Boosting Adversarial Training via Fisher-Rao Norm-based Regularization

Supplementary Material

In this appendix, we provide detailed descriptions of the methodologies and experimental procedures used in our study, which encompasses:

- Theoretical proofs in detail.
- Additional experimental settings and results.
- Limitations and future directions.

The aim is to ensure the transparency and reproducibility of our research, while providing sufficient information for readers interested in the technical intricacies of our work.

8. Proofs of Lemma. 2 and Thm. 1

According to Lemma. 1, we can denote the Fisher-Rao norm of $\mathcal{W}$ w.r.t $\mathcal{L}_{ce}$ as:

$$\|\mathcal{W}\|^2_{FR\circ\mathcal{L}_{ce}} = L^2\mathbb{E}_{(\mathbf{x},y)}\left[\left(\langle\sigma(f^L_{\mathcal{W}}(\mathbf{x})), f^L_{\mathcal{W}}(\mathbf{x})\rangle - f^L_{\mathcal{W}}(\mathbf{x})_y\right)^2\right] \tag{16}$$

Then we can conclude that

$$\begin{aligned}
\gamma_{ce} &= \mathbb{E}_{(\mathbf{x},y)}\left[\left|\langle\sigma(f^L_{\mathcal{W}}(\mathbf{x})), f^L_{\mathcal{W}}(\mathbf{x})\rangle - f^L_{\mathcal{W}}(\mathbf{x})_y\right|\right] \\
&= \mathbb{E}_{(\mathbf{x},y)}\left[\left|\sum_{k=1}^{K}\frac{(f^L_{\mathcal{W}}(\mathbf{x})_k - f^L_{\mathcal{W}}(\mathbf{x})_y)e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}{\sum_{k=1}^{K}e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}\right|\right] \\
&= \mathbb{E}_{(\mathbf{x},y)}\left[\left|\sum_{k\neq y}\frac{(f^L_{\mathcal{W}}(\mathbf{x})_k - f^L_{\mathcal{W}}(\mathbf{x})_y)e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}{\sum_{k=1}^{K}e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}\right|\right] \\
&\leq \mathbb{E}_{(\mathbf{x},y)}\left[\frac{\sum_{k\neq y}\left|(f^L_{\mathcal{W}}(\mathbf{x})_k - f^L_{\mathcal{W}}(\mathbf{x})_y)\right|e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}{\sum_{k\neq y}e^{f^L_{\mathcal{W}}(\mathbf{x})_k}}\right] \\
&\leq \mathbb{E}_{(\mathbf{x},y)}\left[\max_{k\neq y}\left|f^L_{\mathcal{W}}(\mathbf{x})_k - f^L_{\mathcal{W}}(\mathbf{x})_y\right|\right]
\end{aligned} \tag{17}$$

Then the standard Rademacher complexity w.r.t the CE loss $\mathcal{R}_{N_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}})$ can be denoted as:

$$\begin{aligned}
&\mathbb{E}_\xi \sup_{f^L_{\mathcal{W}}\in\mathcal{F}_{\hat{\gamma}_{ce}}}\frac{1}{N}\sum_{i=1}^{N}\xi_i\mathcal{L}_{ce}\left(f^L_{\mathcal{W}}(\mathbf{x}_i), y_i\right) \\
&= \mathbb{E}_\xi \sup_{f^L_{\mathcal{W}}\in\mathcal{F}_{\hat{\gamma}_{ce}}}\frac{1}{N}\sum_{i=1}^{N}\xi_i\ln\left(\frac{e^{f^L_{\mathcal{W}}(\mathbf{x}_i)_y}}{\sum_{k=1}^{K}e^{f^L_{\mathcal{W}}(\mathbf{x}_i)_k}}\right) \\
&= \mathbb{E}_\xi \sup_{f^L_{\mathcal{W}}\in\mathcal{F}_{\hat{\gamma}_{ce}}}\frac{1}{N}\sum_{i=1}^{N}\xi_i\left(\ln\left(\sum_{k=1}^{K}e^{f^L_{\mathcal{W}}(\mathbf{x}_i)_k}\right) - f^L_{\mathcal{W}}(\mathbf{x}_i)_y\right) \\
&= \mathbb{E}_\xi \sup_{f^L_{\mathcal{W}}\in\mathcal{F}_{\hat{\gamma}_{ce}}}\frac{1}{N}\sum_{i=1}^{N}\xi_i\left(\ln\left(\sum_{k=1}^{K}\left(\frac{1}{e}\right)^{f^L_{\mathcal{W}}(\mathbf{x}_i)_y - f^L_{\mathcal{W}}(\mathbf{x}_i)_k}\right)\right)
\end{aligned} \tag{18}$$

Specifically, for N^M_{tr} misclassified samples, the upper bound for $\mathcal{R}_{N^M_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}})$ can be further derived as:

$$\begin{aligned}
&\mathcal{R}_{N^M_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}}) \\
&\leq \mathbb{E}_\xi \sup_{f^L_{\mathcal{W}}\in\mathcal{F}_{\hat{\gamma}_{ce}}}\frac{1}{N^M_{tr}}\sum_{i=1}^{N^M_{tr}}\xi_i\left(\ln\left(\sum_{k=1}^{K}\left(\frac{1}{e}\right)^{f^L_{\mathcal{W}}(\mathbf{x}_i)_y - f^L_{\mathcal{W}}(\mathbf{x}_i)_k}\right)\right) \\
&\leq \mathbb{E}_\xi \frac{1}{N^M_{tr}}\sum_{i=1}^{N^M_{tr}}\xi_i\left(\ln\left(\sum_{k=1}^{K}\left(\frac{1}{e}\right)^{-\left|f^L_{\mathcal{W}}(\mathbf{x}_i)_y - f^L_{\mathcal{W}}(\mathbf{x}_i)_k\right|}\right)\right) \\
&\leq (\ln K + \hat{\gamma}_{ce})\sqrt{\frac{1}{{N^M_{tr}}^2}(\mathbb{E}_\xi\sum_{i=1}^{N^M_{tr}}\xi_i)^2} \\
&\leq (\ln K + \hat{\gamma}_{ce})\sqrt{\frac{1}{N^M_{tr}}(\mathbb{E}_\xi\sum_{i=1}^{N^M_{tr}}\xi_i)} \\
&\overset{(i)}{\lesssim} \frac{\ln K + \hat{\gamma}_{ce}}{\sqrt{N^M_{tr}}}
\end{aligned} \tag{19}$$

Notice that it is very likely that $0 < \left|\mathbb{E}_\xi\sum_{i=1}^{N^M_{tr}}\xi_i\right| < N^M_{tr}$, we then assume $\left|\mathbb{E}_\xi\sum_{i=1}^{N^M_{tr}}\xi_i\right| \approx \sqrt{N^M_{tr}}$. Due to the fact that we are deriving the upper bound, (i) holds. Meanwhile, for N^C_{tr} correctly classified samples, the following inequalities holds:

$$\begin{aligned}
&\mathcal{R}_{N^C_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}}) \\
&\geq \mathbb{E}_\xi\frac{1}{N^C_{tr}}\sum_{i=1}^{N^C_{tr}}\xi_i\left(\ln\left(\sum_{k=1}^{K}\left(\frac{1}{e}\right)^{\hat{\gamma}_{ce}}\right)\right) \\
&\geq \frac{(\ln K - \hat{\gamma}_{ce})}{N^C_{tr}}\mathbb{E}_\xi\sum_{i=1}^{N^C_{tr}}\xi_i
\end{aligned} \tag{20}$$

Similar to Eq. 19, we assume $\left|\mathbb{E}_\xi\sum_{i=1}^{N^C_{tr}}\xi_i\right| \approx \sqrt{N^C_{tr}}$. Meanwhile, due to variations in network architecture, training data, and training algorithm, etc, the value of $\hat{\gamma}_{ce}$ tends to vary a lot, therefore, we need to make $\mathbb{E}_\xi\sum_{i=1}^{N^C_{tr}}\xi_i$ positive to reach the lower bound, which can be denoted as $\mathcal{R}_{N^C_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}}) \gtrsim \frac{\ln K - \hat{\gamma}_{ce}}{\sqrt{N^C_{tr}}}$. Consequently, we can deduce that:

$$\begin{aligned}
\mathcal{R}_{N_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}}) &\leq \frac{(\sqrt{N^M_{tr}} + \sqrt{N^C_{tr}})\ln K + \hat{\gamma}_{ce}\sqrt{N^M_{tr}}}{N_{tr}} \\
\mathcal{R}_{N_{tr}}(\mathcal{L}_{ce}\circ\mathcal{F}_{\hat{\gamma}_{ce}}) &\geq \frac{(\sqrt{N^M_{tr}} + \sqrt{N^C_{tr}})\ln K - \hat{\gamma}_{ce}\sqrt{N^C_{tr}}}{N_{tr}}
\end{aligned} \tag{21}$$

Then if we define $\Gamma_{ce} = \frac{\hat{\gamma}_{ce}^{N_{tr}^C} - \hat{\gamma}_{ce}^{N_{tr}^M}}{\hat{\gamma}_{ce}^{N_{tr}^M}}$, $\mathcal{C}_C = \frac{N_{tr}}{N_{tr}^C}$, $\mathcal{C}_M = \frac{N_{tr}}{N_{tr}^M}$, $\mathcal{C}_{MC} = \frac{\sqrt{N_{tr}^C}+\sqrt{N_{tr}^M}}{N_{tr}}$, we can easily provide the following bounds:

$$\mathcal{R}_{N_{tr}}\left(\mathcal{L}_{ce} \circ \mathcal{F}_{\hat{\gamma}_{ce}}\right) \leq \mathcal{C}_{MC} \ln K + \frac{\hat{\gamma}_{ce}^{N_{tr}^M} \mathcal{C}_M^{-0.5} (\mathcal{C}_C^{-1} \Gamma_{ce} + 1)}{N_{tr}^{0.5}}$$

$$\mathcal{R}_{N_{tr}}\left(\mathcal{L}_{ce} \circ \mathcal{F}_{\hat{\gamma}_{ce}}\right) \geq \mathcal{C}_{MC} \ln K - \frac{\hat{\gamma}_{ce}^{N_{tr}^M} \mathcal{C}_C^{-0.5} (\mathcal{C}_C^{-1} \Gamma_{ce} + 1)}{N_{tr}^{0.5}} \tag{22}$$

9. Additional Experimental Settings and Results

Experimental Settings. The qualitative Results illustrated in Fig. 2 and 3 are derived from training models on a dataset comprising 4000 samples from MNIST. For the evaluations presented in Fig. 4, we utilized 4,000 training instances from MNIST and 10,000 from the CIFAR10. Furthermore, for all setups in LOAT-boosted training, we consistently set the parameters τ to 1 and γ to 0.05, following the specifications of Alg. 1.

Additional Results. In addition to the LOAT-boosted S2O results on ResNet18, as detailed in Tab. 5, we extended our experiments to WideResNet-34-10. The comprehensive outcomes of these additional experiments are presented in Tab. 6. Our findings indicate that SLORE-enhanced PGD-AT can enhance the test clean accuracy of WideResNet-34-10 by 1.40%. Furthermore, LORE-enhanced PGD-AT contributes to an increase in the model's robustness against PGD^7 attacks by 0.70%, PGD^{20} attacks by 0.93%, and PGD^{40} attacks by 1.04%.

Furthermore, we compare our algorithm with widely used Adversarial Logit Pairing (ALP). Detailed results are shown in Tab. 7.

Additionally, we conducted evaluations of the LOAT-boosted DM-AT on augmented datasets from Cifar100, Tiny-ImageNet and SVHN. The extensive results in Tab. 8, 9 and 10 reinforce the adaptability and efficacy of our approach. To more comprehensively demonstrate the efficacy of our approach, we trained LOAT-boosted TRADES

Table 6. Evaluation of LOAT-boosted S2O on WideResNet-34-10 (%).

Defense	Clean_{te}	PGD^7	PGD^{20}	PGD^{40}
PGD-AT	85.57	53.20	47.84	46.78
PGD-AT+SLORE	**86.97**	*53.51*	*48.20*	*47.23*
PGD-AT+LORE	*86.69*	**53.90**	**48.77**	**47.82**

Table 7. Comparison of SLORE-boosted PGD-AT with ALP-boosted one on ResNet-18 (%).

Defense	Clean_{te}	FGSM	PGD^7	PGD^{20}
PGD-AT (ALP)	81.26	60.92	51.38	45.90
PGD-AT (SLORE)	**83.41**	**61.80**	**51.46**	**46.50**

Table 8. Classification Accuracy of models trained by 1×10^6 EDM-generated images-augmented SVHN (%).

Models	Defense	Clean_{tr}	Clean_{te}	PGD^{20}	PGD^{40}	AA
PreResNet18 (Swish)	TRADES	96.59	96.57	56.74	53.84	36.21
	TRADES+SLORE	96.22	96.14	**61.20**	**57.85**	**38.58**
	TRADES+LORE	**97.79**	**97.39**	55.65	47.33	9.64
	TRADES(S)	96.06	96.20	69.91	66.64	**41.38**
	TRADES(S)+SLORE	**97.12**	**97.01**	**77.73**	**73.38**	13.48
	TRADES(S)+LORE	96.46	96.44	65.25	61.82	37.65
WRN-28-10 (Swish)	TRADES	97.45	**97.69**	59.57	54.75	15.17
	TRADES+SLORE	95.46	95.67	**68.35**	**67.01**	**49.25**
	TRADES+LORE	**97.61**	97.66	43.44	37.51	10.79
	TRADES(S)	**97.43**	**97.75**	**74.98**	**69.10**	15.77
	TRADES(S)+SLORE	97.11	97.31	70.17	65.19	11.27
	TRADES(S)+LORE	97.22	97.36	57.71	50.58	**16.36**
	MART	**97.23**	**97.23**	52.34	49.89	22.89
	MART+SLORE	88.13	93.97	64.22	63.82	**53.26**
	MART+LORE	94.70	96.30	**72.16**	**70.76**	9.03

Table 9. Classification Accuracy of models trained by 1×10^6 EDM-generated images-augmented Cifar100 (%).

Models	Defense	Clean_{tr}	Clean_{te}	PGD^{20}	PGD^{40}	AA
PreResNet18 (Swish)	TRADES	85.27	62.33	34.62	**34.59**	**29.66**
	TRADES+SLORE	84.95	61.63	34.41	34.47	29.40
	TRADES+LORE	**86.35**	**62.79**	**34.63**	34.55	29.21
	TRADES(S)	81.84	60.15	33.82	33.76	28.76
	TRADES(S)+SLORE	**81.85**	**60.22**	33.66	33.73	**28.88**
	TRADES(S)+LORE	81.81	59.86	**34.02**	**34.00**	28.70
	MART	77.78	59.91	34.86	34.83	29.18
	MART+SLORE	77.08	59.86	34.72	34.70	**29.60**
	MART+LORE	**78.57**	**60.18**	**35.20**	**35.22**	29.47
WRN-28-10 (Swish)	TRADES	89.39	65.41	37.88	37.93	33.02
	TRADES+SLORE	88.95	64.95	38.07	38.04	**33.29**
	TRADES+LORE	**90.17**	**66.62**	**38.10**	**37.96**	32.61
	TRADES(S)	**85.22**	63.12	36.62	36.57	31.60
	TRADES(S)+SLORE	84.78	**63.18**	36.50	36.54	**31.97**
	TRADES(S)+LORE	84.82	62.91	**36.78**	**36.65**	31.56
	MART	81.82	65.05	38.56	38.51	33.91
	MART+SLORE	80.35	64.46	38.88	38.81	**34.26**
	MART+LORE	**83.21**	**65.11**	**39.06**	**39.00**	33.76

Table 10. Classification Accuracy of PreResNet18 (Swish) trained by 1×10^6 EDM-generated images-augmented Tiny-ImageNet (%). V, S, and L represent vanilla, SLORE-boosted, and LORE-boosted adversarial training algorithms respectively. C_{tr}, C_{te}, P_{40}, A_{ce}, and A_t indicate Clean_{tr}, Clean_{te}, PGD^{40}, APGD_{ce}, and APGD_t individually.

	TRADES					MART				
	C_{tr}	C_{te}	P_{40}	A_{ce}	A_T	C_{tr}	C_{te}	P_{40}	A_{ce}	A_T
V	64.91	58.66	28.52	28.31	**21.96**	52.52	50.62	28.14	28.02	21.51
S	64.76	58.32	28.33	28.32	21.59	**55.17**	**51.34**	**29.02**	**28.75**	**22.54**
L	**65.23**	**59.08**	**28.78**	**28.47**	21.68	53.04	50.79	28.61	27.74	21.26

Table 11. Classification Accuracy of PreResNet18 (Swish) trained by 1×10^6 EDM-generated images-augmented Cifar10 after 400 epochs (%).

Defense	Clean_{te}	PGD^{10}	PGD^{20}	PGD^{40}	AA
TRADES	88.45	62.62	61.88	61.76	**58.21**
TRADES+SLORE	86.12	62.28	61.73	61.56	57.76
TRADES+LORE	**90.26**	**62.81**	**62.01**	**61.85**	56.51
MART	88.35	61.19	60.18	59.93	53.92
MART+SLORE	88.15	61.99	60.94	60.78	54.74
MART+LORE	**89.08**	**63.81**	**62.83**	**62.57**	**55.28**

and MART on PreResNet18 (Swish) for an extended duration of 400 epochs. The evaluations, as presented in Tab. 11, reveal a more pronounced improvement in performance compared to training for approximately 100 epochs. This enhancement is particularly notable in the case of MART.

10. Limitations and Future Works

- **Mitigating Theoretical Gap Between MLPs and Other Network Structures.** We intend to implement knowledge distillation using a ReLU-activated MLP, aligning its architecture to mirror the layer count of a ResNet residual block or a Transformer stack, and matching the hidden units in each layer to the dimensionality of latent features derived from these structures. This framework will be employed in LOAT-boosted AT algorithms, which can help us investigate the theoretical gap.
- **Hyperparameters in LOAT.** The selection of hyperparameters $\mathcal{E}_1$, $\mathcal{E}_2$, τ, and γ is crucial, as their configuration can significantly impact the final results.
- **Future Directions.** The rising prominence of Transformers and Large Language Models (LLMs) presents new challenges and opportunities. Exploring the application of adversarial training frameworks to these models is both an intriguing and vital avenue for future research.