Over-Training with Mixup May Hurt Generalization

Zixuan Liu 1  Ziqiao Wang 1  Hongyu Guo 1,2  Yongyi Mao 1
1University of Ottawa, 2National Research Council Canada

Summary

Novel Observation
• Over-training with Mixup causes U-shaped test curve.

Explanation
• Mixup induces label noise.
• Overfitting to noise occurs in over-training.

Observations

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

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

$$Pr[Y \neq Y'(X)] \geq \frac{1}{2\lambda} \left| f_j(X) - \frac{1 - \lambda}{\lambda} f_j(X') \right|.$$ 

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

Mixup Induces Label Noises

Dynamics of Learning

Lemma 1:
Consider a least squares regression problem training random feature model $\theta^T \phi(X)$:

$$\theta_t - \theta^* = (\theta_t - \theta^*) e^{-\lambda_t \phi} + \sum_{j \in J} \theta_j e^{-\lambda_t \phi} - \sum_{j \in J} \theta_j \phi - \sum_{j \in J} \theta_j \phi = \hat{V} + \hat{Z},$$

where $\theta^* = \hat{V}' \hat{Y}'$ and $\theta^{(p)} = \hat{V}' \hat{Z}$.

Remarks:
• In the early phase: $\theta_t \rightarrow \theta^*$.
• In the latter phase: $\theta_t \rightarrow \theta^{(p)}$.

Theorem 2:
Assume $\theta_t \rightarrow N(0, \epsilon^2)\phi$, $C_1, C_2 > 0$, then:

$$\begin{align*}
\|R_t - R^*\| &\leq C_1 \sum_{k=1}^d \left( \epsilon^2 \phi \lambda_t^{(k)} \right) e^{-\lambda_t \phi} + C_2 \max \left( \epsilon^{2(1-\lambda_k)} \phi \lambda_t^{(k)} \right) + 2 \sqrt{C_1 C_2},
\end{align*}$$

where $R_t = E_{Y \in \mathcal{X}} \left[ \|Y - \theta^T \phi(X)\|_2^2 \right]$, $\lambda_t = \sum_{j \in J} max(\lambda^j + \theta_j^2 \phi)$, and $\mu_k$ is the $k^{th}$ eigenvalue of the matrix $\lambda \theta_j^2 \phi$.

Remark:
• RHS first decreases then increases.

Experiment Evaluation

Mixup Induces Label Noises

Training Mode Switch (Mixup $\rightarrow$ ERM)

Switch off Mixup at a proper early epoch avoids generalization degradation.

Impact of Data Size on U-shaped Curve

Larger dataset postpones the turning point to present.

Gradient Norm in Mixup Training Does Not Vanish

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

Turning point presents earlier