# Adversarial local distribution regularization for knowledge distillation

## 1. Asymptotic analysis of adversarial local distribution approximation.

The kernel $F$ is the radial basis function kernel used in our paper, as shown in Eq. 1

$$F(\boldsymbol{x}', \boldsymbol{x}) = \exp\left\{\frac{-||\boldsymbol{x}' - \boldsymbol{x}||^2}{2\sigma^2}\right\}. \tag{1}$$

The update function $\phi$ can be rewritten as

$$\phi(\boldsymbol{x}_{adv}) = \frac{1}{K}\sum_{j=1}^{K}\Big[F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\nabla_{\boldsymbol{x}_{adv}^{j,(l)}}\ell(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}; \theta) - F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\frac{(\boldsymbol{x}_{adv}^{j,(l)} - \boldsymbol{x}_{adv})}{\sigma^2}\Big]. \tag{2}$$

When $\sigma \to \infty$, it is obvious that

$$\phi(\boldsymbol{x}_{adv}) \to \frac{1}{K}\sum_{j=1}^{K}\nabla_{\boldsymbol{x}_{adv}^{j,(l)}}\ell(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}; \theta). \tag{3}$$

Therefore, our approach reduces exactly to FGSM [2], PGD [3], and Auto-Attack [1] with $K$ independent particles, where in the update quantity is the average of the gradients at each particle as shown in Eq. (5). Evidently, in the update rule in Eq. (5), there does not exist any term that promotes the particle diversity. In addition, when using a single particle (i.e., $n = 1$), our approach under its asymptotic case reduces exactly to the aforementioned approaches.

Particularly, in our update formula in Eq. (8), the first term encourages the particles to seek the optimal values of the loss surface as in FGSM [2], PGD [3], and Auto-Attack [1], while the second term plays a role of a repulsive term to push the particles away for enhancing the particle diversity. The reason is that when $\boldsymbol{x}_{adv}^{j,(l)}$ moves closer to $\boldsymbol{x}_{adv}$, the weight $k(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})$ becomes larger to push them further away from each other.

We present the asymptotic analysis when $\sigma \to 0$. Considering the RBF kernel, the update function $\phi$ can be rewritten as

$$\phi(\boldsymbol{x}_{adv}) = \frac{1}{K}\sum_{j=1}^{K}\Big[F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\nabla_{\boldsymbol{x}_{adv}^{j,(l)}}\ell(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}; \theta) - F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\frac{(\boldsymbol{x}_{adv}^{j,(l)} - \boldsymbol{x}_{adv})}{\sigma^2}\Big]. \tag{4}$$

When $\sigma \to 0$, it is obvious that

$$\phi(\boldsymbol{x}_{adv}) \to \frac{1}{K}\sum_{j=1}^{K}1_{\boldsymbol{x}_{adv}=\boldsymbol{x}_{adv}^{j,(l)}}\nabla_{\boldsymbol{x}_{adv}^{j,(l)}}\ell(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}; \theta), \tag{5}$$

where $1_A$ is the indicator function which returns 1 if $A$ is true and 0 if otherwise. Here we note that we have used the following equations in the above derivation.

$$\lim_{\sigma\to 0} F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\frac{(\boldsymbol{x}_{adv}^{j,(l)} - \boldsymbol{x}_{adv})}{\sigma^2} = 0. \tag{6}$$

$$\lim_{\sigma\to 0} F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv}) = 0 \tag{7}$$

if $\boldsymbol{x}_{adv} \neq \boldsymbol{x}_{adv}^{j,(l)}$.

$$\lim_{\sigma\to 0} F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv}) = 1 \tag{8}$$

if $\boldsymbol{x}_{adv} = \boldsymbol{x}_{adv}^{j,(l)}$.

Therefore, the update amount $\phi(\boldsymbol{x}_{adv})$ in Eq. (5) reduces to only one gradient. It is evident that when $n = 1$, our approach reduces exactly to FGSM [2], PGD [3], and Auto-Attack [1].

## 2. Experimental setting details

### 2.1. Diversity of teacher adversarial particles vs. random initialization

We set $\epsilon$ = 0.3, $\eta$ =0.01, $L$=200, $\tau$=1.0.

### 2.2. TALD regularization with existing methods on CIFAR-100 and ImageNet

For CIFAR-100, we set $\epsilon$ = 0.3, $\eta$ =0.1, $L$=1, $K$=4, $\tau$=10.0, $\lambda$=0.01. All methods used in our experiments are trained by SGD. The learning rate is initialized as 0.05, and decayed it by 0.1 every 30 epochs after the first 150 epochs until the last 240 epoch. We use a learning rate of 0.01 for MobileNetV2, ShuffleNetV1 and ShuffleNetV2, while 0.05 is optimal for other models. Batch size is 64.

For ImageNet, we use settings from config files of Torchdistill[1]. We set $\epsilon$ = 0.3, $\eta$ =0.2, $L$=1, $K$=4, $\tau$=5.0, $\lambda$=0.01.

## References

[1] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of ICML*, pages 2206–2216. PMLR, 2020. 1

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Proceeding ICLR*, 2014. 1

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[1] github.com/yoshitomo-matsubara/torchdistill/tree/main/torchdistill