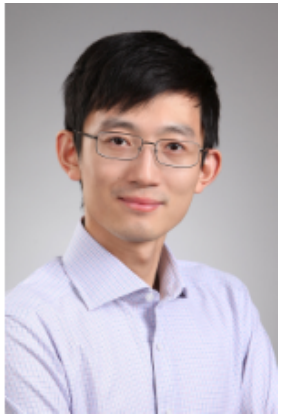
Regularization Effect of Large Initial Learning Rate

Yuanzhi Li*
Carnegie Mellon University

Colin Wei*
Stanford University

Tengyu Ma
Stanford University



Large **Initial** Learning Rate is Crucial for Generalization

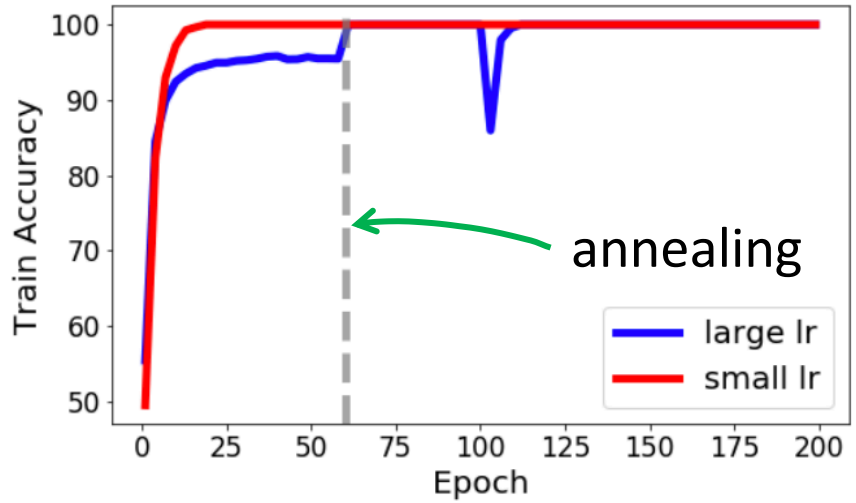
Large **Initial** Learning Rate is Crucial for Generalization

- Common schedule: large initial learning rate + annealing

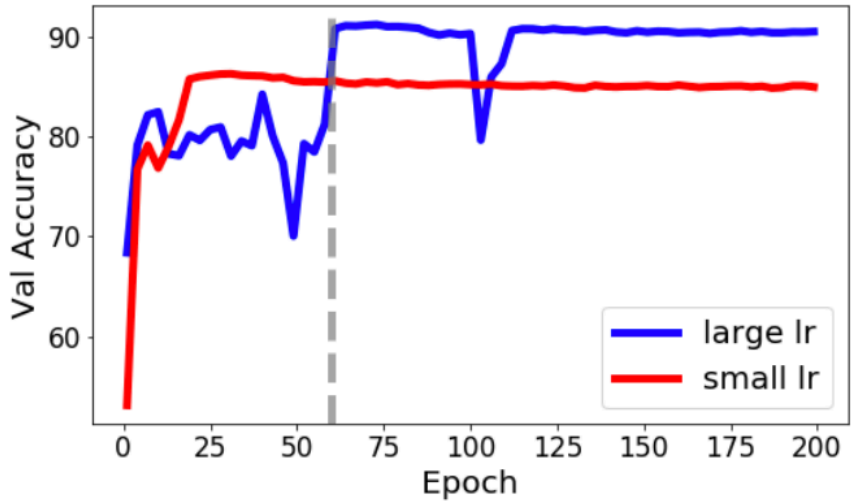
Large **Initial** Learning Rate is Crucial for Generalization

- Common schedule: large initial learning rate + annealing
- … But small learning rate: better train and test performance up until annealing 

Train Accuracy

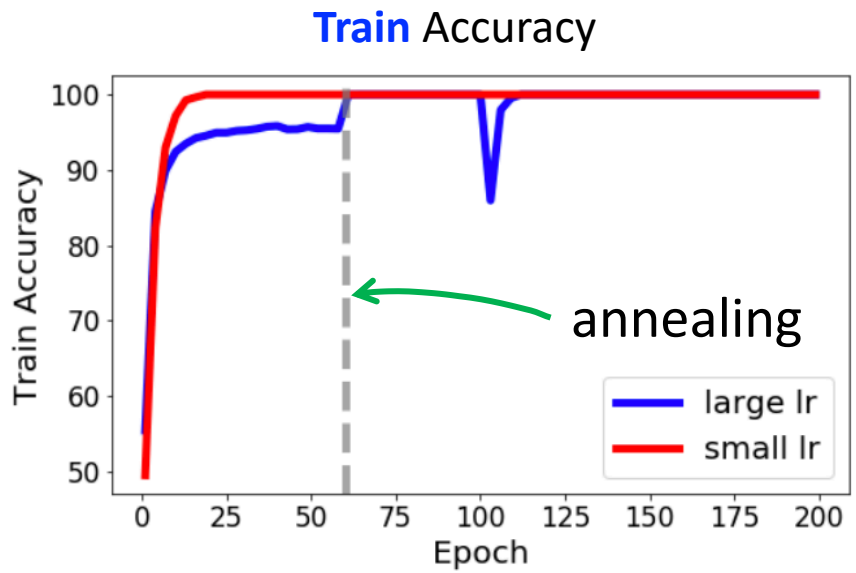


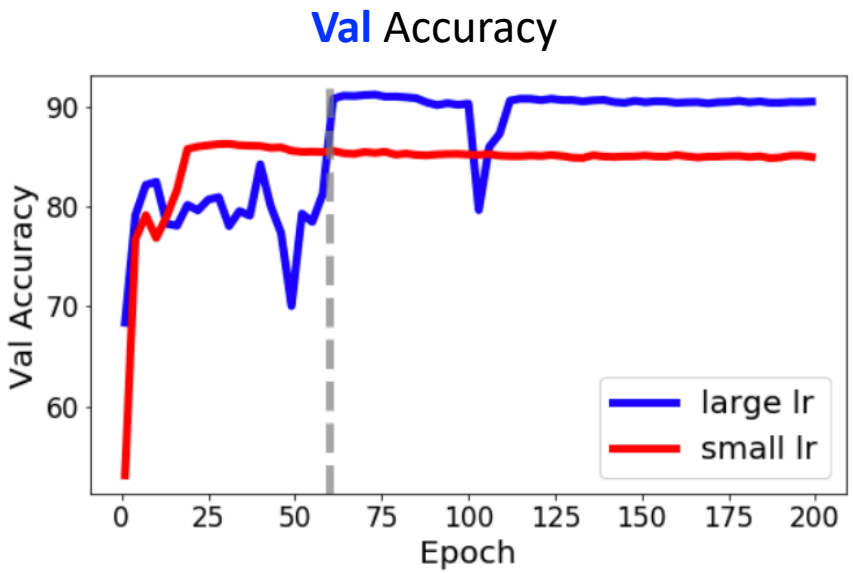
Val Accuracy



Large **Initial** Learning Rate is Crucial for Generalization

- Common schedule: large initial learning rate + annealing
- … But small learning rate: better train and test performance up until annealing





- Large LR outperforms small LR after annealing!

LR schedule changes order of learning patterns => generalization

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing
  - => learns to use all patterns, helping generalization!

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing
  - => learns to use all patterns, helping generalization!

- Intuition: larger LR
  - ⇒ larger noise in activations

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing
  - => learns to use all patterns, helping generalization!

- Intuition: larger LR
  - ⇒ larger noise in activations
  - ⇒ effectively weaker representational power

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing
  - => learns to use all patterns, helping generalization!

- Intuition: larger LR
  - ⇒ larger noise in activations
  - ⇒ effectively weaker representational power
  - ⇒ won’t overfit to “signatures”

LR schedule changes order of learning patterns => generalization

- Small LR quickly memorizes **hard-to-fit** “class signatures”
  - Ignores other patterns, harming generalization
- Large initial LR + annealing learns **easy-to-fit** patterns first
  - Only memorizes hard-to-fit patterns after annealing
  - => learns to use all patterns, helping generalization!

- Intuition: larger LR
  - ⇒ larger noise in activations
  - ⇒ effectively weaker representational power
  - ⇒ won’t overfit to “signatures”

- Non-convexity is crucial: different LR schedules find different solutions

Demonstration on Modified CIFAR10

Demonstration on Modified CIFAR10

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

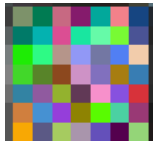
original image

Demonstration on Modified CIFAR10

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

original image



hard-to-fit patch indicating class

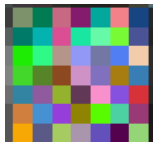
Demonstration on Modified CIFAR10

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

Group 3: 60% examples with both patterns

original image



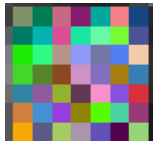
hard-to-fit patch indicating class

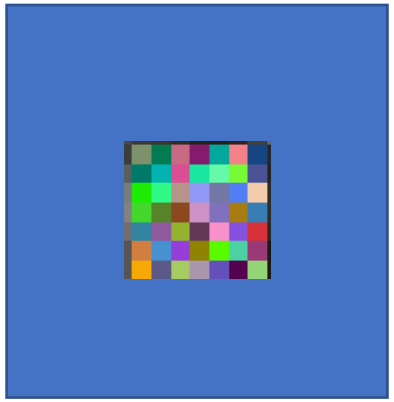
Demonstration on Modified CIFAR10

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

Group 3: 60% examples with both patterns





original image

hard-to-fit patch indicating class

- Small LR memorizes patch, **ignores** rest of the image
  - ⇒ learns image from **20%** examples

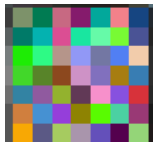
Demonstration on Modified CIFAR10

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

Group 3: 60% examples with both patterns

original image



hard-to-fit patch indicating class

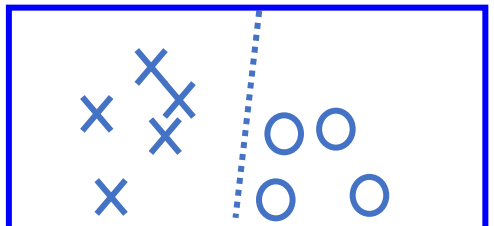
- Small LR memorizes patch, **ignores** rest of the image
  - ⇒ learns image from **20%** examples
- Large initial LR initially ignores patch, only learns it after annealing
  - ⇒ learns image from **80%** examples

Theoretical Setting

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

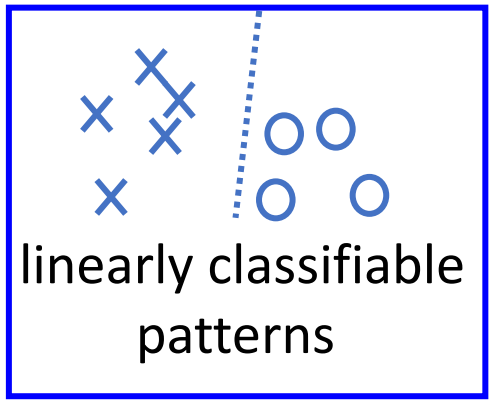
Group 3: 60% examples with both patterns



linearly classifiable patterns

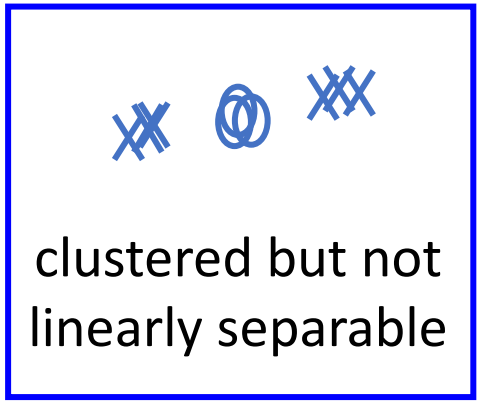
Theoretical Setting

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns



linearly classifiable patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

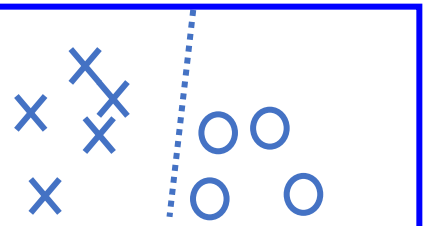


clustered but not linearly separable

Group 3: 60% examples with both patterns

Theoretical Setting

Group 1: 20% examples with hard-to-generalize, easy-to-fit patterns



linearly classifiable patterns

Group 2: 20% examples with easy-to-generalize, hard-to-fit patterns

clustered but not linearly separable

Group 3: 60% examples with both patterns

Contains both patterns

Conclusion

Conclusion

- Small LR optimizes faster, but generalizes worse than large initial LR + annealing

Conclusion

- Small LR optimizes faster, but generalizes worse than large initial LR + annealing
- Explanation: order of learning pattern types
  - Easy-to-generalize, hard-to-fit patterns
  - Hard-to-generalize, easy-to-fit patterns

Conclusion

- Small LR optimizes faster, but generalizes worse than large initial LR + annealing
- Explanation: order of learning pattern types
  - Easy-to-generalize, hard-to-fit patterns
  - Hard-to-generalize, easy-to-fit patterns
- SGD noise from large LR is mechanism for regularization

Conclusion

- Small LR optimizes faster, but generalizes worse than large initial LR + annealing
- Explanation: order of learning pattern types
  - Easy-to-generalize, hard-to-fit patterns
  - Hard-to-generalize, easy-to-fit patterns
- SGD noise from large LR is mechanism for regularization

Come find our poster: 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #144!