# Over-Training with Mixup May Hurt Generalization

Zixuan Liu[1] Ziqiao Wang[1] Hongyu Guo[1,2] Yongyi Mao[1]

[1]University of Ottawa, [2]National Research Council Canada

National Research Council Canada

uOttawa

## Summary

**Novel Observation**

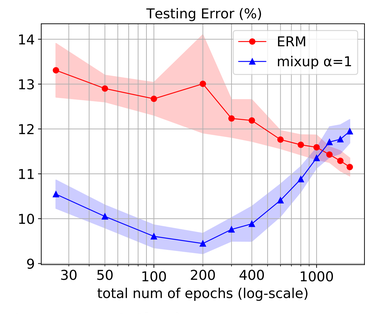
- Over-training with Mixup causes U-shaped test error curve.

**Explanation**

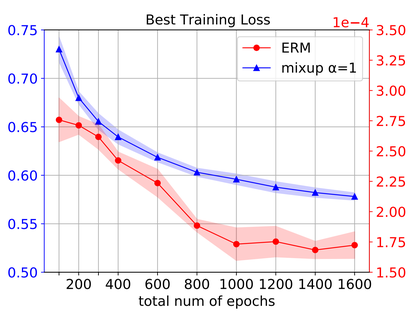
- Mixup induces label noise.
- Overfitting to noise occcurs in over-training.

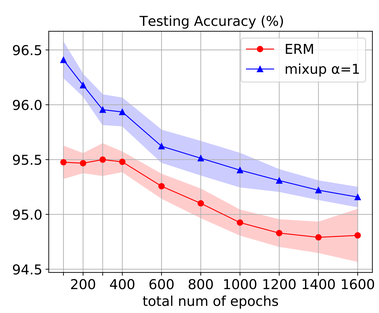
## Observations

**As the training loss continuously decays (left), the testing error first decreases then increases (right).**

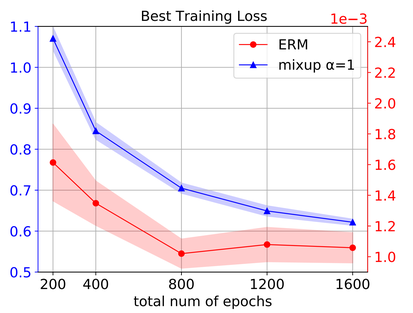



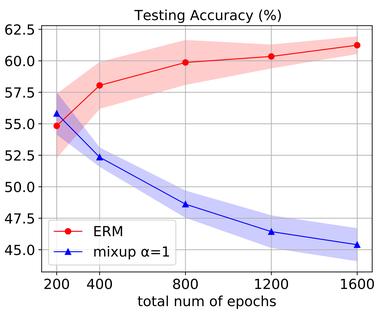

**ResNet18 on CIFAR10 (w/o augmentation)**




**ResNet18 on SVHN (w/o augmentation)**




**ResNet34 on CIFAR100 (w/o augmentation)**

## Mixup Induces Label Noises

**Theorem 1**

For $\widetilde{X} = \lambda X + (1-\lambda)X'$ with a fixed $\lambda \in [0,1]$, the probability of assigning a noisy label is lower bounded by

$$Pr(\widetilde{Y}_{\mathrm{h}} \neq \widetilde{Y}^*_{\mathrm{h}} | \widetilde{X}) \geq \frac{1}{2} \sup_{j \in \mathcal{Y}} \left| f_j(\widetilde{X}) - [(1-\lambda) f_j(X) + \lambda f_j(X')] \right|.$$

**Remark**:

Mixup induces label noises as long as the ground-truth function $f$ is not target-linear.

## Dynamics of Learning

**Lemma 1**:

Consider a least squares regression problem training random feature model $\theta^T \phi(X)$:

$$\theta_t - \theta^* = (\theta_0 - \theta^*) e^{-\frac{\eta}{m}\widetilde{\Phi}\widetilde{\Phi}^T t} + (\mathbf{I}_d - e^{-\frac{\eta}{m}\widetilde{\Phi}\widetilde{\Phi}^T t})\theta^{\mathrm{noise}},$$

where $\theta^* = \widetilde{\Phi}^\dagger \widetilde{\mathbf{Y}}^*$ and $\theta^{\mathrm{noise}} = \widetilde{\Phi}^\dagger \mathbf{Z}$.

**Remarks**:

- In the early phase: $\theta_t \to \theta^*$.
- In the latter phase: $\theta_t \to \theta^* + \theta^{\mathrm{noise}}$.

**Theorem 2**:

Assume $\theta_0 \sim \mathcal{N}(0, \xi^2 \mathrm{I}_d)$, $C_1, C_2 > 0$, then:

$$R_t - R^* \leq C_1 \sum_{k=1}^{d} \left[ \left(\xi_k^2 + \theta_k^{*2}\right) e^{-2\eta\mu_k t} + \frac{C_2}{\mu_k} \left(1 - e^{-\eta\mu_k t}\right)^2 \right] + 2\sqrt{C_1 R^* \zeta},$$

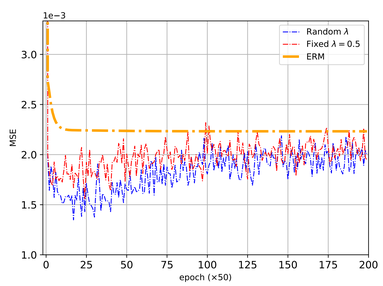
where $R^* = \mathbb{E}_{X,Y} \left|\left| Y - \theta^{*T}\phi(X) \right|\right|_2^2$, $\zeta = \sum_{k=1}^{d} \max\{\xi_k^2 + \theta_k^{*2}, \frac{C_2}{\mu_k}\}$ and $\mu_k$ is the $k^{\mathrm{th}}$ eigenvalue of the matrix $\frac{1}{m}\widetilde{\Phi}\widetilde{\Phi}^T$.

**Remark**:

- RHS first decreases then increases.

## Experimental Verification

- **Teacher** network: provides ground-truth training targets.
- **Student** network: be trained as random feature model.



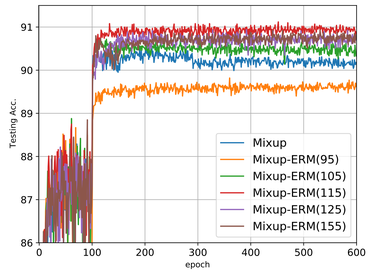

**Fixing $\lambda$ = 0.5 increases the severity of label noises**

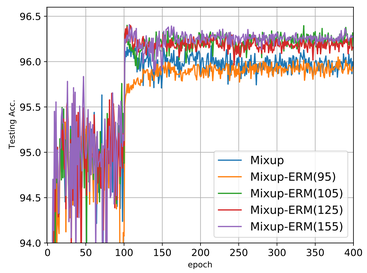
⇓

**Turning point presents earlier**

## Training Mode Switch (Mixup → ERM)

**Switch off Mixup at a proper early epoch avoids genralization degradation.**

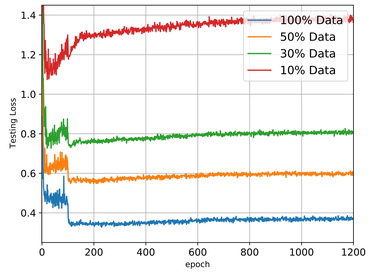


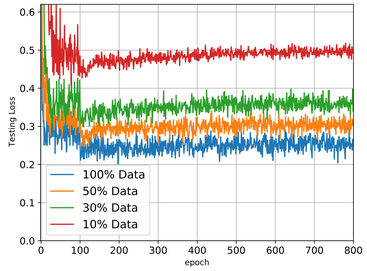




**Left**: CIFAR10; **Right**: SVHN; **Numbers in brackets**: switching epoch.

## Impact of Data Size on U-shaped Curve

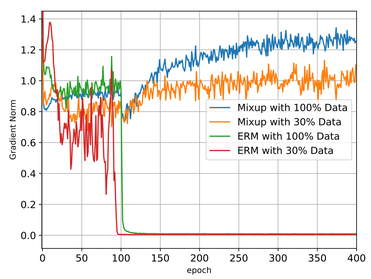
**Larger dataset postpones the turning point to present.**

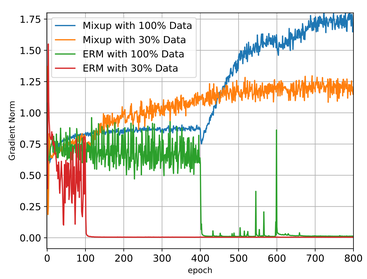



**Left**: CIFAR10; **Right**: SVHN.

## Gradient Norm in Mixup Training Does Not Vanish




**Left**: CIFAR10; **Right**: SVHN.

NEURAL INFORMATION PROCESSING SYSTEMS