Particle-based Adversarial Local Distribution Regularization

Thanh Nguyen-Duc, Trung Le, He Zhao, Jianfei Cai, Dinh Phung

Department of DSAI, Monash University

AISTATS 2022

09/03/2022

THANH NGUYEN

The minmax optimization

- This optimization improves model robustness and generalization.
- PGD[1] and TRADES[2] for robust ML, VAT[3] for semi-supervised learning, and VADA[4] for domain adaptation.

The minmax optimization

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y)\sim P_{\mathbb{D}}} \left[\max_{\boldsymbol{x}' \in B_{\epsilon}(\boldsymbol{x})} \ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta) \right]$$

$(\boldsymbol{x}, y) \sim P_{\mathbb{D}}$

f_θ model parameterized by θ

$\boldsymbol{x}'$ adversarial example

$B_{\epsilon}(\boldsymbol{x}) = \{\boldsymbol{x}' \in \boldsymbol{X} : ||\boldsymbol{x}' - \boldsymbol{x}||_p \leq \epsilon\}$

- The loss function can be various:
  - PGD uses cross-entropy loss

$$\ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta) = CE(f_\theta(\boldsymbol{x}'), y)$$

  - TRADES, VAT and VADA use KL-divergence

$$\ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta) = D_{\mathrm{KL}}(f_\theta(\boldsymbol{x}'), f_\theta(\boldsymbol{x}))$$

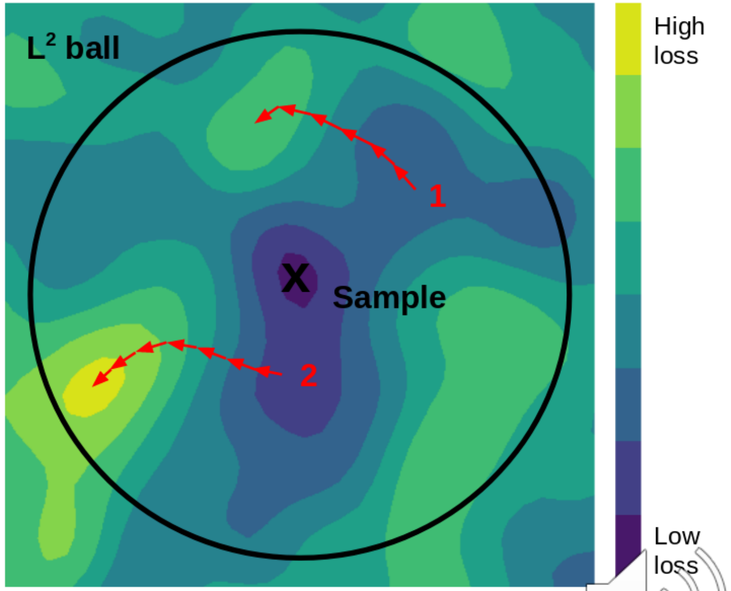
[1] Towards deep learning models resistant to adversarial attacks. ICLR, 2018
[2] Theoretically principled trade-off be- tween robustness and accuracy. ICML, 2019
[3] Virtual adversarial training: a regularization method for supervised and semi-supervised learning, TPAMI, 2018
[4] A dirt-t approach to unsupervised domain adaptation. ICLR, 2018

Adversarial local distribution

- Adversarial local distribution:
  - Consists of all adversarial examples inside the ball constraint.
  - Replaces the maximization optimization by a conditional distribution.

$$P_\theta(\boldsymbol{x}'|\boldsymbol{x},y) := \frac{e^{\ell(\boldsymbol{x}',\boldsymbol{x},y;\theta)}}{\int_{B_\epsilon(\boldsymbol{x})} e^{\ell(\boldsymbol{x}'',\boldsymbol{x},y;\theta)} d\boldsymbol{x}''} = \frac{e^{\ell(\boldsymbol{x}',\boldsymbol{x},y;\theta)}}{Z(\boldsymbol{x},y;\theta)}$$

- $P_\theta(\cdot|\boldsymbol{x},y)$ is the conditional local distribution over $B_\epsilon(\boldsymbol{x})$
- $Z(\boldsymbol{x},y;\theta)$ is a normalization function



https://towardsdatascience.com/know-you...en emy-7f7c5038bdf3

Adversarial local distribution regularization

- The regularization:
  - improves robustness and generalization.
  - can be applied to robust ML, semi-supervised learning and domain adaptation
  - is a general regularization of PGD, TRADES, VAT and VADA.

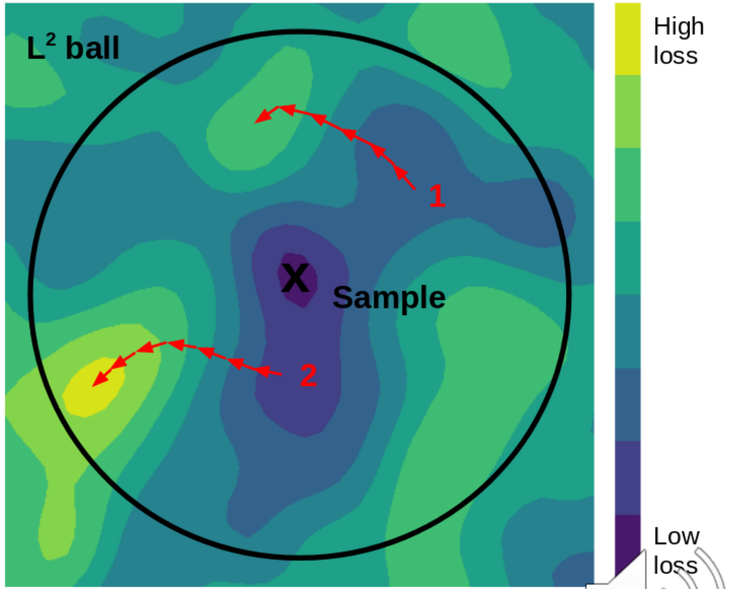
$$R(\theta, \boldsymbol{x}, y) := \mathbb{E}_{\boldsymbol{x}' \sim P_\theta(\cdot|\boldsymbol{x}, y)}[\log P_\theta(\boldsymbol{x}'|\boldsymbol{x}, y)]$$

$$= -H(P_\theta(\cdot|\boldsymbol{x}, y)),$$

$$B_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}' \in \boldsymbol{X} : ||\boldsymbol{x}' - \boldsymbol{x}||_p \leq \epsilon\}$$

- Minimize $R(\theta, \boldsymbol{x}, y)$ or equivalent to maximize $H(P_\theta(\cdot|\boldsymbol{x}, y))$
  - $P_\theta(\cdot|\boldsymbol{x}, y)$ to be more uniform distribution
  - $\ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta) = \ell(\boldsymbol{x}'', \boldsymbol{x}, y; \theta) = c(x, y; \theta)$

- Minimizing the regularization loss leads to an enhancement in the model output smoothness



https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

Adversarial local distribution regularization

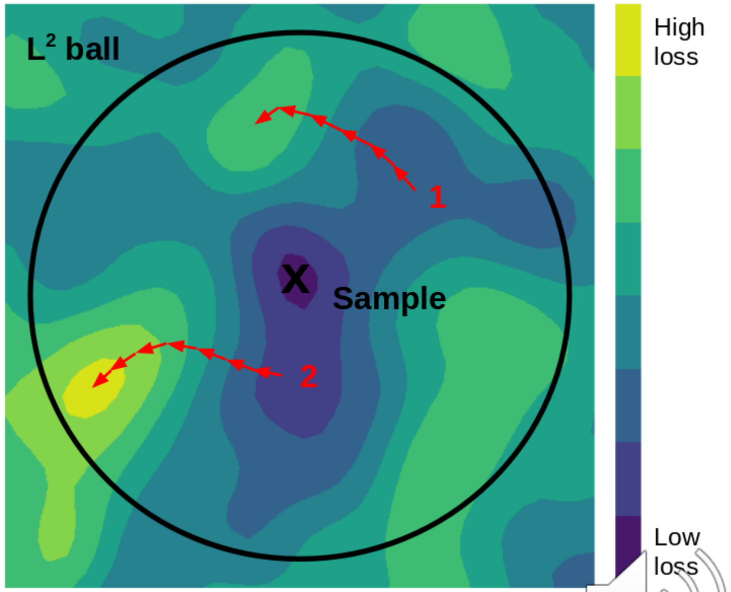
- $Z(\boldsymbol{x}, y; \theta)$ is intractable to find

$$P_\theta(\boldsymbol{x}'|\boldsymbol{x}, y) := \frac{e^{\ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta)}}{\int_{B_\epsilon(\boldsymbol{x})} e^{\ell(\boldsymbol{x}'', \boldsymbol{x}, y; \theta)} d\boldsymbol{x}''} = \frac{e^{\ell(\boldsymbol{x}', \boldsymbol{x}, y; \theta)}}{Z(\boldsymbol{x}, y; \theta)}$$

- Solve by a particle-based method to sample

$$\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_n \sim P_\theta(\cdot|\boldsymbol{x}, y))$$

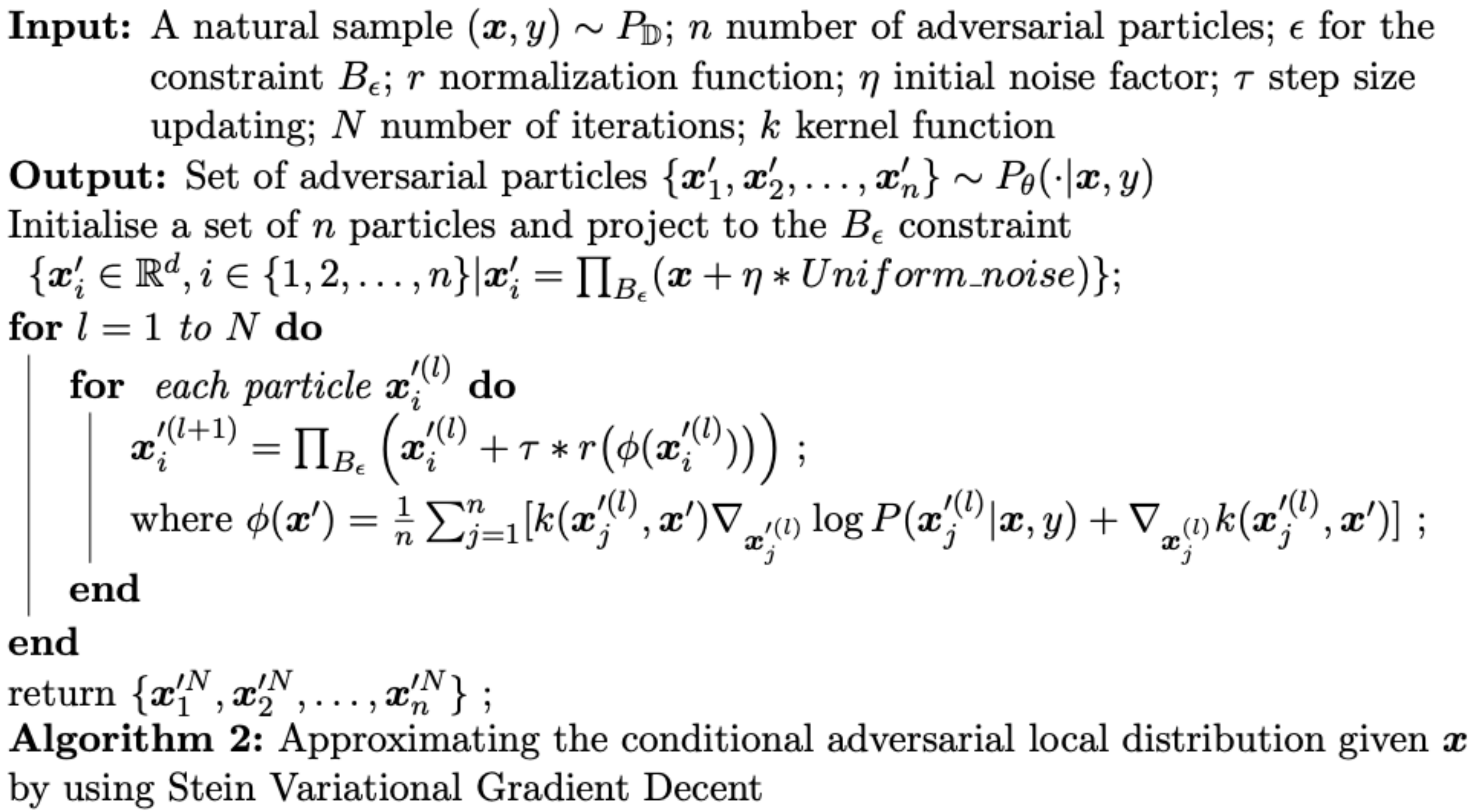
n: is the number of samples (or adversarial particles)



https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

Approximate the adversarial local distribution

- Stein Variational Gradient Decent (SVGD) is a particle-based inference method using a functional gradient decent to approximate a ground-truth distribution.



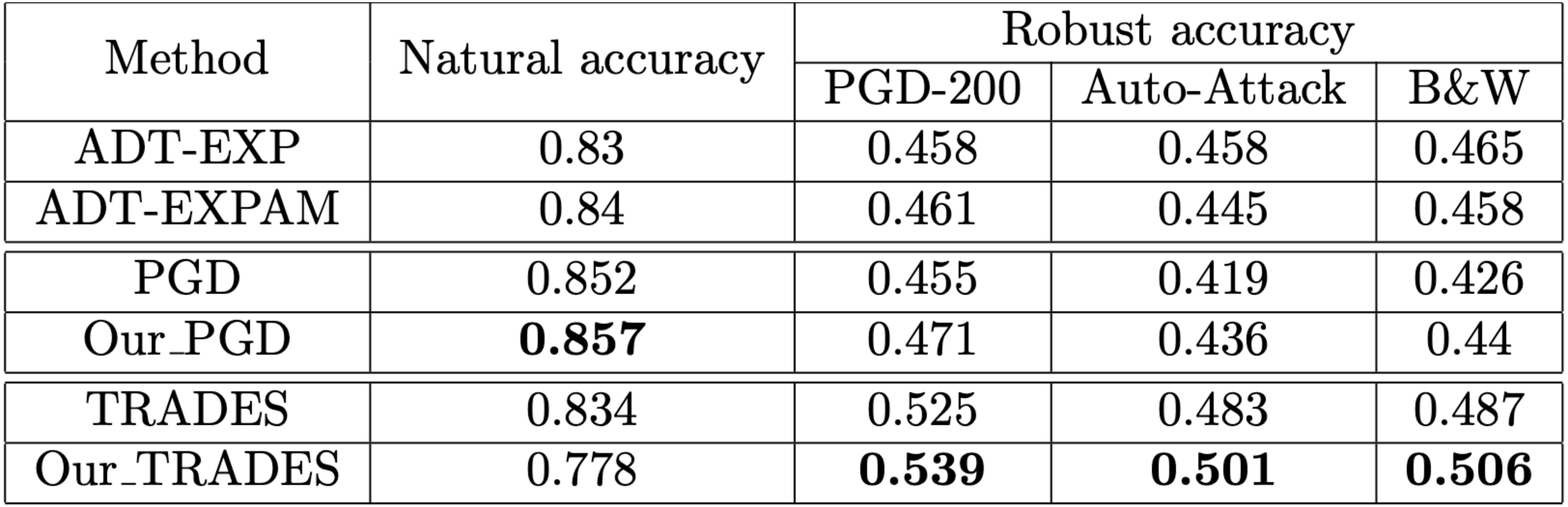
```
**Input:** A natural sample $(\boldsymbol{x}, y) \sim P_{\mathbb{D}}$; $n$ number of adversarial particles; $\epsilon$ for the
        constraint $B_\epsilon$; $r$ normalization function; $\eta$ initial noise factor; $\tau$ step size
        updating; $N$ number of iterations; $k$ kernel function
**Output:** Set of adversarial particles $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_n\} \sim P_\theta(\cdot|\boldsymbol{x}, y)$
Initialise a set of $n$ particles and project to the $B_\epsilon$ constraint
 $\{\boldsymbol{x}'_i \in \mathbb{R}^d, i \in \{1, 2, \ldots, n\} | \boldsymbol{x}'_i = \prod_{B_\epsilon}(\boldsymbol{x} + \eta * Uniform\_noise)\}$;
**for** $l = 1$ *to* $N$ **do**
    **for** *each particle* $\boldsymbol{x}_i'^{(l)}$ **do**
        $\boldsymbol{x}_i'^{(l+1)} = \prod_{B_\epsilon}\left(\boldsymbol{x}_i'^{(l)} + \tau * r(\phi(\boldsymbol{x}_i'^{(l)}))\right)$ ;
        where $\phi(\boldsymbol{x}') = \frac{1}{n}\sum_{j=1}^{n}[k(\boldsymbol{x}_j'^{(l)}, \boldsymbol{x}')\nabla_{\boldsymbol{x}_j'^{(l)}} \log P(\boldsymbol{x}_j'^{(l)}|\boldsymbol{x}, y) + \nabla_{\boldsymbol{x}_j^{(l)}} k(\boldsymbol{x}_j'^{(l)}, \boldsymbol{x}')]$ ;
    **end**
**end**
return $\{\boldsymbol{x}_1'^N, \boldsymbol{x}_2'^N, \ldots, \boldsymbol{x}_n'^N\}$ ;
```

Algorithm 2: Approximating the conditional adversarial local distribution given $\boldsymbol{x}$ by using Stein Variational Gradient Decent

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Proceedings of NeurIPS, volume 29, 2016

Adversarial local distribution regularization

- Robust machine learning
  - CIFAR10 dataset
  - ResNet18



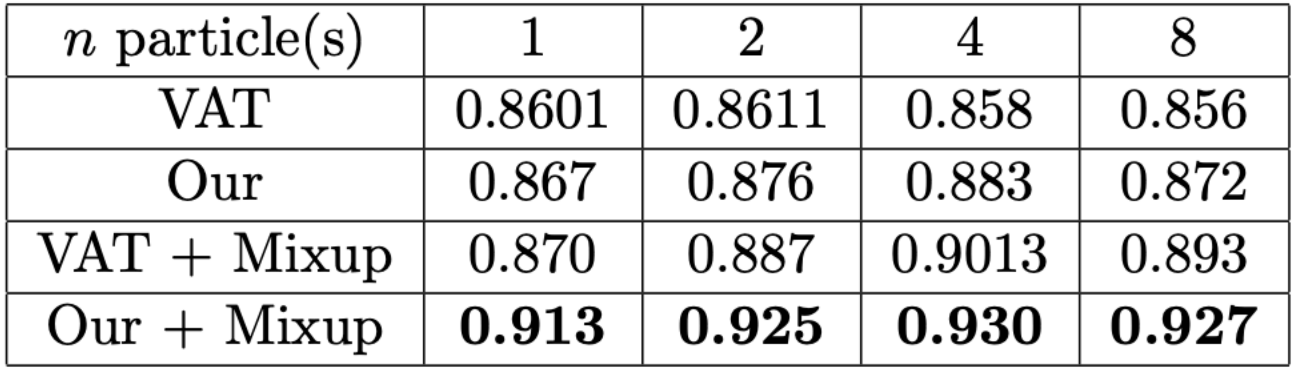
Method	Natural accuracy	Robust accuracy		
		PGD-200	Auto-Attack	B&W
ADT-EXP	0.83	0.458	0.458	0.465
ADT-EXPAM	0.84	0.461	0.445	0.458
PGD	0.852	0.455	0.419	0.426
Our_PGD	**0.857**	0.471	0.436	0.44
TRADES	0.834	0.525	0.483	0.487
Our_TRADES	0.778	**0.539**	**0.501**	**0.506**

Natural and robust accuracy comparison using CIFAR10 with ResNet18.

Robust accuracy: accuracy of model with perturbed images

Adversarial local distribution regularization

- Semi-supervised learning
- CIFAR10 dataset
  - Trainset: 4,000 labeled samples, 56,000 unlabeled samples
  - Testset: 10,000 samples



n particle(s)	1	2	4	8
VAT	0.8601	0.8611	0.858	0.856
Our	0.867	0.876	0.883	0.872
VAT + Mixup	0.870	0.887	0.9013	0.893
Our + Mixup	**0.913**	**0.925**	**0.930**	**0.927**

Performance comparison between our method and VAT using mixup technique for all adversarial particles

Thank You