How does mix-up help with robustness and generalization?

-Summary-

Introduction

Notation & Result

- Example : $(x_i, y_i) \sim P_{x,y}\ (i.i.d)$, where $x \in \mathbb{R}^p$, $y \in \mathbb{R}^m$ and $z_i = (x_i, y_i)$
- Set of training data : $S = \{(x_i, y_i)\}_{i=1}^n$, where $(x_i, y_i) \sim P_{x,y}$
- Pair of example : Mix-up example : $\tilde{x}_{i,j}(\lambda) = \lambda x_i + (1-\lambda)x_j$, $\tilde{y}_{i,j}(\lambda) = \lambda y_i + (1-\lambda)y_j$
- Standard population loss : $L(\theta) = \mathbb{E}_{z\sim P_{x,y}} l(\theta, z)$
- Standard empirical loss : $L_n^{std}(\theta, S) = \frac{1}{n}\sum_{i=1}^n l(\theta, z_i)$
- Mix-up loss : $L_n^{mix}(\theta, S) = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}_{\lambda \sim D_\lambda}[l(\theta, \tilde{z}_{i,j}(\lambda)]$, where $\lambda \sim D_\lambda = Beta(\alpha, \beta), \alpha > 0, \beta > 0$
- Gradient : $\nabla f_\theta(x)$, $\nabla_\theta f_\theta(x)$ is gradient with respect to x and θ
- Cosine : $\cos(x, y) = \frac{<x,y>}{\|x\|\cdot\|y\|}$
- Empirical Rademacher complexity of a function class $\mathcal{F}$: $\mathcal{R}_S(\mathcal{F}) = Rad(\mathcal{F}, S) = \frac{1}{n}\mathbb{E}_\epsilon[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i)]$

Result: Mix-up training minimizes an upper bound on the adversarial loss (=> robustness against FGSM)

+ Regularization terms are related with over-fitting and achieving better generalization behaviors

Results – The regularization effect of mix-up

Claim : $L_n^{mix}(\theta,S) = L_n^{std}(\theta,S) + regularization\ term$

Lemma 3.1 (By Taylor theorem)

Consider the loss function $l(\theta,(x,y)) = h(f_\theta(x)) - yf_\theta(x)$, where h, f are twice differentiable for all $\theta \in \Theta$.
Let us denote $\tilde{D}_\lambda = \frac{\alpha}{\alpha+\beta} Beta(\alpha+1,\beta) + \frac{\beta}{\alpha+\beta} Beta(\beta+1,\alpha)$, D_X = empirical distribution of $S = \{(x_i,y_i)\}_{i=1}^n$
Then, the following holds :

$$L_n^{mix}(\theta,S) = L_n^{std}(\theta,S) + \sum_{i=1}^{3} R_i(\theta,S) + \mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[(1-\lambda)^2\varphi(1-\lambda)]$$

where $\lim_{\lambda\to 0}\varphi(\lambda) = 0$, and

$$R_1(\theta,S) = \frac{\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[1-\lambda]}{n}\sum_{i=1}^{n}(h'(f_\theta(x_i)) - y_i)\,\nabla f_\theta(x_i)^T\,\mathbb{E}_{r_x\sim D_x}[r_x - x_i]$$

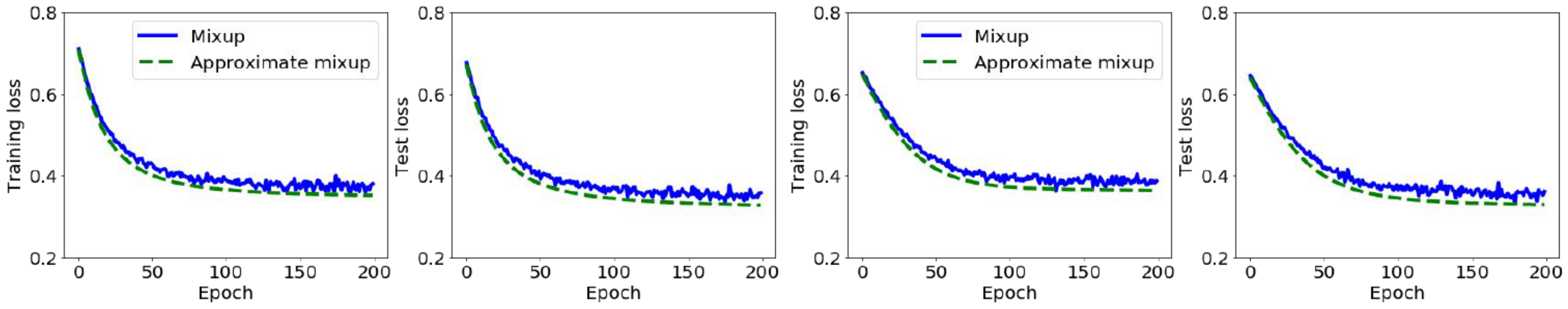
$$R_2(\theta,S) = \frac{\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[(1-\lambda)^2]}{2n}\sum_{i=1}^{n} h''(f_\theta(x_i))\,\nabla f_\theta(x_i)^T\,\mathbb{E}_{r_x\sim D_x}[(r_x - x_i)\,(r_x - x_i)^T]\,\nabla f_\theta(x_i)$$

$$R_3(\theta,S) = \frac{\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[(1-\lambda^2)]}{2n}\sum_{i=1}^{n}(h'(f_\theta(x_i)) - y_i)\,\mathbb{E}_{r_x\sim D_x}[(r_x - x_i)\nabla^2 f_\theta(x_i)(r_x - x_i)^T]$$

Results – The regularization effect of mix-up

Note

1. The loss function class $\mathcal{L} = \{l(\theta, (x, y)) | \, l(\theta, (x, y)) = h(f_\theta(x)) - y f_\theta(x)$ for some function h} includes many commonly used loss functions
   (ex : $h(x) = \log(1 + \exp(x))$ for logistic loss (= negative log-likelihood) or loss function induced by GLMs)
2. Quadratic approximation of $L_n^{mix}(\theta, S)$: $\tilde{L}_n^{mix}(\theta, S) = L_n^{std}(\theta, S) + \sum_{i=1}^{3} R_i(\theta, S)$ [regularization terms]
   (Empirically, $\tilde{L}_n^{mix}(\theta, S)$ is very close to $L_n^{mix}(\theta, S)$)



Logistic Regression

Two Layer ReLU Neural Network

Results – Mix-up and Adversarial robustness

Analysis setting

Analysis setting :

1. Consider logistic regression :

   $l(\theta, z) = \log\big(1 + \exp\big(f_\theta(x)\big)\big) - y f_\theta(x)$, where $y \in \{0,1\}$, $f_\theta(x) = \theta^T x$

2. Consider the case where $\theta \in \Theta = \left\{\theta \in \mathbb{R}^d \,\middle|\, y_i f_\theta(x_i) + (y_i - 1) f_\theta(x_i) \geq 0 \text{ for all } i = 1, \dots, n\right\}$

   (Note : Θ includes the set of all θ with zero training errors => $y_i = 1 \Rightarrow f_\theta(x_i) \geq 0$, $y_i = 0 \Rightarrow f_\theta(x_i) \leq 0$)

3. Consider the adversarial loss with l_2-attack of size $\epsilon\sqrt{d}$: $L_n^{adv}(\theta, S) = \frac{1}{n}\sum_{i=1}^{n} \max_{||\delta_i||_2 \leq \epsilon\sqrt{d}} l(\theta, (x_i + \delta_i, y_i))$

Results – Mix-up and Adversarial robustness

Lemma 3.2

The second order Taylor approximation of $L_n^{adv}(\theta, S)$ is $\frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\epsilon\sqrt{d}, (x_i, y_i)\right)$, where for any $\eta > 0, x \in \mathbb{R}^p$ and $y \in \{0,1\}$,

$$\tilde{l}_{adv}(\eta, (x, y)) = l(\theta, (x, y)) + \eta\left|g(x^T\theta) - y\right| \cdot \|\theta\|_2 + \frac{\eta^2}{2} g(x^T\theta)(1 - g(x^T\theta)) \cdot \|\theta\|_2^2$$

where $g(s) = \frac{e^s}{1+e^s}$ is logistic function

Theorem 3.1

Suppose there exists a constant $c_x > 0$ such that $\|x_i\|_2 \geq c_x\sqrt{d}$ for all $i \in \{1, \dots, n\}$. Then, for any $\theta \in \Theta$, we have

$$\tilde{L}_n^{mix}(\theta, S) \geq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\epsilon_i\sqrt{d}, (x_i, y_i)\right) \geq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\epsilon_{mix}\sqrt{d}, (x_i, y_i)\right)$$

where $\epsilon_i = R_i c_x \mathbb{E}_{\lambda \sim \tilde{D}_\lambda}[1 - \lambda]$ with $R_i = |\cos(\theta, x_i)|$ and $\epsilon_{mix} = R c_x \mathbb{E}_{\lambda \sim \tilde{D}_\lambda}[1 - \lambda]$ with $R = \min_{i \in \{1,\dots,n\}} |\cos(\theta, x_i)|$

Question : Can it be generalized in non-logistic circumstance, also some cases when $\nabla^2 f_\theta(x) \neq 0$??

Results – Mix-up and Adversarial robustness

Note

1. $\tilde{L}_n^{mix}(\theta, S)$ is upper bound of the second order taylor expansion of $L_n^{adv}(\theta, S)$ with l_2-attack size $\epsilon_{mix}\sqrt{d}$

   => Minimizing the Mix-up loss would result in a small adversarial loss

2. $\epsilon_{mix}\sqrt{d}$ seems to be small attack in d-dimensional data (tends to be single-step attacks, such as FGSM),

   So future works for exploring robustness against large and sophisticated multiple-step attacks (ex : I-FGSM) are required.

3. ϵ_{mix} depends on θ, but we need constant lower bound => Theorem 3.2

Assumption for theorem 3.2 (Assumption 3.1)

Denote $\widehat{\Theta}_n = \{\theta \in \Theta \mid minimizer\ of\ \tilde{L}_n^{mix}(\theta, S)\}$, and assume there exists a set Θ^* such that for all $n \geq N \in \mathbb{N}$, $\widehat{\Theta}_n \subseteq \Theta^*$ with probability at least $1 - \delta_n$, where $\delta_n \to 0\ as\ n \to \infty$. Moreover, there exists a $\tau \in (0,1)$ such that

$$p_\tau = \mathbb{P}(\{x \in \mathcal{X} : |\cos(x, \theta)| \geq \tau \text{ for all } \theta \in \Theta^*\}) \in (0,1]$$

Results – Mix-up and Adversarial robustness

Theorem 3.2

Under assumption 3.1, if there exists constants $b_x, c_x > 0$ such that $c_x\sqrt{d} \leq \|x_i\|_2 \leq b_x\sqrt{d}$ for all $i \in \{1, \dots, n\}$. Then,

with probability at least $1 - \delta_n - 2\exp(-\frac{np_\tau^2}{2})$, there exists constant $\kappa > 0, \kappa_2 > \kappa_1 > 0$, such that for any $\theta \in \widehat{\Theta}_n$, we have

$$\tilde{L}_n^{mix}(\theta, S) \geq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\tilde{\epsilon}_{mix}\sqrt{d}, (x_i, y_i)\right)$$

where $\tilde{\epsilon}_{mix} = \tilde{R}c_x\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[1-\lambda]$ and $\tilde{R} = \min\{\frac{p_\tau\kappa_1}{2\kappa_2 - p_\tau(\kappa_2-\kappa_1)}, \sqrt{\frac{4\kappa p_\tau}{2-p_\tau+4\kappa p_\tau}}\} \cdot \tau$

Note

Now, consider more general case : NN with ReLU/Max-pooling

$\Rightarrow f_\theta(x) = \beta^T\sigma(W_{N-1}\cdots(W_2\sigma(W_1x)))$, where σ = nonlinear function via ReLU / max-pooling

Note that $f_\theta(x) = \nabla f_\theta(x)^T x$ and $\nabla^2 f_\theta(x) = 0$ (almost everywhere)

Results – Mix-up and Adversarial robustness

Theorem 3.3

Assume that $f_\theta(x) = \nabla f_\theta(x)^T x$ and $\nabla^2 f_\theta(x) = 0$ and there exists a constant $c_x > 0$ such that $\|x_i\|_2 \geq c_x\sqrt{d}$ for all $i \in \{1, \dots, n\}$. Then, for any $\theta \in \Theta$, we have

$$\tilde{L}_n^{mix}(\theta, S) \geq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\epsilon_i\sqrt{d}, (x_i, y_i)\right) \geq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\left(\epsilon_{mix}\sqrt{d}, (x_i, y_i)\right)$$

where $\epsilon_i = R_i c_x \mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[1-\lambda]$ with $R_i = |\cos(\nabla f_\theta(x_i), x_i)|$ and $\epsilon_{mix} = R c_x \mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[1-\lambda]$ with $R = \min_{i\in\{1,\dots,n\}} |\cos(\nabla f_\theta(x_i), x_i)|$

Assumption for theorem 3.A (Assumption 3.A)

Denote $\widehat{\Theta}_n = \{\theta \in \Theta \mid minimizer\ of\ \tilde{L}_n^{mix}(\theta, S)\}$, and assume there exists a set Θ^* such that for all $n \geq N \in \mathbb{N}$, $\widehat{\Theta}_n \subseteq \Theta^*$ with probability at least $1 - \delta_n$, where $\delta_n \to 0\ as\ n \to \infty$. Moreover, there exists a $\tau, \tau' \in (0,1)$ such that

$$p_{\tau,\tau'} = \mathbb{P}\left(\left\{x \in \mathcal{X}: \left|\cos\left(x, \nabla f_\theta(x)\right)\right| \geq \tau, \|\nabla f_\theta(x)\| \geq \tau' \text{ for all } \theta \in \Theta^*\right\}\right) \in (0,1)$$

Results – Mix-up and Adversarial robustness

Theorem 3.A

Under assumption 3.A, if there exists constants $b_x, c_x > 0$ such that $c_x\sqrt{d} \le \|x_i\|_2 \le b_x\sqrt{d}$ for all $i \in \{1, \dots, n\}$. Then, with probability at least $1 - \delta_n - 2\exp(-\frac{np_{\tau,\tau'}^2}{2})$, there exists constant $\kappa > 0, \kappa_2 > \kappa_1 > 0$, such that for any $\theta \in \widehat{\Theta}_n$, we have

$$\tilde{L}_n^{mix}(\theta, S) \ge \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{adv}\,(\tilde{\epsilon}_{mix}\sqrt{d}, (x_i, y_i))$$

where $\tilde{\epsilon}_{mix} = \tilde{R}c_x\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}[1-\lambda]$ and $\tilde{R} = \min\{\frac{p_{\tau,\tau'}\kappa_1\tau}{p_{\tau,\tau'}\kappa_1\tau+\left(2-p_{\tau,\tau'}\right)\kappa_2\tau'}, \sqrt{\frac{p_{\tau,\tau'}\kappa\tau^2}{\frac{2-p_{\tau,\tau'}}{4\tau'^2}+p_{\tau,\tau'}\kappa\tau^2}}\} \cdot \tau$

Results – Mix-up and Generalization

Note and analysis setting

Here, We show that the data-dependent regularization induced by Mix-up directly controls the Rademacher complexity of the underlying function classes => yield concrete generalization error bounds.

Analysis setting :

1. GLM case : $l\big(\theta, (x, y)\big) = A\big(\theta^T x\big) - y\theta^T x,\ f_\theta(x) = \theta^T x$
2. Non-linear two-layer NN case : $l\big(\theta, (x, y)\big) = \big(y - f_\theta(x)\big)^2,\ f_\theta(x) = \theta_1^T \sigma(Wx) + \theta_0$

Notations : (related with regularization terms obtained by the second-order approximation of $\tilde{L}_n^{mix}(\theta, S)$)

- $\mathcal{W}_\gamma = \{x \to \theta^T x \mid \mathbb{E}_x[A''(\theta^T x) \cdot \theta^T \Sigma_X \theta \leq \gamma\}$: function class in GLM (regularization induced by mix-up)

  where $\Sigma_X = \mathbb{E}[x_i x_i^T]$
- $\mathcal{W}_\gamma^{NN} = \{x \to f_\theta(x) \mid \theta_1^T \Sigma_X^\sigma \theta_1 \leq \gamma\}$: function class in NN (regularization induced by mix-up)

  where $\Sigma_X^\sigma = \mathbb{E}[\hat{\Sigma}_X^\sigma]$, and $\hat{\Sigma}_X^\sigma$ = sample covariance matrix of $\{\sigma(Wx_i)\}_{i=1}^n$

Results – Mix-up and Generalization

Lemma 3.3/3.4 (by Talyor theorem)

Consider the centralized dataset S, that is, $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$, and denote $\hat{\Sigma}_X = \frac{1}{n} x_i x_i^T$, $\hat{\Sigma}_X^{\sigma}$ = sample covariance matrix of $\{\sigma(W x_i)\}_{i=1}^{n}$.

For GLM, if $A(\cdot)$ is twice differentiable, then the regularization term (=$\sum_{i=1}^{3} R_i(\theta, S)$) obtained by the second-order approximation of $\tilde{L}_n^{mix}(\theta, S)$ is given by

$$\frac{1}{2n}[\sum_{i=1}^{n} A''(\theta^T x_i)] \cdot \mathbb{E}_{\lambda\sim\tilde{D}_\lambda}\left[\frac{(1-\lambda)^2}{\lambda^2}\right]\theta^T \hat{\Sigma}_X \theta$$

For NN, the regularization term is given by

$$\mathbb{E}_{\lambda\sim\tilde{D}_\lambda}\left[\frac{(1-\lambda)^2}{\lambda^2}\right]\theta_1^T\, \hat{\Sigma}_X^{\sigma} \theta_1$$

Recall : $\tilde{D}_\lambda = \frac{\alpha}{\alpha+\beta} Beta(\alpha+1, \beta) + \frac{\beta}{\alpha+\beta} Beta(\beta+1, \alpha)$

Results – Mix-up and Generalization

Def : ρ-retentive distribution (for theorem 3.4)

The distribution of x is ρ-retentive for some $\rho \in \left(0, \frac{1}{2}\right]$ if for any non-zero vector $v \in \mathbb{R}^d$,

$$\left[\mathbb{E}_x\left[A''\left(x^T v\right)\right]\right]^2 \geq \rho \cdot \min\{1, \mathbb{E}_x[\left(v^T x\right)^2]\}$$

Theorem 3.4 / 3.B

The empirical Rademacher complexity of $\mathcal{W}_\gamma$ satisfies (when the distribution of x_i is ρ-retentive)

$$Rad\left(\mathcal{W}_\gamma, S\right) \leq \max\{\left(\frac{\gamma}{\rho}\right)^{\frac{1}{4}}, \left(\frac{\gamma}{\rho}\right)^{\frac{1}{2}}\} \cdot \sqrt{\frac{rank(\Sigma_X)}{n}}$$

The empirical Rademacher complexity of $\mathcal{W}_\gamma^{NN}$ satisfies

$$Rad\left(\mathcal{W}_\gamma^{NN}, S\right) \leq 2\sqrt{\frac{\gamma \cdot rank\left(\Sigma_X^\sigma\right) + \left\|(\Sigma_X^{\sigma\frac{\dagger}{2}}\mathbb{E}_x[\sigma(Wx)]\right\|^2}{n}}$$

Recall : $\mathcal{W}_\gamma = \{x \to \theta^T x \mid \mathbb{E}_x[A''(\theta^T x) \cdot \theta^T \Sigma_X \theta \leq \gamma\}$, $\mathcal{W}_\gamma^{NN} = \{x \to f_\theta(x) \mid \theta_1^T \Sigma_X^\sigma \theta_1 \leq \gamma\}$

$\Sigma_X = E[xx^T]$, $\Sigma_X^\sigma = \mathbb{E}[\hat{\Sigma}_X^\sigma]$, $\hat{\Sigma}_X^\sigma$ = sample covariance matrix of $\{\sigma(Wx_i)\}_{i=1}^n$.

Results – Mix-up and Generalization

Corollary 3.1/ Theorem 3.5 (Apply Lemma A.1 to theorem 3.4/3.B)

Suppose $\mathcal{X}, \mathcal{Y}, \Theta$ are all bounded, then

For GLM, if $A(\cdot)$ is L_A-Lipschitz continuous, there exists constants $L, B > 0$, such that for all $f_\theta \in \mathcal{W}_\gamma$, we have ,with probability at least $1-\delta$,

$$L(\theta) \leq L_n^{std}(\theta, S) + 2L \cdot L_A \left(\max\left\{ \left(\frac{\gamma}{\rho}\right)^{\frac{1}{4}}, \left(\frac{\gamma}{\rho}\right)^{\frac{1}{2}} \right\} \cdot \sqrt{\frac{rank(\Sigma_X)}{n}} \right) + B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}$$

For NN, there exists constants $L, B > 0$, such that for all $f_\theta \in \mathcal{W}_\gamma^{NN}$, we have ,with probability at least $1-\delta$,

$$L(\theta) \leq L_n^{std}(\theta, S) + 4L\sqrt{\frac{\gamma \cdot rank\left(\Sigma_X^\sigma\right) + \left\| (\Sigma_X^{\sigma\frac{\dagger}{2}} \mathbb{E}_x[\sigma(Wx)] \right\|^2}{n}} + B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}$$

Recall : $L(\theta)$ is standard population loss

Conclusion (+appendix)

Appendix : Lemma A.1(Result from Bartlett & Mendelson, 2002)

For any B-uniformly bounded and L-Lipschitz function l, for all $f \in \mathcal{F}$, with probability at least 1-δ,

$$\mathbb{E}\left[l\big(f(x)\big)\right] \leq \frac{1}{n}\sum_{i=1}^{n} l\big(f(x_i)\big) + 2L \cdot Rad(\mathcal{F}, S) + B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}$$

Conclusion

- Mix-up training is approximately a regularized loss minimization
- Derived regularization terms are used to demonstrate why Mix-up has improved generalization and robustness against one-step adversarial examples (small l_2 norm attack)