

# On Mixup Regularization

[Carratino et al., JMLR 2022]

-Summary-

# Introduction – Current paper for mix-up theory

- There are some papers dealing with theoretical analysis of mix-up technique :

## [Brief summary & idea]

### 1. How Does Mixup Help With Robustness And Generalization [Zhang et al., ICLR 2021]

- ① Showed regularization effect of mix-up using Taylor expansion on mix-up loss.
- ② Given adversarial attack size, demonstrated mix up loss is the upper bound of adversarial loss.
- ③ Under two-layer ReLU network setup with some assumption, they showed :  
Decreasing mix-up loss -> Decreasing ERM loss -> improve generalization performance as ERM minimization even using mix up training

# Introduction – Current paper for mix-up theory

## [Brief summary & idea]

### 2. Towards Understanding the Data Dependency of Mixup-style Training

[Chidambaram et al., ICLR 2022]

- ① Showed why mix up technique still works even though model encounter very few true data points during training using theoretical analysis.
- ② Demonstrated that **if collinearity ( = manifold intrusion) of mix up point is expected, then model cannot achieve zero training error even with very long training.**
- ③ On high dimensional dataset [MNIST, CIFAR-10/100], showed empirically that collinearity rarely occurs => reason why training loss -> 0 on CIFAR-10 mixup training  
(By computing minimum distance between each mixup points and points from classes other than the two mixed classes)

# Introduction – Current paper for mix-up theory

## [Brief summary & idea]

### 3. On Mixup Regularization [Carratino et al., JMLR 2022]

- ① Showed Mixup can be written as a perturbed ERM loss.
- ② **Showed regularization effect** of mix-up using Taylor expansion on various training case : Cross entropy loss / logistic regression loss / MSE loss  
(Slight different approach, but involves interesting terms to study)
- ③ Suggested '**Approximated Mixup**' by dropping out the regularization term, which is an intermediate compromise of Mixup and ERM training in the view of regularization.  
(Not our interest...)

**Our current goal** : Want to focus on part ① to study the detailed regularization effect of mix-up and find some reasonable intuition to modify original mix-up.  
(to improve generalization performance)

# Theorems on the paper [On Mixup Regularization]

- Learning problem & Notations:

1. Training set :  $S_n = \{(x_1, y_1), \dots (x_n, y_n)\}$ , and  $x_i \in \mathcal{X} \subset \mathbb{R}^d$ ,  $y_i \in \mathcal{Y} \subset \mathbb{R}^c$

2. Sample mean :  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

3. Sample covariance :  $\Sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})^T$

4. ERM loss :  $\mathcal{E}^{ERM}(f) = \frac{1}{n} \sum_{i=1}^n l(y_i, f(x_i))$

Usually  $\lambda \sim \text{Beta}(\alpha, \alpha)$

5. Mix-up loss :  $\mathcal{E}^{Mixup}(f) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_{\lambda} [l(\lambda y_i + (1 - \lambda)y_j, f(\lambda x_i + (1 - \lambda)x_j))]$

- Now, we reformulate mix-up loss as a perturbed ERM by following procedure.

# Theorems on the paper [On Mixup Regularization]

1. Define  $m_{ij}(\lambda) = l(\lambda y_i + (1 - \lambda)y_j, f(\lambda x_i + (1 - \lambda)x_j))$  to simplify  $\mathcal{E}^{Mixup}(f)$ :

$$\mathcal{E}^{Mixup}(f) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_{\lambda}[m_{ij}(\lambda)]$$

2. Separate  $\lambda$  by symmetry of Beta distribution as follows :

Beta distribution constricted on given domain

$$\lambda = \pi\lambda_0 + (1 - \pi)\lambda_1, \quad \lambda_0 \sim \text{Beta}_{\left[0, \frac{1}{2}\right]}(\alpha, \alpha), \quad \lambda_1 \sim \text{Beta}_{\left[\frac{1}{2}, 1\right]}(\alpha, \alpha), \quad \pi \sim \text{Bern}\left(\frac{1}{2}\right)$$

**Critical note : It is possible to use any pdf symmetric w.r.t 0.5 (restricted on  $[0,1]$ ) for  $\lambda$**

3. Using the fact  $\lambda'_1 := 1 - \lambda_0 \sim \text{Beta}_{\left[\frac{1}{2}, 1\right]}(\alpha, \alpha)$ , we further simplify as below :

$$\mathbb{E}_{\lambda}[m_{ij}(\lambda)] = \mathbb{E}_{\lambda_0, \lambda_1, \pi}[m_{ij}(\pi\lambda_0 + (1 - \pi)\lambda_1)] = \frac{1}{2} [\mathbb{E}_{\lambda'_1}[m_{ji}(\lambda'_1)] + \mathbb{E}_{\lambda_1}[m_{ij}(\lambda_1)]]$$

# Theorems on the paper [On Mixup Regularization]

4. Finally, we can simply  $\mathcal{E}^{Mixup}(f)$  by following :

Recall :  $m_{ij}(\lambda) = l(\lambda y_i + (1 - \lambda)y_j, f(\lambda x_i + (1 - \lambda)x_j))$

$$\mathcal{E}^{Mixup}(f) = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \left[ \mathbb{E}_{\lambda'_1} [m_{ji}(\lambda'_1)] + \mathbb{E}_{\lambda_1} [m_{ij}(\lambda_1)] \right] = \frac{1}{n} \sum_{i=1}^n l_i$$

where  $l_i = \mathbb{E}_{\theta, j} \left[ l(\theta y_i + (1 - \theta)y_j, f(\theta x_i + (1 - \theta)x_j)) \right]$ ,  $\theta \sim \text{Beta}[\frac{1}{2}, 1](\alpha, \alpha)$ ,  $j \sim \text{Unif}([n])$

5. Define  $\tilde{x}_i := \mathbb{E}_{\theta, j} [\theta x_i + (1 - \theta)x_j]$  and  $\tilde{y}_i := \mathbb{E}_{\theta, j} [\theta y_i + (1 - \theta)y_j]$

6. Define  $\delta_i := \theta x_i + (1 - \theta)x_j - \mathbb{E}_{\theta, j} [\theta x_i + (1 - \theta)x_j]$   
 $\epsilon_i := \theta y_i + (1 - \theta)y_j - \mathbb{E}_{\theta, j} [\theta y_i + (1 - \theta)y_j]$

Note :  $\mathbb{E}_{\theta, j} [\delta_i] = \mathbb{E}_{\theta, j} [\epsilon_i] = 0$

5. Then,  $l_i = \mathbb{E}_{\theta, j} [l(\tilde{y}_i + \epsilon_i, f(\tilde{x}_i + \delta_i))]$

# Theorems on the paper [On Mixup Regularization]

- Summarizing this result, we get following theorem :

**Theorem 1** *Let  $\theta \sim \text{Beta}_{[\frac{1}{2}, 1]}(\alpha, \alpha)$  and  $j \sim \text{Unif}([n])$  be two random variables with  $\alpha > 0$ ,  $n > 0$  and let  $\bar{\theta} = \mathbb{E}_{\theta}\theta$ . For any training set  $\mathcal{S}_n$ , let  $(\tilde{x}_i, \tilde{y}_i)$  for any  $i \in [n]$  be the modified input/output pair given by*

$$\begin{cases} \tilde{x}_i &= \bar{x} + \bar{\theta}(x_i - \bar{x}) , \\ \tilde{y}_i &= \bar{y} + \bar{\theta}(y_i - \bar{y}) , \end{cases}$$

*and  $(\delta_i, \varepsilon_i)$  be the random perturbations given by:*

$$\begin{cases} \delta_i &= (\theta - \bar{\theta})x_i + (1 - \theta)x_j - (1 - \bar{\theta})\bar{x} , \\ \varepsilon_i &= (\theta - \bar{\theta})y_i + (1 - \theta)y_j - (1 - \bar{\theta})\bar{y} . \end{cases}$$

*Then for any  $i \in [n]$ ,  $\mathbb{E}_{\theta, j}\delta_i = \mathbb{E}_{\theta, j}\varepsilon_i = 0$ , and for any function  $f \in \mathcal{H}$ ,*

$$\mathcal{E}^{\text{Mixup}}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta, j} \ell(\tilde{y}_i + \varepsilon_i, f(\tilde{x}_i + \delta_i)) .$$

**Mixup loss as perturbed ERM**



# Theorems on the paper [On Mixup Regularization]

- Now, we want to approximate the mix-up loss via 2<sup>nd</sup> order Taylor expansion :

$$\varepsilon_Q^{Mixup}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta,j} [l_Q^{(i)}(\tilde{y}_i + \epsilon_i, f(\tilde{x}_i + \delta_i))]$$

where  $l_Q^{(i)}$  is given as following :

$$\begin{aligned} l_Q^{(i)}(\tilde{y}_i + \varepsilon, f(\tilde{x}_i + \delta)) &= \ell(\tilde{y}_i, f(\tilde{x}_i)) + \nabla_y \ell(\tilde{y}_i, f(\tilde{x}_i)) \varepsilon + \nabla_u \ell(\tilde{y}_i, f(\tilde{x}_i)) \nabla_x f(\tilde{x}_i) \delta \\ &\quad + \frac{1}{2} \left\langle \delta \delta^\top, \nabla f(\tilde{x}_i)^\top \nabla_{uu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \nabla f(\tilde{x}_i) + \nabla_u \ell(\tilde{y}_i, f(\tilde{x}_i)) \nabla^2 f(\tilde{x}_i) \right\rangle \\ &\quad + \frac{1}{2} \left\langle \varepsilon \varepsilon^\top, \nabla_{yy}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \right\rangle + \left\langle \varepsilon \delta^\top, \nabla_{yu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \nabla f(\tilde{x}_i) \right\rangle, \end{aligned}$$

Note :  $\langle A, B \rangle := \text{Tr}(A^\top B)$  (Frobenius inner product)

This term :  $1 \times c$  by  $c \times |x| \times |x|$  calculation (tensor product)  
In detail, it is the same as following :

$$\sum_{k=1}^c \nabla_u \ell(\tilde{y}_i, f(\tilde{x}_i))^{(k)} \nabla^2 f(\tilde{x}_i)^{(k)} \in \mathbb{R}^{|x| \times |x|}$$

# Theorems on the paper [On Mixup Regularization]

- To proceed further, we define 3 covariance related term as belows :

**Lemma 2** *Let  $\bar{\theta}$  and  $\sigma^2$  be respectively the mean and variance of a  $\text{Beta}_{[\frac{1}{2},1]}(\alpha, \alpha)$  distributed random variable, and  $\gamma^2 = \sigma^2 + (1 - \bar{\theta})^2$ . For any  $i \in [n]$ , let*

$$\Sigma_{\widetilde{x}\widetilde{x}}^{(i)} = \frac{\sigma^2(\widetilde{x}_i - \bar{x})(\widetilde{x}_i - \bar{x})^\top + \gamma^2 \Sigma_{\widetilde{x}\widetilde{x}}}{\bar{\theta}^2},$$

$$\Sigma_{\widetilde{y}\widetilde{y}}^{(i)} = \frac{\sigma^2(\widetilde{y}_i - \bar{y})(\widetilde{y}_i - \bar{y})^\top + \gamma^2 \Sigma_{\widetilde{y}\widetilde{y}}}{\bar{\theta}^2},$$

$$\Sigma_{\widetilde{x}\widetilde{y}}^{(i)} = \frac{\sigma^2(\widetilde{x}_i - \bar{x})(\widetilde{y}_i - \bar{y})^\top + \gamma^2 \Sigma_{\widetilde{x}\widetilde{y}}}{\bar{\theta}^2}.$$

*Then, for any  $i \in [n]$ , the random perturbations defined in (6) satisfy*

$$\mathbb{E}_{\theta,j} \delta_i \delta_i^\top = \Sigma_{\widetilde{x}\widetilde{x}}^{(i)}, \quad \mathbb{E}_{\theta,j} \varepsilon_i \varepsilon_i^\top = \Sigma_{\widetilde{y}\widetilde{y}}^{(i)}, \quad \text{and} \quad \mathbb{E}_{\theta,j} \delta_i \varepsilon_i^\top = \Sigma_{\widetilde{x}\widetilde{y}}^{(i)}.$$

# Theorems on the paper [On Mixup Regularization]

- Further cleaning up the terms induces following approximation of Mixup loss :

**Theorem 3** For any twice continuously differentiable loss  $\ell(y, u)$ , the approximate Mixup risk at any twice differentiable  $f \in \mathcal{H}$  satisfies

$$\mathcal{E}_Q^{\text{Mixup}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(\tilde{y}_i, f(\tilde{x}_i)) + R_1(f) + R_2(f) + R_3(f) + R_4(f),$$

where

$$R_1(f) = \frac{1}{2n} \sum_{i=1}^n \left\| \left( \nabla f(\tilde{x}_i) - J^{(i)} \right)^\top \left( \nabla_{uu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \right)^{\frac{1}{2}} \right\|_{\Sigma_{\tilde{x}\tilde{x}}^{(i)}}^2,$$

Note :  $\|A\|_Z^2 := \langle A, ZA \rangle = \text{Tr}(A^T Z A)$   
(Squared frobenius norm with metric  $Z$ )

$$R_2(f) = \frac{1}{2n} \sum_{i=1}^n \left\langle \Sigma_{\tilde{x}\tilde{x}}^{(i)}, \nabla_u \ell(\tilde{y}_i, f(\tilde{x}_i)) \nabla^2 f(\tilde{x}_i) \right\rangle,$$

$$R_3(f) = -\frac{1}{2n} \sum_{i=1}^n \left\| \Sigma_{\tilde{x}\tilde{y}}^{(i)} \nabla_{yu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \left( \nabla_{uu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \right)^{-\frac{1}{2}} \right\|_{\left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1}}^2,$$

$$R_4(f) = \frac{1}{2n} \sum_{i=1}^n \left\langle \Sigma_{\tilde{y}\tilde{y}}^{(i)}, \nabla_{yy}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \right\rangle,$$

**Hard to interpret**, but we can do something  
when loss is Cross-entropy loss

and

$$\forall i \in [n], \quad J^{(i)} = - \left( \nabla_{uu}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \right)^{-1} \nabla_{uy}^2 \ell(\tilde{y}_i, f(\tilde{x}_i)) \Sigma_{\tilde{y}\tilde{x}}^{(i)} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1}.$$

# Theorems on the paper [On Mixup Regularization]

- If we use Cross-entropy loss  $l^{CE}(y, u)$ , we can get following facts :

$$\begin{aligned}\nabla_y \ell^{CE}(y, u) &= -u^\top, \\ \nabla_u \ell^{CE}(y, u) &= (\mathcal{S}(u) - y)^\top \\ \nabla_{yy}^2 \ell^{CE}(y, u) &= \mathbf{0}_c, \\ \nabla_{yu}^2 \ell^{CE}(y, u) &= -\mathbf{I}_c, \\ \nabla_{uu}^2 \ell^{CE}(y, u) &= H(u).\end{aligned}$$

where  $H(u) = \text{diag}(\mathcal{S}(u)) - \mathcal{S}(u)\mathcal{S}(u)^\top \in \mathbb{R}^{c \times c}$  [the Jacobian of softmax function ( $\mathcal{S}$ )]

- By putting these terms into previous theorem, we achieve following approximate mix up loss on Cross entropy loss setup

# Theorems on the paper [On Mixup Regularization]

**Corollary 4** Let  $\mathcal{S} : \mathbb{R}^c \rightarrow \mathbb{R}^c$  be the softmax operator, i.e., for any  $i \in [c]$  and  $u \in \mathbb{R}^c$ ,  $\mathcal{S}(u)_i = e^{u_i} / \sum_{j=1}^c e^{u_j}$ , and let  $H(u) = \text{diag}(\mathcal{S}(u)) - \mathcal{S}(u)\mathcal{S}(u)^\top \in \mathbb{R}^{c \times c}$ . The approximate Mixup risk for the cross-entropy loss satisfies

$$\mathcal{E}_Q^{\text{Mixup}}(f) = \frac{1}{n} \sum_{i=1}^n \ell^{\text{CE}}(\tilde{y}_i, f(\tilde{x}_i)) + R_1^{\text{CE}}(f) + R_2^{\text{CE}}(f) + R_3^{\text{CE}}(f),$$

where

$$R_1^{\text{CE}}(f) = \frac{1}{2n} \sum_{i=1}^n \left\| \left( \nabla f(\tilde{x}_i) - J^{(i)} \right)^\top H(f(\tilde{x}_i))^{-\frac{1}{2}} \right\|_{\Sigma_{\tilde{x}\tilde{x}}^{(i)}}^2,$$

$$R_2^{\text{CE}}(f) = \frac{1}{2n} \sum_{i=1}^n \left\langle \Sigma_{\tilde{x}\tilde{x}}^{(i)}, \left( \mathcal{S}(f(\tilde{x}_i)) - \tilde{y}_i \right)^\top \nabla^2 f(\tilde{x}_i) \right\rangle,$$

$$R_3^{\text{CE}}(f) = -\frac{1}{2n} \sum_{i=1}^n \left\| \Sigma_{\tilde{x}\tilde{y}}^{(i)} H(f(\tilde{x}_i))^{-\frac{1}{2}} \right\|_{\left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1}}^2,$$

This term is related with EL2N score.

=> Let's try upper bound analysis for each term

with

$$\forall i \in [n], \quad J^{(i)} = H(f(\tilde{x}_i))^{-1} \Sigma_{\tilde{y}\tilde{x}}^{(i)} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1}.$$

# Analysis beyond the paper

1.  $R_1^{CE}(f)$  analysis : focus on  $\left\| (\nabla f(\tilde{x}_i) - J^{(i)})^T H(f(\tilde{x}_i))^{\frac{1}{2}} \right\|_{\Sigma_{\tilde{x}\tilde{x}}^{(i)}}^2$  term

2. Note the following :

$$\begin{aligned} \left\| (\nabla f(\tilde{x}_i) - J^{(i)})^T H(f(\tilde{x}_i))^{\frac{1}{2}} \right\|_{\Sigma_{\tilde{x}\tilde{x}}^{(i)}}^2 &= \text{Tr} \left[ H(f(\tilde{x}_i))^{\frac{1}{2}} (\nabla f(\tilde{x}_i) - J^{(i)}) \Sigma_{\tilde{x}\tilde{x}}^{(i)} (\nabla f(\tilde{x}_i) - J^{(i)})^T H(f(\tilde{x}_i))^{\frac{1}{2}} \right] \\ &= \text{Tr} \left[ \Sigma_{\tilde{x}\tilde{x}}^{(i)} (\nabla f(\tilde{x}_i) - J^{(i)})^T H(f(\tilde{x}_i)) (\nabla f(\tilde{x}_i) - J^{(i)}) \right] \\ &\leq \text{Tr} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right) \cdot \text{Tr} \left( H(f(\tilde{x}_i)) \right) \cdot \text{Tr} \left[ (\nabla f(\tilde{x}_i) - J^{(i)})^T (\nabla f(\tilde{x}_i) - J^{(i)}) \right] \\ &\leq \text{Tr} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right) \cdot \mathcal{H} \left( S(f(\tilde{x}_i)) \right) \cdot \text{Tr} \left[ (\nabla f(\tilde{x}_i) - J^{(i)})^T (\nabla f(\tilde{x}_i) - J^{(i)}) \right] \end{aligned}$$

As the entropy of prediction of  $\tilde{x}_i$  get higher (close to decision boundary ~ Hard example), It gets stronger regularization from mix-up

$R_1^{CE}$  terms forces the Jacobian of  $f(\tilde{x}_i)$  to be close to  $J^{(i)}$ , which is interpreted as weighted Multivariate OLS estimator in this paper (MLOS ???)

# Analysis beyond the paper

3.  $R_2^{CE}(f)$  analysis : focus on  $\langle \Sigma_{\tilde{x}\tilde{x}}^{(i)}, (S(f(\tilde{x}_i)) - \tilde{y}_i)^T \nabla^2 f(\tilde{x}_i) \rangle$  term

4. Note the following :

$$\begin{aligned} \langle \Sigma_{\tilde{x}\tilde{x}}^{(i)}, (S(f(\tilde{x}_i)) - \tilde{y}_i)^T \nabla^2 f(\tilde{x}_i) \rangle &= \text{Tr} \left[ \Sigma_{\tilde{x}\tilde{x}}^{(i)} (S(f(\tilde{x}_i)) - \tilde{y}_i)^T \nabla^2 f(\tilde{x}_i) \right] \\ &\leq \text{Tr} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right) \cdot \text{Tr} \left[ (S(f(\tilde{x}_i)) - \tilde{y}_i)^T \nabla^2 f(\tilde{x}_i) \right] \\ &= \text{Tr} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right) \cdot (S(f(\tilde{x}_i)) - \tilde{y}_i)^T \text{Tr}^{tensor} \left( \nabla^2 f(\tilde{x}_i) \right) \\ &\leq \text{Tr} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right) \|S(f(\tilde{x}_i)) - \tilde{y}_i\|_2 \cdot \|\text{Tr}^{tensor} \left( \nabla^2 f(\tilde{x}_i) \right)\|_2 \end{aligned}$$

where  $\text{Tr}^{tensor}(\nabla^2 f(\tilde{x}_i)) = \begin{bmatrix} \text{Tr}(\nabla^2 f(\tilde{x}_i)^{(1)}) \\ \dots \\ \text{Tr}(\nabla^2 f(\tilde{x}_i)^{(c)}) \end{bmatrix} \in \mathbb{R}^c$

This term is EL2N score of expected mix-up point of  $x_i$  ( $=\tilde{x}_i$ ). As the  $\tilde{x}_i$  gets harder, the regularization effect of  $\tilde{x}_i$  gets stronger.

$R_2^{CE}$  terms forces the input Laplacian of  $f(\tilde{x}_i)^{(k)}$  to be zero. ( $k \in [c]$ )  
[Mustafal et al., ICLM 2020] claims that input hessian regularization improves robustness of model

# Analysis beyond the paper

5.  $R_3^{CE}(f)$  analysis : focus on  $\left\| \Sigma_{\tilde{x}\tilde{y}}^{(i)} H(f(\tilde{x}_i))^{-\frac{1}{2}} \right\|_{(\Sigma_{\tilde{x}\tilde{x}}^{(i)})^{-1}}^2$  term

6. Note the following :

$$\begin{aligned} \left\| \Sigma_{\tilde{x}\tilde{y}}^{(i)} H(f(\tilde{x}_i))^{-\frac{1}{2}} \right\|_{(\Sigma_{\tilde{x}\tilde{x}}^{(i)})^{-1}}^2 &= \text{Tr} \left[ H(f(\tilde{x}_i))^{-\frac{1}{2}} \Sigma_{\tilde{y}\tilde{x}}^{(i)} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1} \Sigma_{\tilde{x}\tilde{y}}^{(i)} H(f(\tilde{x}_i))^{-\frac{1}{2}} \right] \\ &= \text{Tr} \left[ \Sigma_{\tilde{y}\tilde{x}}^{(i)} \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1} \Sigma_{\tilde{x}\tilde{y}}^{(i)} H(f(\tilde{x}_i)) \right] \\ &\leq \text{Tr} \left( \Sigma_{\tilde{y}\tilde{x}}^{(i)} \Sigma_{\tilde{x}\tilde{y}}^{(i)} \right) \text{Tr} \left( \left( \Sigma_{\tilde{x}\tilde{x}}^{(i)} \right)^{-1} \right) \mathcal{H} \left( S(f(\tilde{x}_i)) \right) \end{aligned}$$

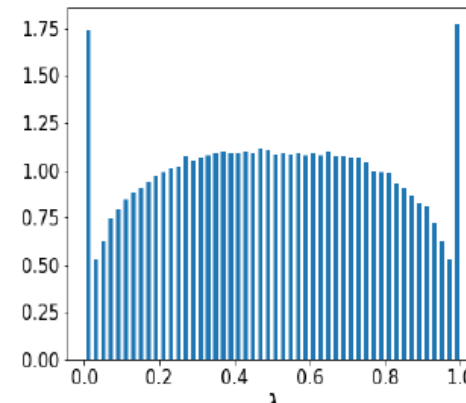
Since  $R_3^{CE}$  term regularize  $\mathcal{H} \left( S(f(\tilde{x}_i)) \right)$ , the regularization effect of  $R_1^{CE}$  ( $\sim$ Jacobian regularization) becomes smaller as the training loss gets minimized

Regularize the entropy of prediction on  $\tilde{x}_i$  and force the entropy to be lower as possible for expected mix-up point for  $x_i$ .



# Towards stronger regularization

- Our natural question on Mix-up is ‘Why we sample  $\lambda$  for  $Beta(\alpha, \alpha)$ ?’ (Obviously, this is the well-known distribution restricted on  $[0,1]$ )



Learned  $\lambda$  distribution from MetaMixup

- Some papers on mix-up shows test accuracy improvement via adopting new  $\lambda$  distribution:
  1. **RegMixup** [Pinto et al., NeurIPS 2022]  
: Propose to mix-up training with original training samples (with high  $\alpha \sim 10$ )
  2. **MetaMixup** [Mai et al., IEEE TNNLS]  
: Using Meta learning to learn good  $\lambda$  selection policy based on input  $x_i$ , demonstrated test accuracy improvement via W shaped  $\lambda$  distribution (learned from meta learning)

# Towards stronger regularization

- Both paper showed the results by experiments, but we may be able to explain this phenomenon using our previous analysis on mix-up.

- Note the following terms appeared in  $R_1^{CE}$ ,  $R_2^{CE}$ ,  $R_3^{CE}$ :  $Tr\left(\Sigma_{\tilde{y}\tilde{x}}^{(i)}\Sigma_{\tilde{x}\tilde{y}}^{(i)}\right)$  or  $Tr\left(\Sigma_{\tilde{x}\tilde{x}}^{(i)}\right)$ :

(Note :  $\gamma^2 = \sigma^2 + (1 - \bar{\theta})^2$  and  $\theta \sim \text{Beta}_{[\frac{1}{2}, 1]}(\alpha, \alpha)$ )

- Recall :

$$\begin{aligned}\Sigma_{\tilde{x}\tilde{x}}^{(i)} &= \frac{\sigma^2(\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^\top + \gamma^2 \Sigma_{\tilde{x}\tilde{x}}}{\bar{\theta}^2}, \\ \Sigma_{\tilde{y}\tilde{y}}^{(i)} &= \frac{\sigma^2(\tilde{y}_i - \bar{y})(\tilde{y}_i - \bar{y})^\top + \gamma^2 \Sigma_{\tilde{y}\tilde{y}}}{\bar{\theta}^2}, \\ \Sigma_{\tilde{x}\tilde{y}}^{(i)} &= \frac{\sigma^2(\tilde{x}_i - \bar{x})(\tilde{y}_i - \bar{y})^\top + \gamma^2 \Sigma_{\tilde{x}\tilde{y}}}{\bar{\theta}^2}.\end{aligned}$$

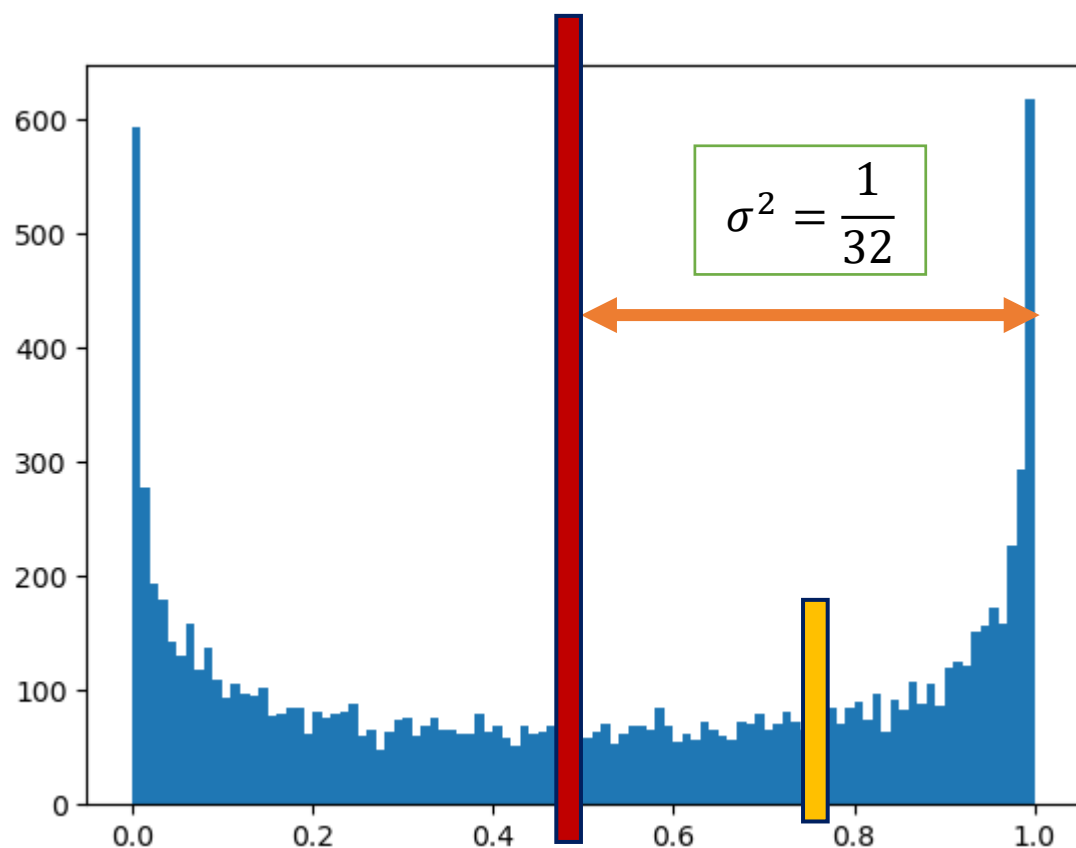
By maximizing,  $\frac{\sigma^2}{\bar{\theta}^2}$  we can enhance the effect of regularization effect induced by all  $R_1^{CE}$ ,  $R_2^{CE}$ ,  $R_3^{CE}$  term

**Note : There is no problem to replace  $\text{Beta}_{[\frac{1}{2}, 1]}$  distribution to any  $[\frac{1}{2}, 1]$  restricted distribution symmetric to  $\frac{1}{2}$ . (This is our idea)**

# Towards stronger regularization

- (Experiment) Let  $\lambda \sim \text{Beta}(0.5, 0.5)$  as in naïve mixup.

$$dbeta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$



In other words,  $p(\theta) = 2 \times dbeta(0.5, 0.5)$ .  
Then,

$$\bar{\theta} = \mathbb{E}[\theta] = \frac{3}{4}$$

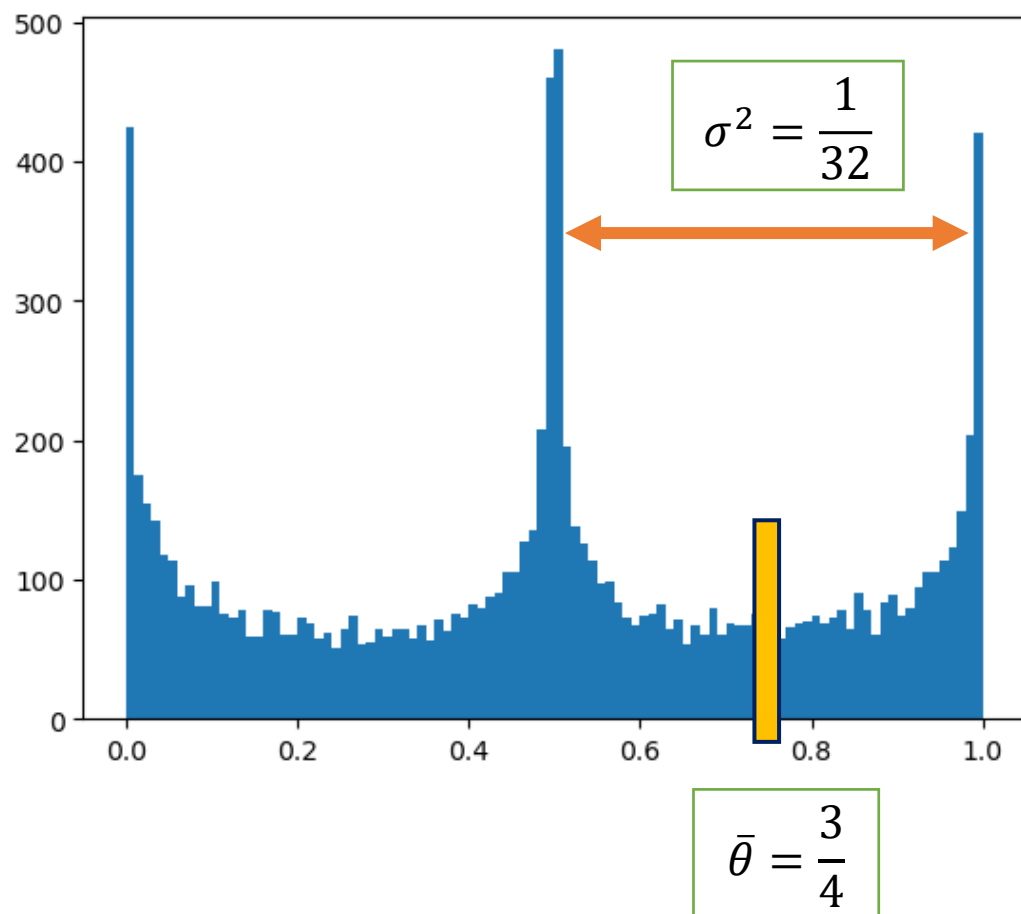
$$\sigma^2 = \text{Var}(\theta) = \frac{1}{4} \text{Var}(X) = \frac{1}{32}$$

Thus,  $\frac{\sigma^2}{\bar{\theta}} = \frac{1}{24}$

Note :  $X \sim \text{Beta}(\alpha, \alpha) \Rightarrow \text{Var}(X) = \frac{1}{4(2\alpha+1)}$

# Towards stronger regularization

- (Experiment) Let  $\lambda|Z = \begin{cases} \frac{1}{2}X & Z = 0 \\ \frac{1}{2}(X + 1) & Z = 1 \end{cases}$ , where  $X \sim \text{Beta}(0.5, 0.5)$ ,  $Z \sim \text{Bern}\left(\frac{1}{2}\right)$



In other words,  $\theta = \frac{1}{2}(X + 1)$ .  
Then,

$$\bar{\theta} = \mathbb{E}[\theta] = \frac{3}{4}$$
$$\sigma^2 = \text{Var}(\theta) = \frac{1}{4} \text{Var}(X) = \frac{1}{32}$$

Thus,  $\frac{\sigma^2}{\bar{\theta}} = \frac{1}{24}$

Note :  $X \sim \text{Beta}(\alpha, \alpha) \Rightarrow \text{Var}(X) = \frac{1}{4(2\alpha+1)}$