Towards Understanding the Data Dependency of Mixup-style Training

Muthu Chidambaram[1] Xiang Wang[1] Yuzheng Hu[2] Chenwei Wu[1] Rong Ge[1]

[1]Duke University

[2]University of Illinois Urbana-Champaign

Background on Mixup

- **Mixup** [ZCDLP18] is a data augmentation technique that uses random convex combinations of the original data points and their labels for training.

$$(x_i, y_i),\ (x_j, y_j) \longrightarrow \big(\lambda x_i + (1-\lambda)x_j, \lambda y_i + (1-\lambda)y_j\big), \quad \lambda \sim \mathbb{P}_f$$

Background on Mixup

- **Mixup** [ZCDLP18] is a data augmentation technique that uses random convex combinations of the original data points and their labels for training.

$$(x_i, y_i),\ (x_j, y_j) \longrightarrow \big(\lambda x_i + (1-\lambda)x_j, \lambda y_i + (1-\lambda)y_j\big), \quad \lambda \sim \mathbb{P}_f$$

Benefits and Surprising Facts of Mixup Training

- Mixup training has been empirically shown to improve **generalization** and **robustness** [ZCDLP18, HZZ+19, ZDK+20, LVKB19] while still minimizing the training error on the original data.

Benefits and Surprising Facts of Mixup Training

- Mixup training has been empirically shown to improve **generalization** and **robustness** [ZCDLP18, HZZ$^+$19, ZDK$^+$20, LVKB19] while still minimizing the training error on the original data.
- **Question 1:** When and why does Mixup minimize the original risk?

Benefits and Surprising Facts of Mixup Training

- Mixup training has been empirically shown to improve **generalization** and **robustness** [ZCDLP18, HZZ$^+$19, ZDK$^+$20, LVKB19] while still minimizing the training error on the original data.
- **Question 1:** When and why does Mixup minimize the original risk?
- **Question 2:** Does Mixup always improve generalization?

Main Contributions

A **Relationship between Mixup and Empirical Risk Minimization (ERM)**

Main Contributions

A. **Relationship between Mixup and Empirical Risk Minimization (ERM)**
   1. Template for dataset on which Mixup fails to classify original data correctly

Main Contributions

A. **Relationship between Mixup and Empirical Risk Minimization (ERM)**
   1. Template for dataset on which Mixup fails to classify original data correctly
   2. Sufficient conditions for Mixup to also minimize original risk

Main Contributions

A. **Relationship between Mixup and Empirical Risk Minimization (ERM)**
   1. Template for dataset on which Mixup fails to classify original data correctly
   2. Sufficient conditions for Mixup to also minimize original risk

B. **When and why Mixup improves generalization**

Main Contributions

A. **Relationship between Mixup and Empirical Risk Minimization (ERM)**
   1. Template for dataset on which Mixup fails to classify original data correctly
   2. Sufficient conditions for Mixup to also minimize original risk

B. **When and why Mixup improves generalization**
   1. Data and model conditions under which Mixup does not improve generalization

Main Contributions

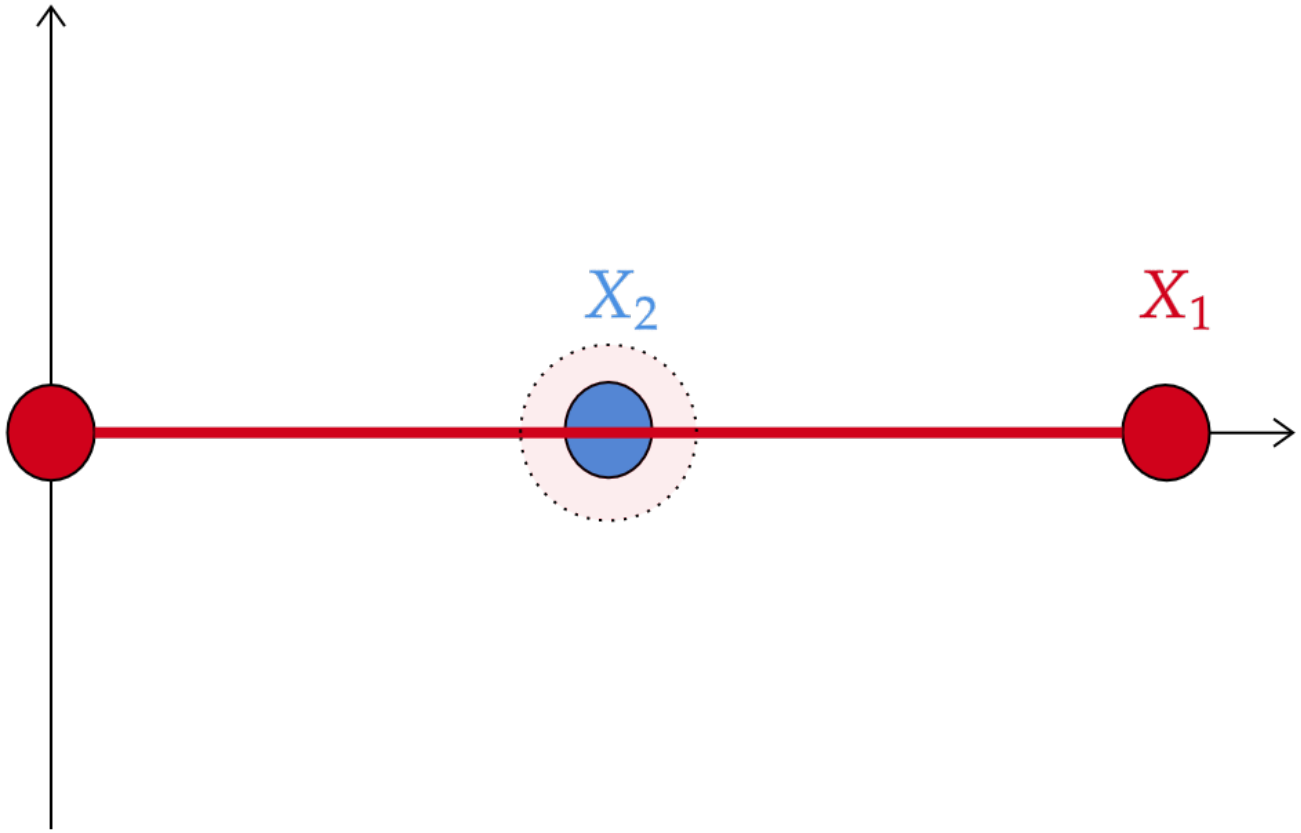
A. **Relationship between Mixup and Empirical Risk Minimization (ERM)**
   1. Template for dataset on which Mixup fails to classify original data correctly
   2. Sufficient conditions for Mixup to also minimize original risk

B. **When and why Mixup improves generalization**
   1. Data and model conditions under which Mixup does not improve generalization
   2. Margin properties of Mixup that may explain why Mixup has better generalization in other cases

Mixup and Empirical Risk Minimization

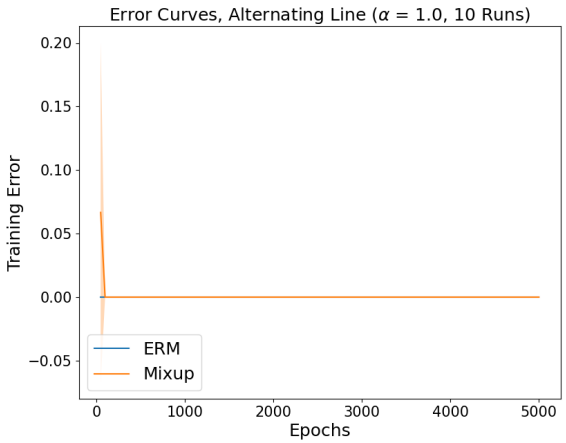
- **Relationship between Mixup and Empirical Risk Minimization (ERM)**

A Mixup "Counterexample"

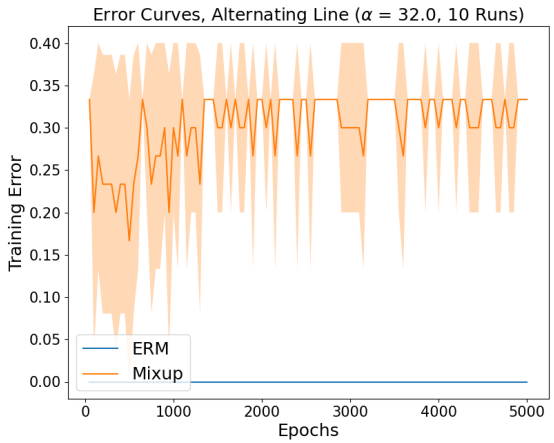


Empirical Performance on Counterexample

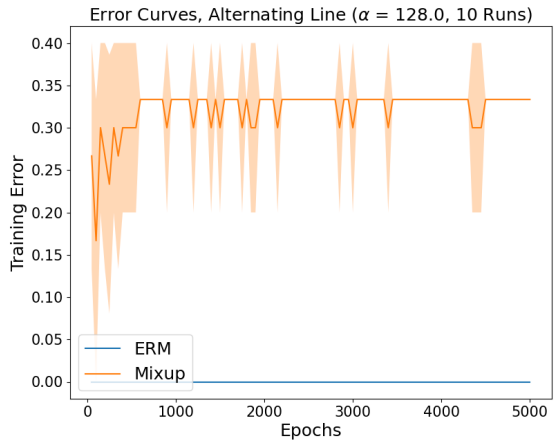
We train a 2-layer ReLU network on the previous dataset mixing using Beta(α, α), results shown below.



(a) $\alpha = 1$



(b) $\alpha = 32$



(c) $\alpha = 128$

Sufficient Condition for Mixup to Minimize Original Risk

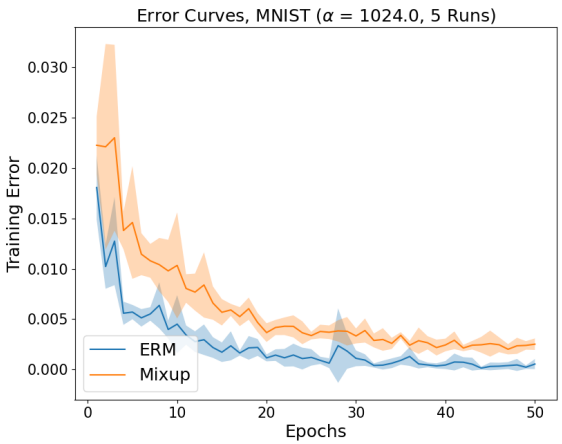
On the other hand, when **inter-class collinearity is not present**, Mixup minimizes the original risk.

Theorem 1 (Theorem 2.10 in Paper, Informal)

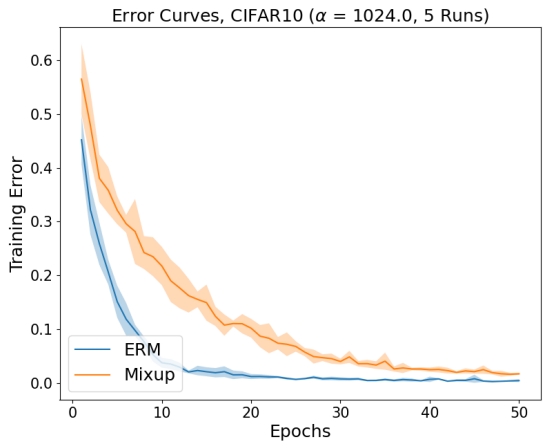
For each point x in a class i, if x does not fall on a line segment which has an endpoint in a class j for $j \neq i$, then the Mixup-optimal classifier will also minimize the original risk.

Empirical Results on Image Classification Benchmarks

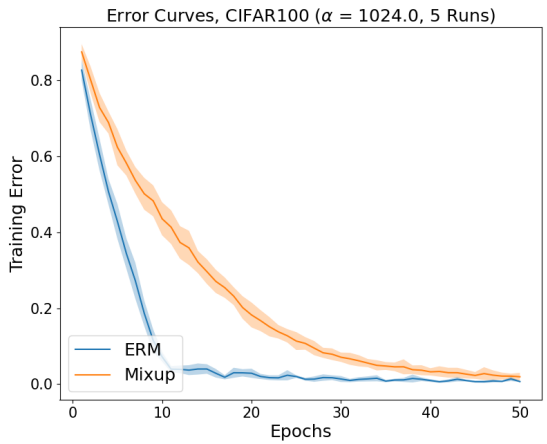
Empirical results for ResNet-18 [HZRS15] on MNIST, CIFAR-10, and CIFAR-100 while using Mixup with Beta$(1024, 1024)$.



(a) MNIST



(b) CIFAR-10



(c) CIFAR-100

Mixup and Generalization

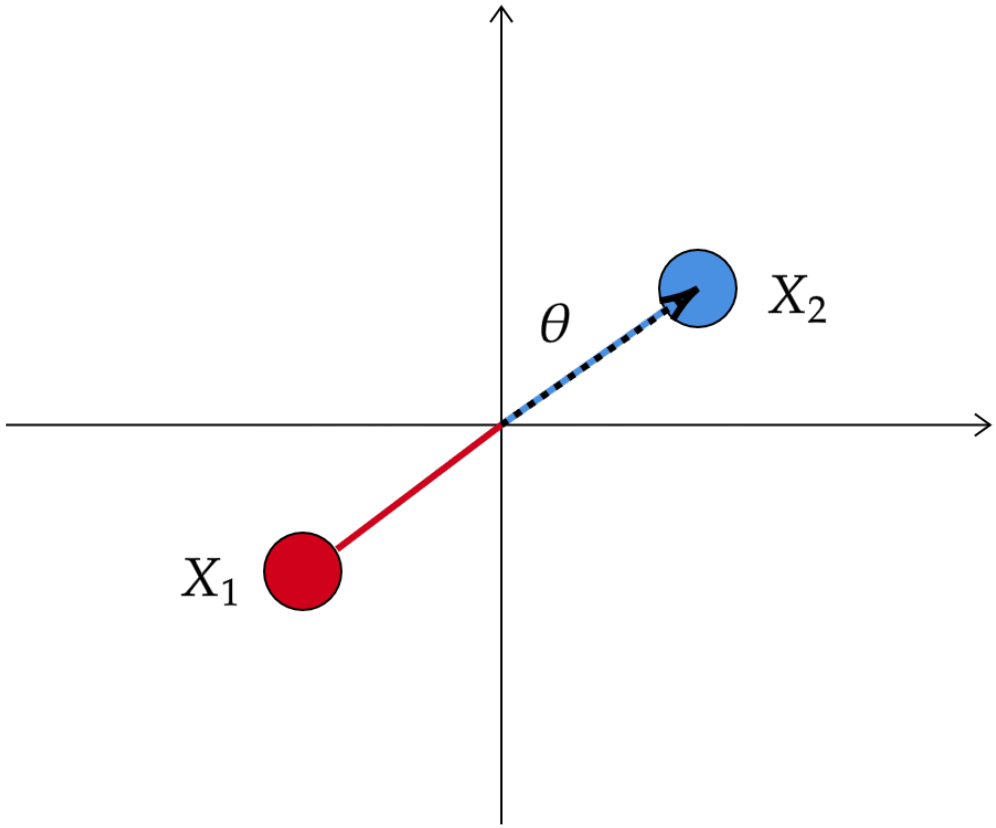
- **When and why Mixup improves generalization**

When Mixup Generalizes Identically to ERM

- For linear models and binary classification datasets for which all data points are support vectors, Mixup and ERM trained with gradient descent **learn the same classifier**.

When Mixup Generalizes Identically to ERM

- For linear models and binary classification datasets for which all data points are support vectors, Mixup and ERM trained with gradient descent **learn the same classifier**.



Mixup Margin Characterization

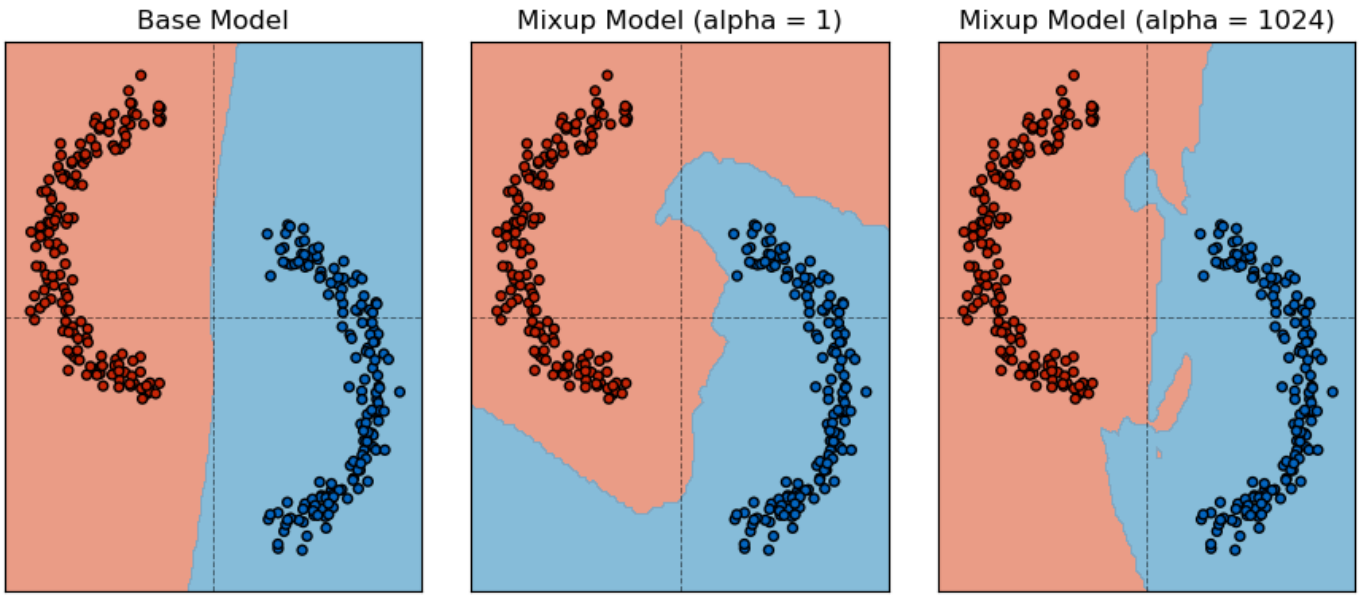
In other cases, however, Mixup can have better **margin properties**.

Theorem 2 (Theorem 3.2 in Paper, Informal)

If for a point x we have that x falls closer to the data points in class i than any other class for every line segment between points in the dataset that intersect it, and we have that x does not fall on any line segments between points in classes j, q for $j, q \neq i$, then x will be classified as class i by the Mixup-optimal classifier.

Mixup Decision Boundaries on Two Moons

- Here we compare standard training to Mixup with Beta$(1, 1)$ and Beta$(1024, 1024)$ on the two moons dataset.



Summary

- Collinearity can cause problems for Mixup minimizing the original risk on a dataset

Summary

- Collinearity can cause problems for Mixup minimizing the original risk on a dataset
- Mixup can have better margin properties than ERM, but there exist cases where this need not be true

References I


Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, 2015.


Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li, *Bag of tricks for image classification with convolutional neural networks*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.


Alex Lamb, Vikas Verma, Juho Kannala, and Yoshua Bengio, *Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy*, Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 2019, pp. 95–103.


Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, *mixup: Beyond empirical risk minimization*, International Conference on Learning Representations, 2018.

References II


Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou, *How does mixup help with robustness and generalization?*, 2020.