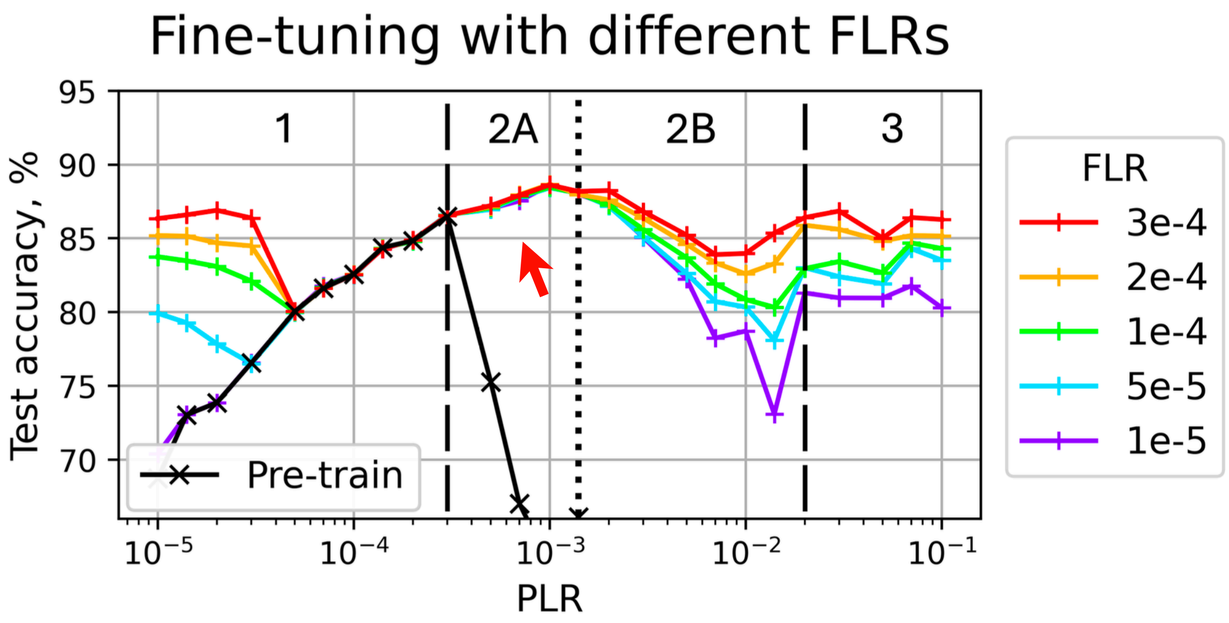
- **Fine-tune** with another LR (**FLR**)
- **FLR** is taken from regime 1 to ensure convergence

Best LRs for generalization



SI ResNet-18, CIFAR-10

Best LRs for generalization

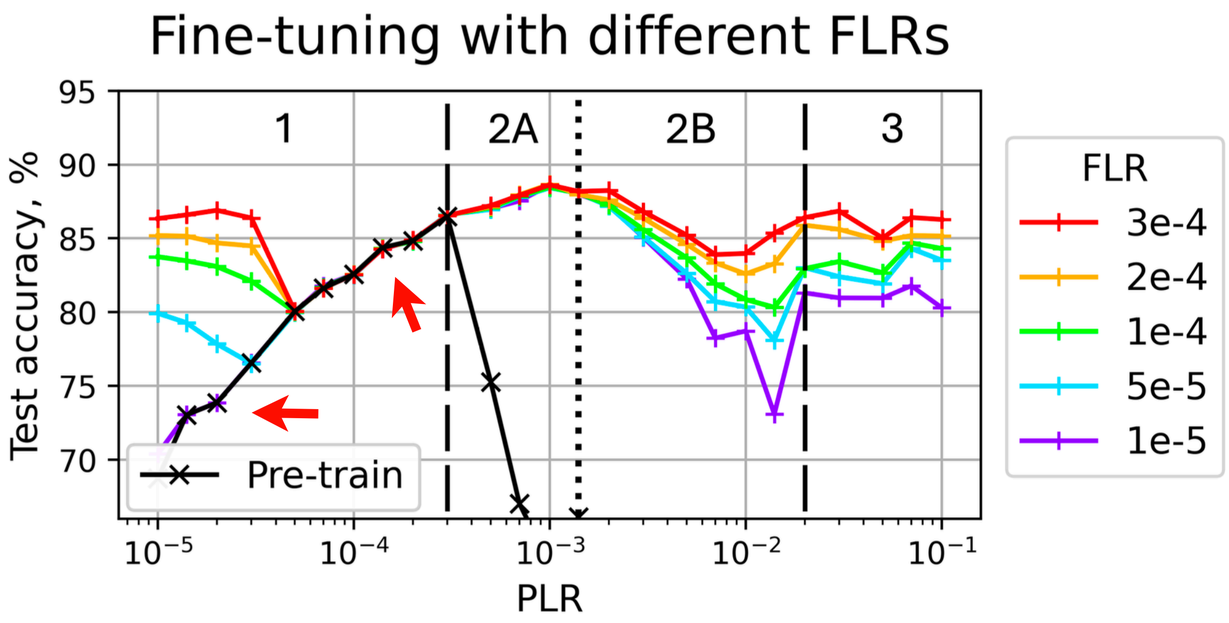


SI ResNet-18, CIFAR-10

PLR from 2A:

- optimal quality

Best LRs for generalization



SI ResNet-18, CIFAR-10

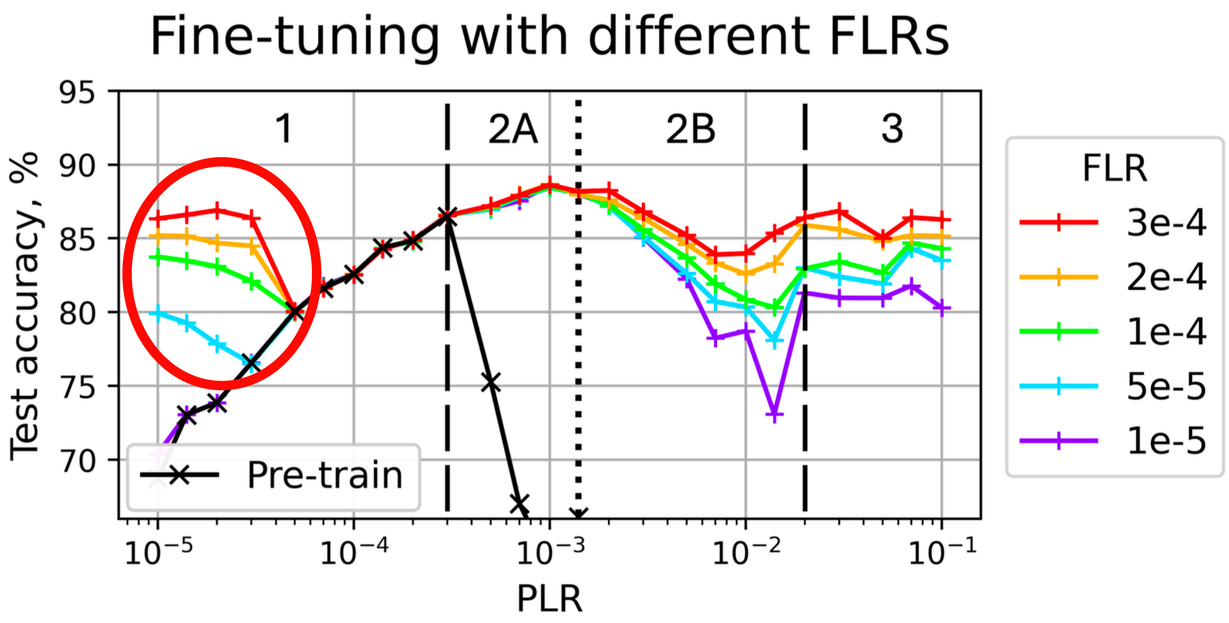
PLR from 2A:

- optimal quality

Smaller PLR (regime 1):

- FLR ≤ PLR — no improvement

Best LRs for generalization



SI ResNet-18, CIFAR-10

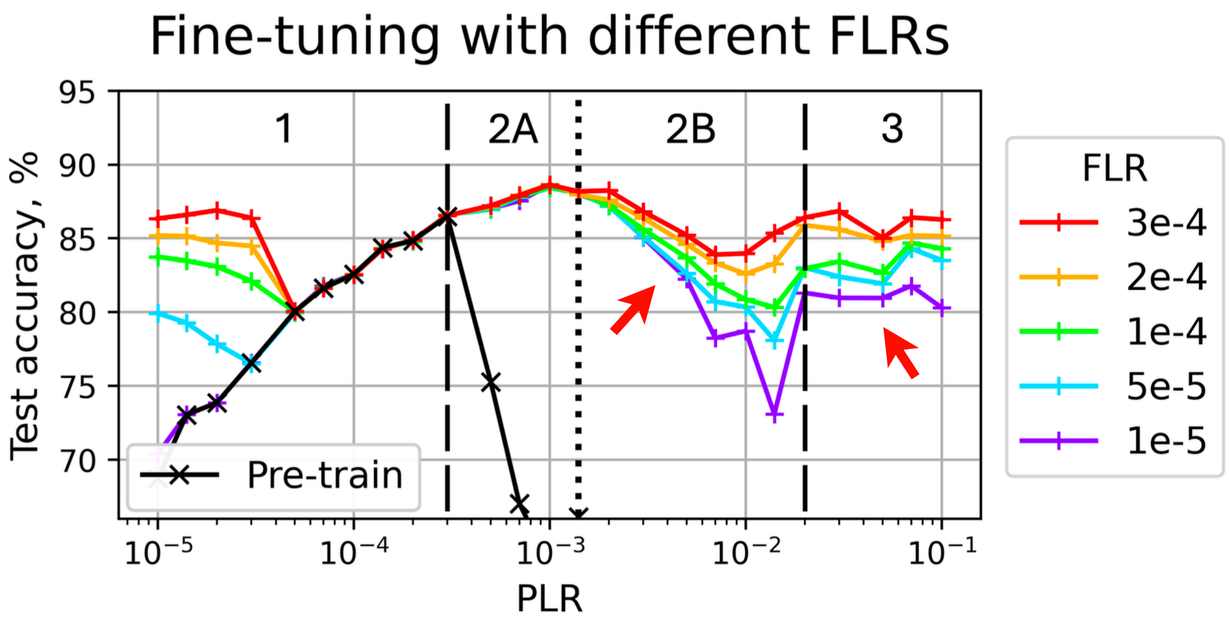
PLR from 2A:

- optimal quality

Smaller PLR (regime 1):

- FLR ≤ PLR — no improvement
- FLR > PLR — jump to better minimum

Best LRs for generalization



SI ResNet-18, CIFAR-10

PLR from 2A:

- optimal quality

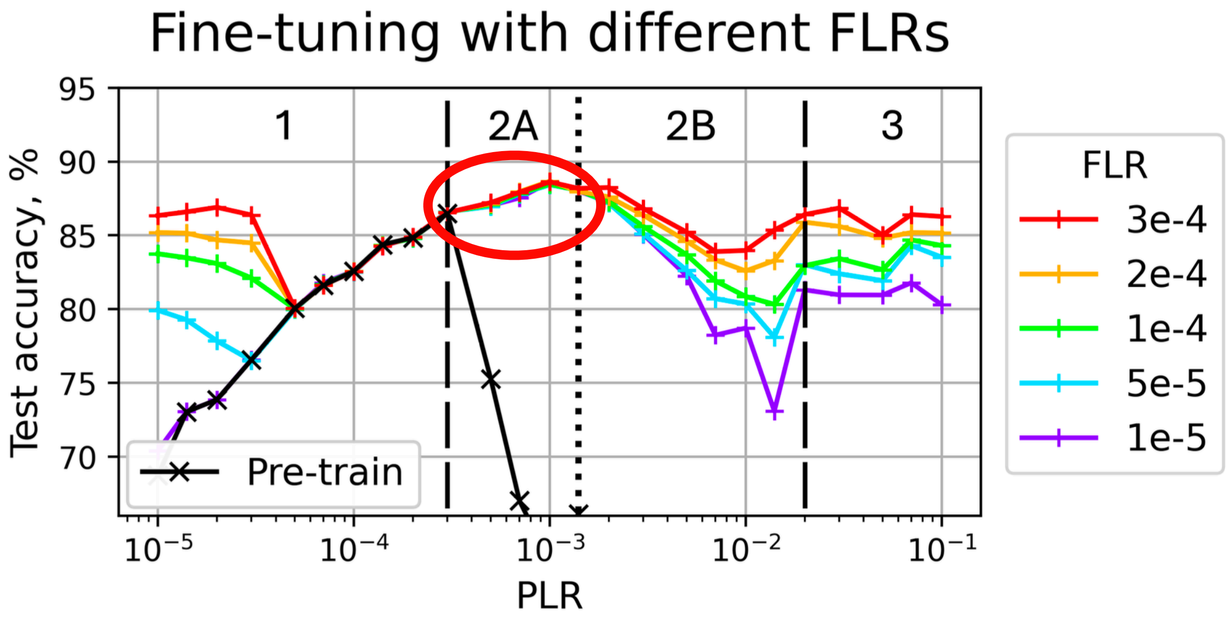
Smaller PLR (regime 1):

- FLR ≤ PLR — no improvement
- FLR > PLR — jump to better minimum

Larger PLR (regimes 2B, 3):

- lower quality

Best LRs for generalization



SI ResNet-18, CIFAR-10

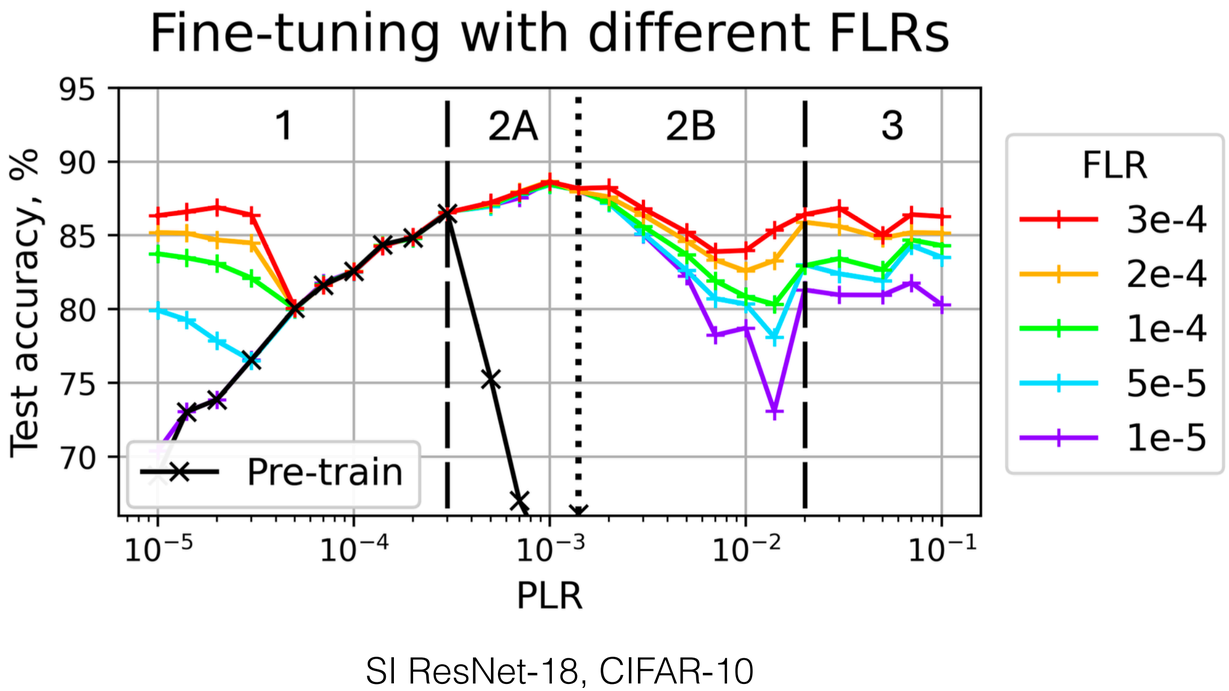
PLR from 2A:

- same quality for all FLRs

Other PLRs (regime 1, 2B, 3):

- quality depends on FLR

Best LRs for generalization

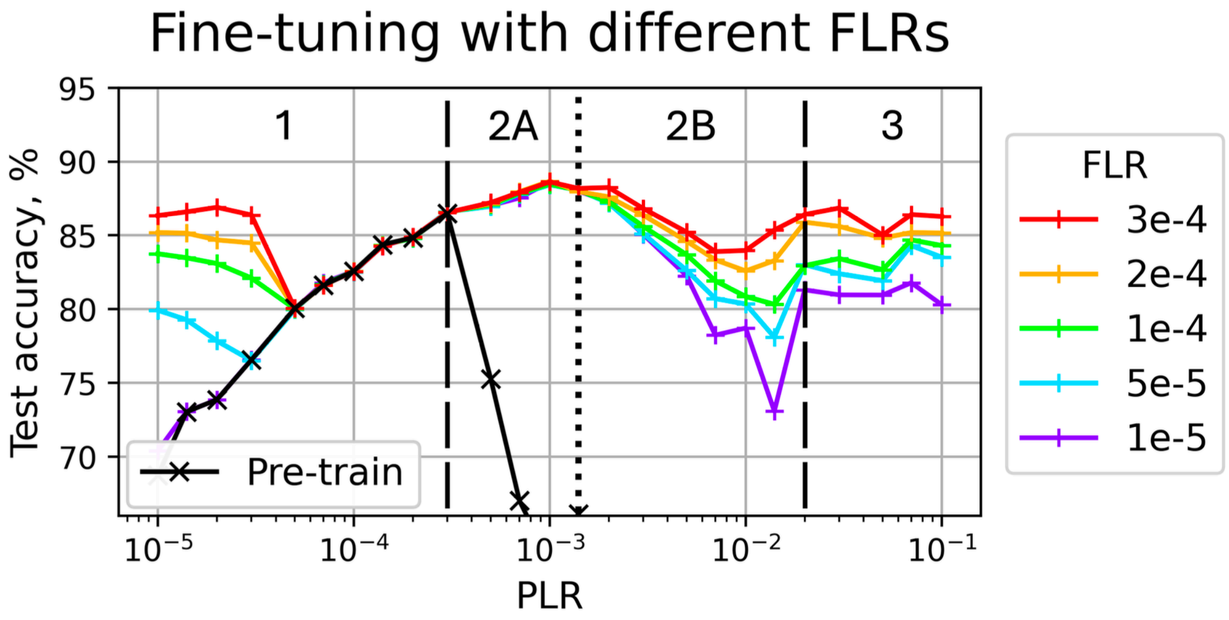


SI ResNet-18, CIFAR-10

— *Which LRs are best to start with?*

— A narrow range just above the convergence threshold (subregime 2A)

Best LRs for generalization



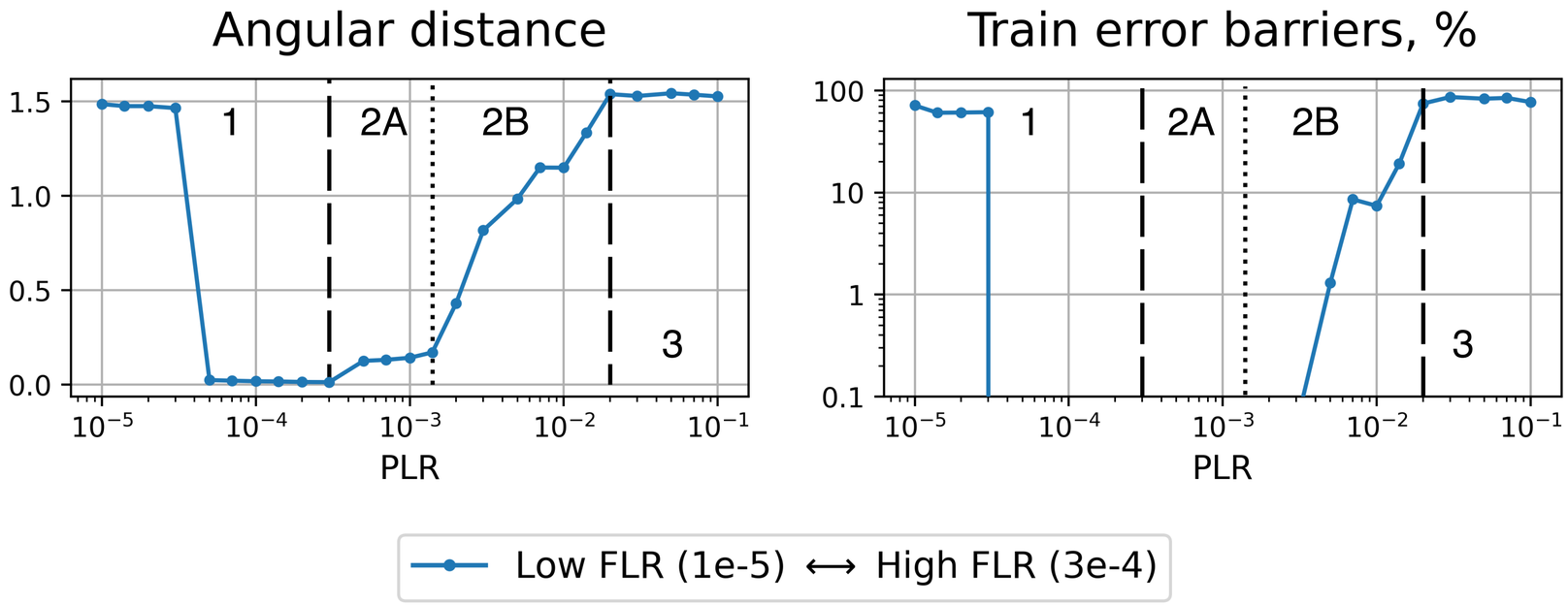
SI ResNet-18, CIFAR-10

— *Which LRs are best to start with?*

— A narrow range just above the convergence threshold (subregime 2A)

— *Why?..*

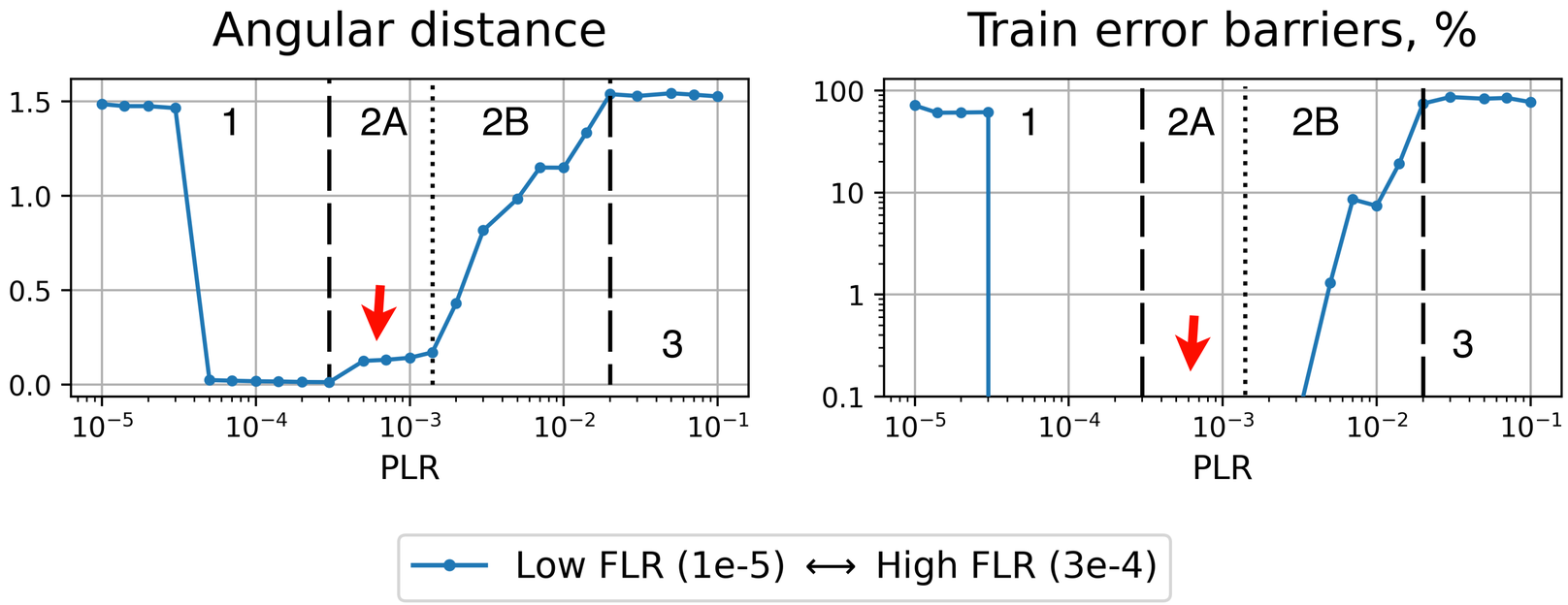
Loss landscape perspective



SI ResNet-18, CIFAR-10

- ↓ angle — closer in weight space
- ↓ error barrier — same basin

Loss landscape perspective



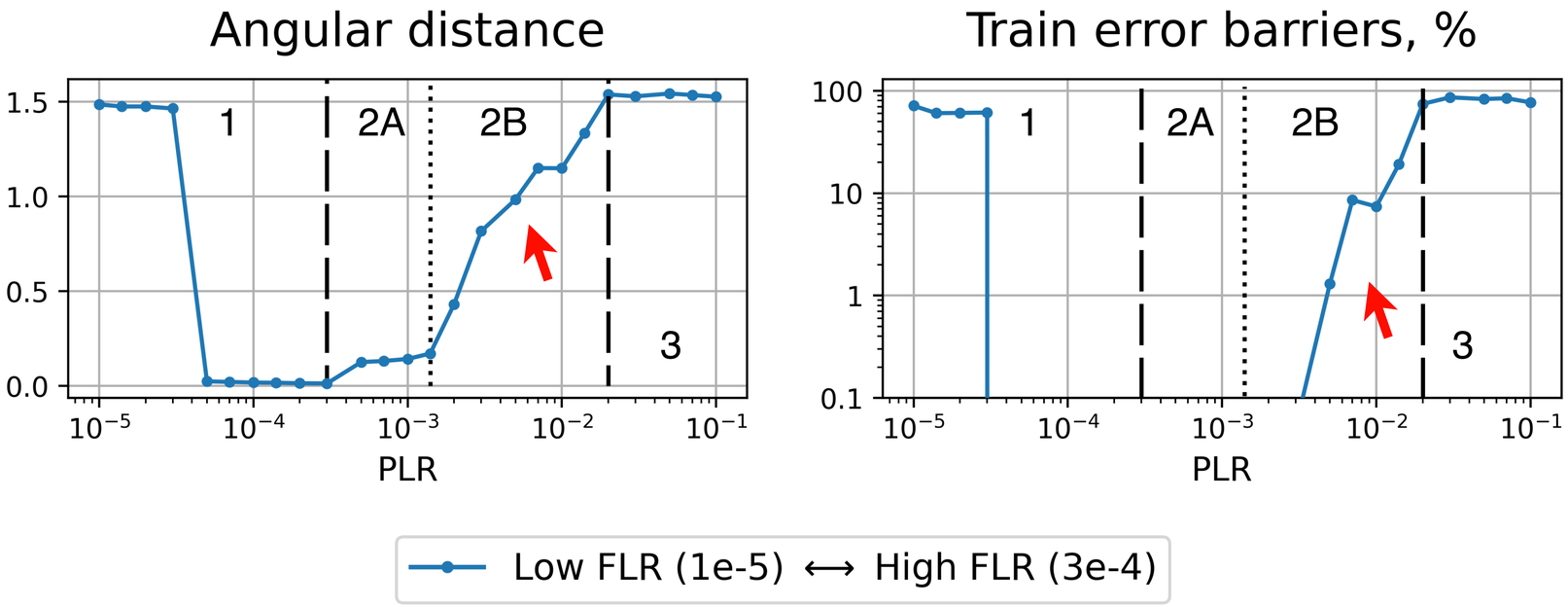
SI ResNet-18, CIFAR-10

PLR from 2A:

- same basin

- ↓ angle — closer in weight space
- ↓ error barrier — same basin

Loss landscape perspective



SI ResNet-18, CIFAR-10

PLR from 2A:

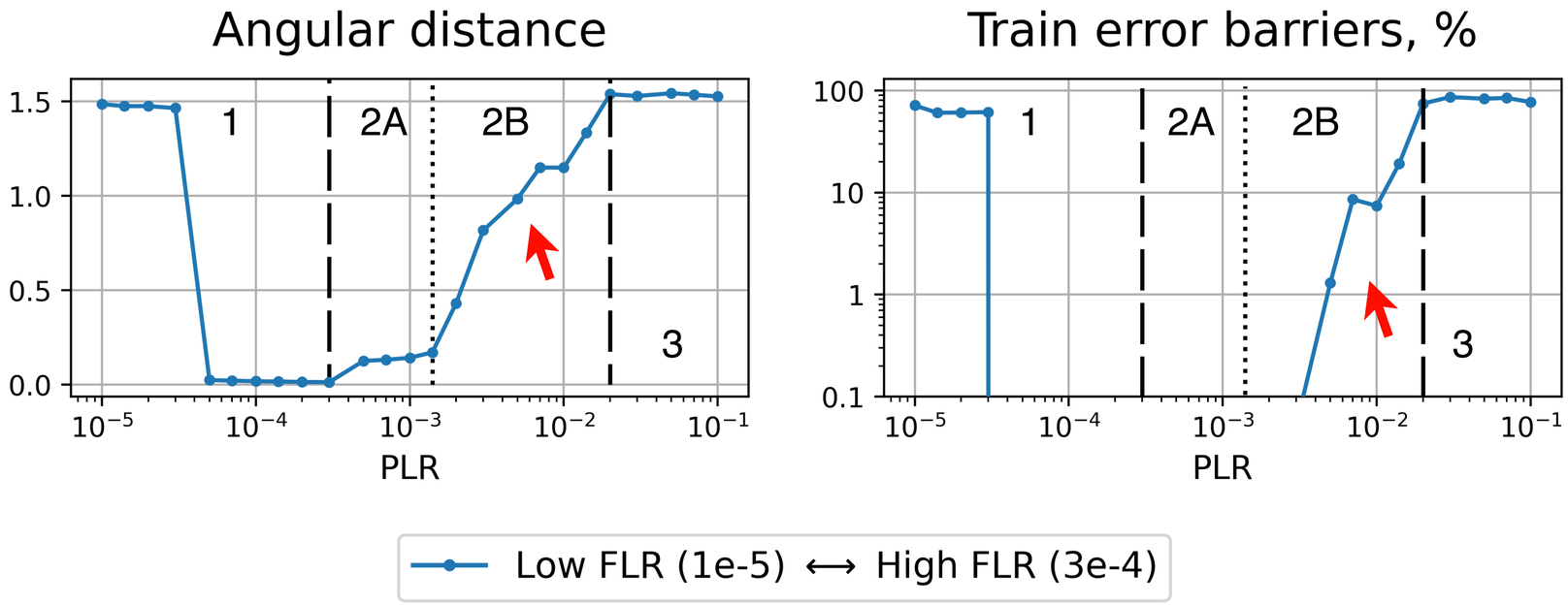
- same basin

PLR from 2B:

- distinct minima

- ↓ angle — closer in weight space
- ↓ error barrier — same basin

Loss landscape perspective



SI ResNet-18, CIFAR-10

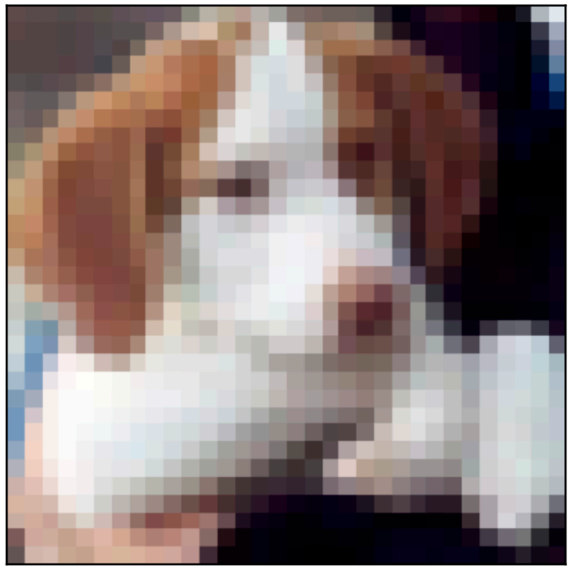
— *Why is 2A better than 2B?*

— It locates a convex basin of good solutions

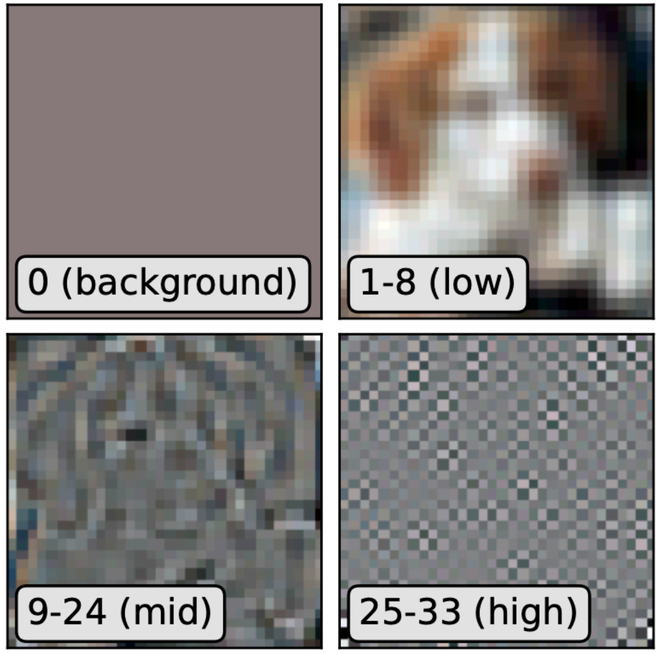
- ↓ angle — closer in weight space
- ↓ error barrier — same basin

Feature learning perspective

Original image

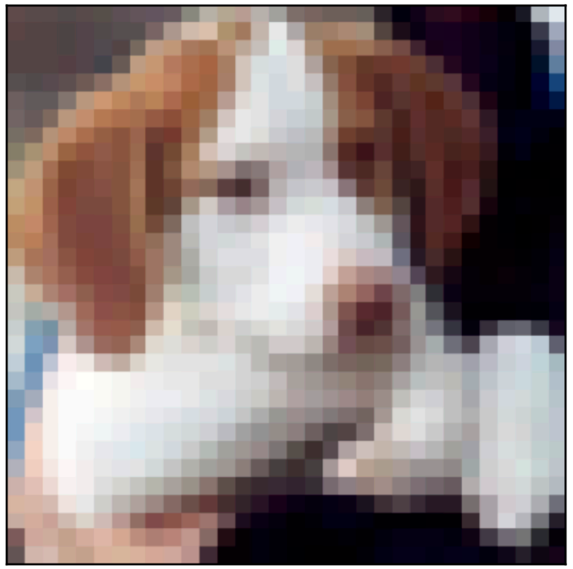


Frequency components

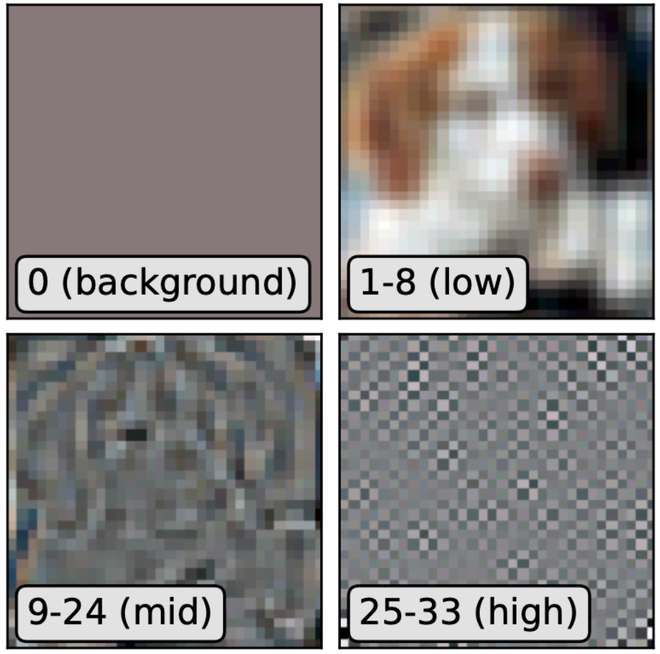


Feature learning perspective

Original image



Frequency components

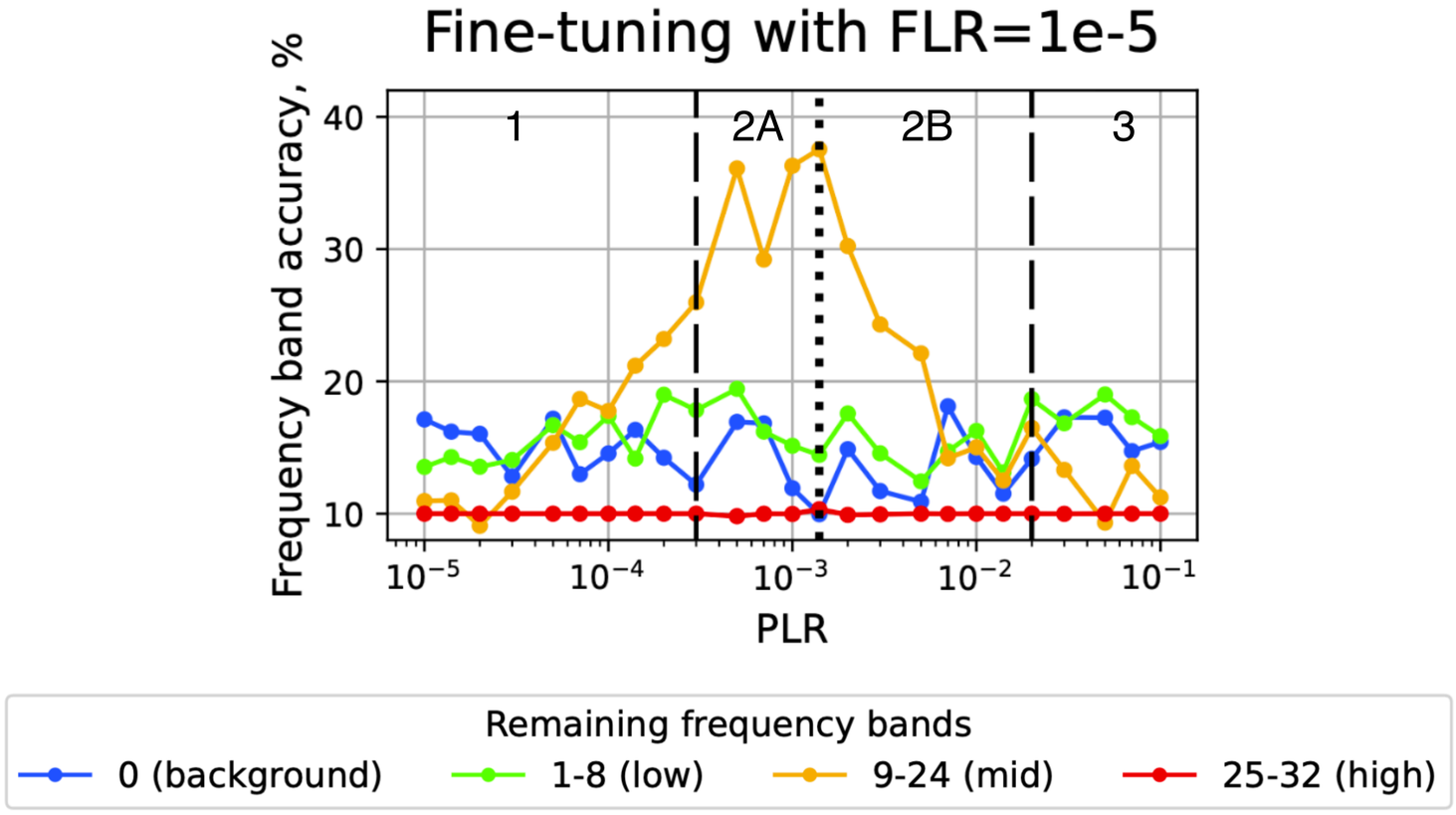


Apply to test images

4 new test sets:

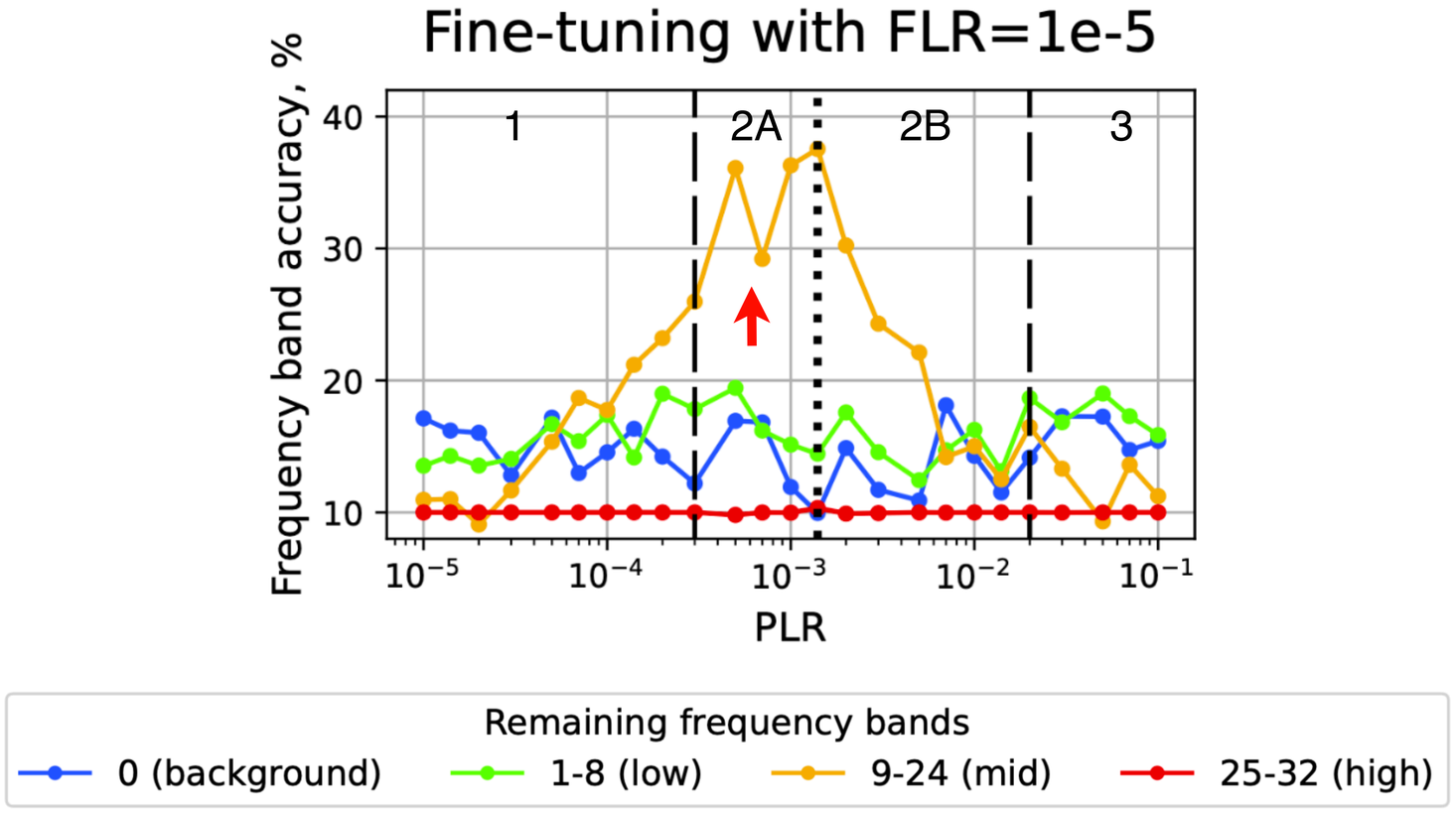
- background
- low
- mid
- high

Feature learning perspective



SI ResNet-18, CIFAR-10

Feature learning perspective

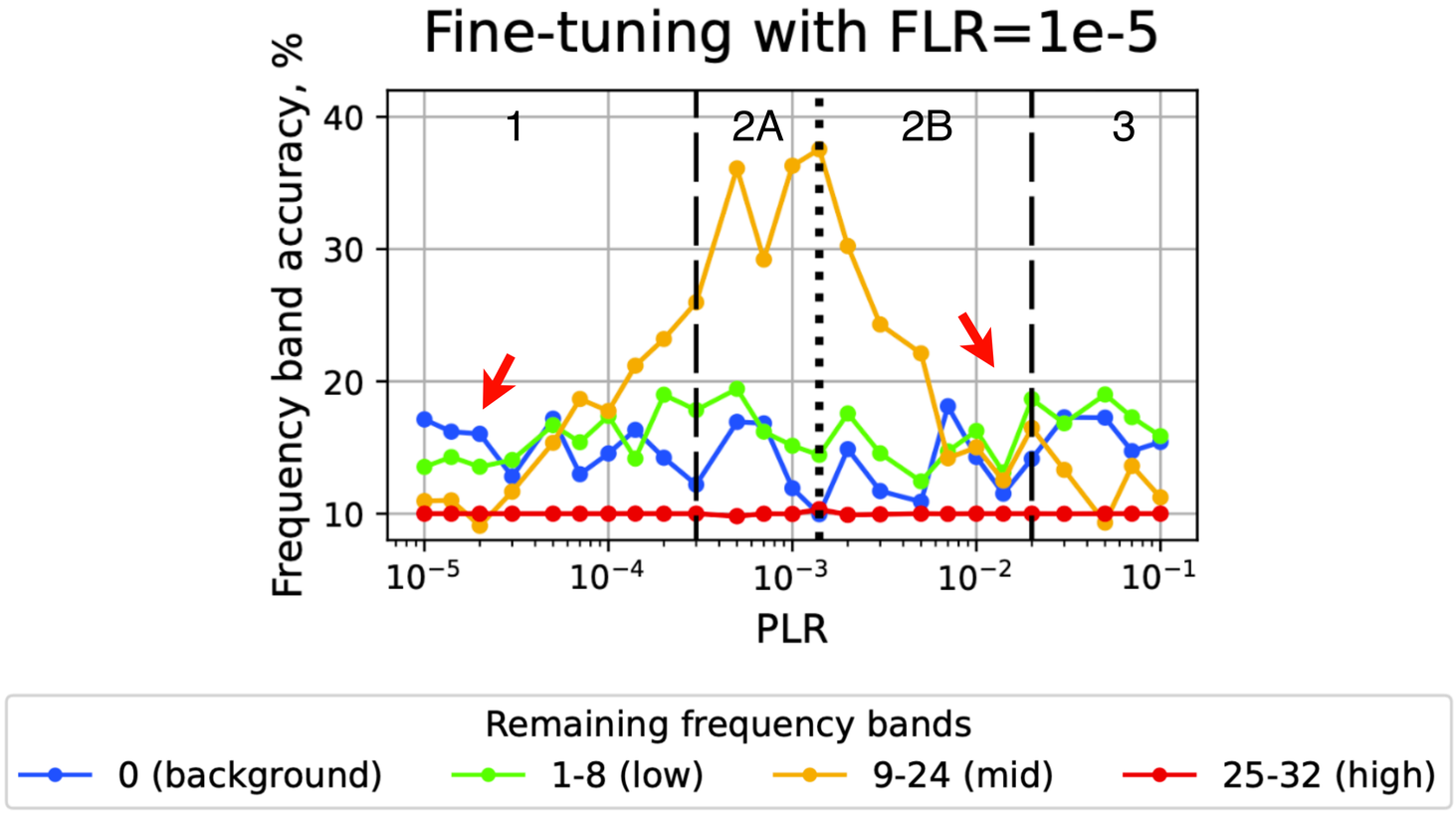


SI ResNet-18, CIFAR-10

PLR from 2A:

- feature sparsity
- prefers mid-frequencies

Feature learning perspective



SI ResNet-18, CIFAR-10

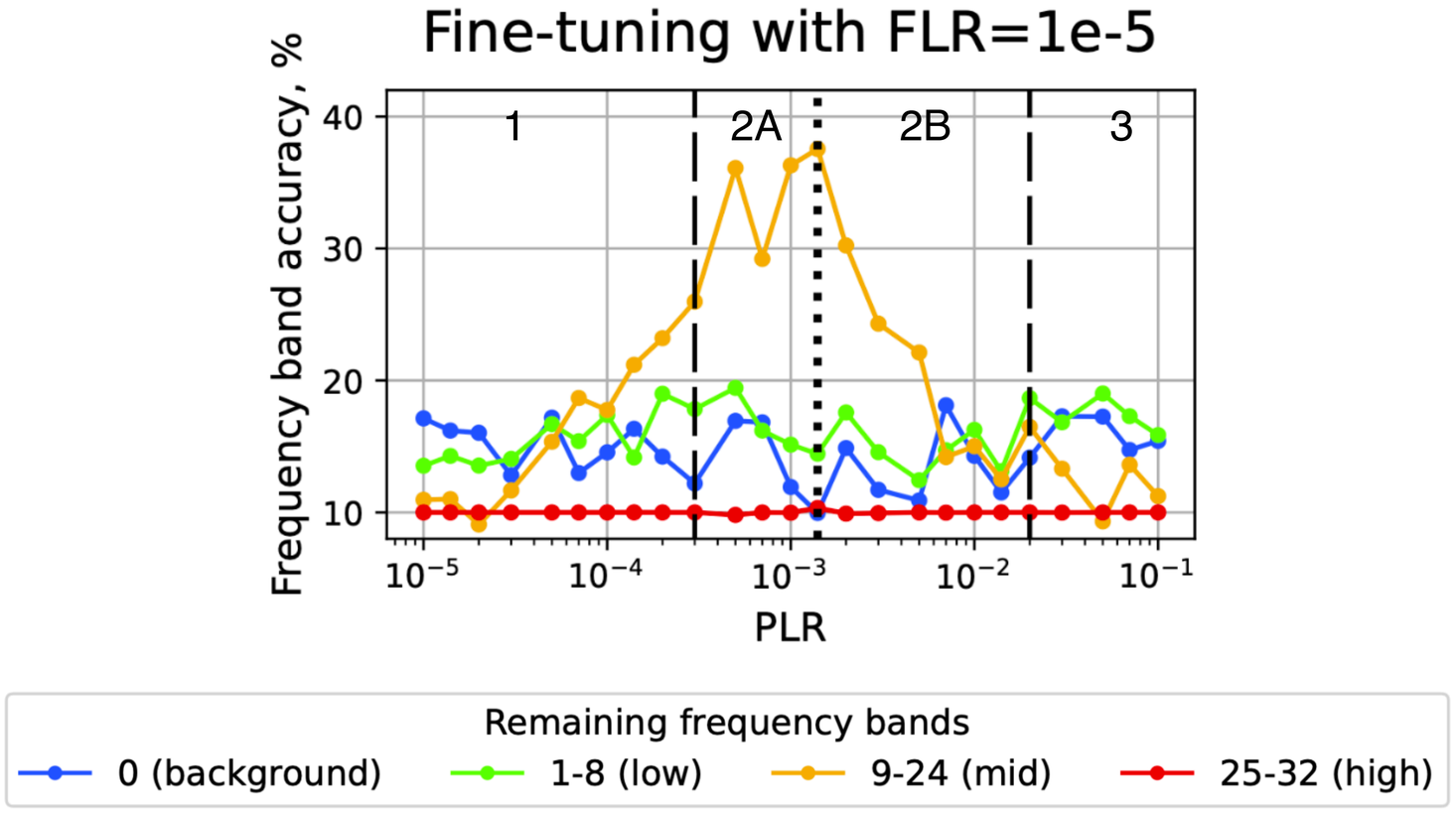
PLR from 2A:

- feature sparsity
- prefers mid-frequencies

Smaller and larger PLR

- no sparsity

Feature learning perspective



SI ResNet-18, CIFAR-10

— *Why is 2A so special?*

— It learns a sparse set of the most relevant features

Conclusion

Best LRs to start training (subregime 2A)

- ✓ narrow range just above convergence threshold
- ✓ locate a convex basin with good solutions
- ✓ learn a sparse set of the most relevant features

arXiv

code

$p(B|A)$yesgroup

Conclusion

Best LRs to start training (subregime 2A)

- ✓ narrow range just above convergence threshold
- ✓ locate a convex basin with good solutions
- ✓ learn a sparse set of the most relevant features

Additional results in paper

- synthetic example with controlled feature learning
- conventional training setups, other architectures/datasets
- Stochastic Weight Averaging (SWA) instead of fine-tuning

arXiv

code

$p(B|A)$yesgroup