# Reproducing <Certifying Some Distributional Robustness with Principled Adversarial Training>

**Xiao Fan (20086722)** **Zhibin Lu (20091078)**

## 1 Paper Summary

We reproduced the paper "Certifying Some Distributional Robustness with Principled Adversarial Training" [SND17]. We will firstly introduce the background of the problem, the existing solutions and their drawbacks and then describe the approach proposed by the paper.

### 1.1 Motivation

Robustness of deep learning systems is critical as model failure is extremely costly in situations like self-driving cars, automated detection of tumors. However, recent work shows that neural networks are vulnerable to adversarial examples which means imperceptible perturbations to data can lead to misbehavior of the model such as misclassification.

Consider the supervised learning problem, in which we minimize an expected loss $\mathbb{E}_{P_0}[l(\theta; Z)]$ over a parameter $\theta \in \Theta$, where $Z \sim P_0$ and $l$ is the loss function. Robustness to changes in the data-generating distribution $P_0$ is desirable, whether they be from covariant shifts, changes in the underlying domain, or adversarial attacks. From the perspective of distributionally robust optimization, we consider a distribution $P$ around $P_0$ and the problem is now :

$$\min_{\theta \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(\theta; Z)] \tag{1}$$

However, the inner supremum problem is intractable except for specially structured loss.

One standard **robust-optimization approach** simplifies the problem and considers the loss $l(\theta; (x_i + \Delta_{x_i}(\theta), y_i)$ where $\Delta_{x_i}(\theta)$ is seen as perturbation to data. One heuristic is "fast gradient sign method" (FGM) [GSS15] where

$$\Delta_{x_i} := \arg\max_{\|\eta\|_p \le \epsilon} \{\nabla_x l(\theta; (x_i, y_i))^T \eta\} \tag{2}$$

It extends to multi-step variant "iterative fast gradient sign method" (IFGM) [KGB16] and "projected-gradient method" (PGM) [MMS$^+$17] where

$$\Delta_{x_i}^{t+1}(\theta) := \arg\max_{\|\eta\|_p \le \epsilon} \{\nabla_x l(\theta; (x_i^t, y_i))^T \eta\} \quad \text{and} \quad x_i^{t+1} := \Pi_{x:\|x - x_i\|_p \le \epsilon}\{x_i^t + \alpha \Delta x_i^t(\theta)\} \tag{3}$$

These methods have no guarantee to converge and there are possibly worst-case perturbations that are not found, especially with deep networks use ReLU activations.

Another approach is **distributionally robust optimization** in which researchers consider finite-dimensional parametrization for $\mathcal{P}$, such as constraint sets for moments, support, as well as non-parametric distances for probability measures such as f-divergences and Wasserstein distances. The proposed approach in this paper uses similar techniques but is able to treat larger classes of losses and costs and defends against imperceptible adversarial perturbations.

Submitted as course project for IFT6135 H2018 Representation Learning

### 1.2 Proposed approach

The proposed approach introduces a Lagrangian relaxation into 1 and perturbs the underlying data distribution in a Wasserstein ball, the problem is reformulated as :

$$\min_{\theta\in\Theta}\left\{F(\theta) := \sup_{P}\{\mathrm{E}_P[l(\theta;Z)] - \gamma W_c(P,P_0)\}\} = \mathbb{E}_{P_0}[\phi_\gamma(\theta;Z)]\right\} \tag{4a}$$

$$\text{where} \quad \phi_\gamma(\theta;z_0) := \sup_{z\in\mathcal{Z}}\{l(\theta;z) - \gamma c(z,z_0)\} \tag{4b}$$

The formula constrains robustness to region $\mathcal{P} = \{P : W_c(P,P_0) \le \rho\}$, $W_c(\cdot,\cdot)$ is the Wasserstein metric. $c(z,z_0)$ is the cost to perturb $z_0$ to $z$ and $c(z,z_0) = \|z - z_0\|_p^2$. The empirical counterpart is to replace $P_0$ with the training data distribution $\hat{P}_n$.

A stochastic gradient-type algorithm is developed for the relaxation formula 4

**Algorithm 1** Distributionally robust optimization with adversarial training

```
Input : Sampling distribution $P_0$, constraint sets $\Theta$ and $\mathcal{Z}$, stepsize sequence $\{\alpha_t > 0\}_{t=0}^{T-1}$
for $t = 0,\ldots,T-1$ do
    Sample $z^t \sim P_0$ and find an $\epsilon$-approximation maximizer $\hat{z}^t$ of $l(\theta;z) - \gamma c(z,z^t)$
    $\theta^{t+1} \leftarrow \mathrm{Proj}_\Theta(\theta^t - \alpha_t \nabla_\theta l(\theta^t;\hat{z}^t))$
```

Whenever the loss is smooth enough and the penalty $\gamma$ is large enough (corresponding to small perturbation), the surrogate in 4b is strongly concave and is therefore efficiently computable. Robustness is guaranteed as an upper bound of worst-case perturbation $\sup_{P\in\mathcal{P}}\mathbb{E}_P[l(\theta;Z)]$ is provided

$$\sup_{P:W_c(P,P_0)\le\rho} \mathbb{E}_P[l(\theta;Z)] \le \gamma\rho + \mathbb{E}_{\hat{P}_n}[\phi_\gamma(\theta;Z)] + O(1/\sqrt{n}) \quad \forall\theta\in\Theta \tag{5}$$

What's more, the author demonstrates that the algorithm generalizes. Robustness achieved on training set also helps test set to protect from adversarial attack.

## 2 Experiments

### 2.1 Motivation for the experiments

In original paper, three empirical evaluations are conducted respectively on synthetic data, MNIST and Markov decision process. The first two are supervised learning task and the third is reinforcement learning task. Performance of WRM (Algorithm 1) are compared with Empirical risk minimization (ERM), the fast-gradient method (FGM) and its iterated variant (IFGM) and the projected-gradient method (PGM).

The good properties of this approach, robustness guarantee, efficient computation, generalization, rely greatly on the concavity of the surrogate 4b. Smoothness of the loss function and large $\gamma$ are two important preconditions. The first experiment is designed to illustrate the impact of smoothness of the loss function and what the benefits of the certificate. It uses synthetic data and compares the decision boundary of a small neural network use either ReLU (not smooth) or ELU (smooth) activations. Three methods ERM, FGM and WRM are compared. Large $\gamma$ is used in all experiments because WRM is designed for small to moderate perturbations. However, more experiments with small $\gamma$ are provided in appendix to test the performance on larger perturbation.

The second experiment provides a standard benchmark which trains a neural network classifier on the MNIST dataset. Several figures are plotted to demonstrate different benefits of WRM.

- Figure 2 in original paper plots the certificate vs. the worst-case perturbation. This helps to validate the upper bound proposed in 5. Same plot is provided with synthetic data.
- Figure 3 plots the test error vs. $\epsilon_{\text{adv}}$ ($\epsilon = \sqrt{\rho}$, $\rho$ indicates the level of perturbation) under PGM attacks with respect to 2- and $\infty$-norms. Since models are trained with 2-norm attack,

performance under $\infty$-norm helps to demonstrate the extensive ability under other attacks. Performance on test set shows that the model generalizes. However, it's better to provide performance on training set to be more convinced. Different adversarial training techniques like FGM, IFGM, PGM and WRM are compared to show that WRM outperforms the others with smaller test error.

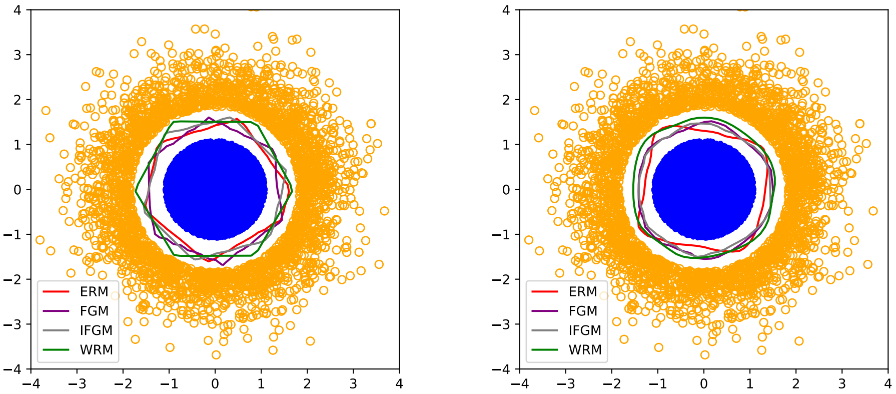
- Figure 4 studies the loss surface with respect to perturbations to inputs. It illustrates the defense mechanism of WRM by reducing magnitudes of gradients (corresponding to small values of $\hat{\rho}_{\text{test}}$).

The third experiment is a reinforcement learning task to show that technique for distributionally robust optimization extends beyond supervised learning. We omit the details here.

## 2.2 Reproducing the main results

### 2.2.1 Supervised learning with synthetic data

First experiment is to train a classifier with synthetic data which contains two classes. We generated $10^4$ data for training and 4000 data for test. We used batch-size of 128, $\gamma = 2$, learning rate of 0.01 and 15 iterations with stepsize of $0.08/\sqrt{t}$ for iterated method (IFGM and WRM). All models are trained with 2-norm PGM attack. All methods achieved 100% accuracy on both training and test set after 30 epochs. The training process sometimes gets stuck with no improving accuracy and needs a restart, this may due to the specific property of the synthetic data because we didn't observe the same problem in MNIST.

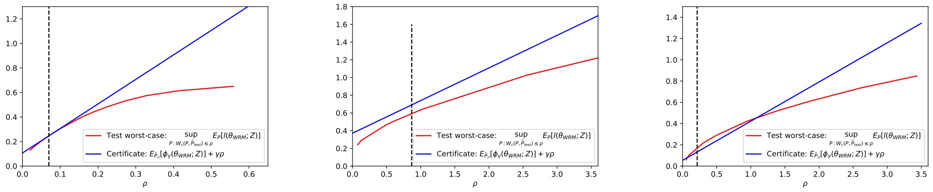


(a) ReLU model

(b) ELU model

Figure 1: Experimental results on synthetic data. Training data are in blue and orange. Decision boundaries are in red, purple, gray and blue for ERM, FGM, IFGM and WRM respectively.

Decision boundaries are shown in figure 1 for both ReLU and ELU activations. As 70% data are in class blue, classification boundaries should push toward class orange. In both ReLU and ELU, WRM pushes the boundary further outwards than the other methods indicating better robustness. Differences between ReLU and ELU activations are obvious. With ReLU, all boundaries have sharp angles where the models are easily attacked. All methods suffer from asymmetry due to the lack of smoothness of ReLU activation. With ELU which provides smoothness, same level of robustness is guaranteed with symmetric boundaries, especially for WRM. For all experiments followed, ELU activation is used.



(a) Synthetic data

(b) MNIST at Epoch 2

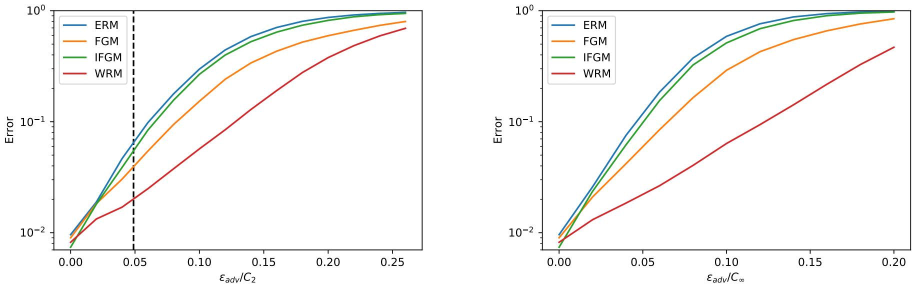
(c) MNIST at Epoch 30

Figure 2: Certificate of robustness 5 and worst-case perturbation. The vertical bar indicates the achieved level of robustness on the training set $\hat{\rho}_n(\theta_{\text{WRM}})$.

Figure 2 plots the robustness certificate and the worst-case perturbation to show the validation of equation 5. Figure 2(a) for the synthetic data is perfect. The certificate (upper bound) is always above the worst-case perturbation and is tightest at $\hat{\rho}_n$. For MNIST, we plotted the same figures at epoch 2 (accuracy=97%) and epoch 30 (accuracy=99%) respectively. We observed that with more epochs, the line of certificate gradually approaches the curve of the worst-case perturbation. Figure 2(b) is aligned with the paper while in figure 2(c), the worst-case perturbation is sometimes slightly above the certificate. One possible reason is that we omitted the term $O(1/\sqrt{n})$ in equation 5. Or we didn't find the best hyper-parameters as the paper and we didn't train the model properly. The author should provide more details about how they train MNIST and how they produce this figure since we can't easily reproduce the same result.

### 2.2.2 Supervised learning with MNIST

For synthetic data, we perfectly reproduced the results in the paper, however, for MNIST, we experienced more discrepancy. We used $5 \times 10^4$ examples for training, $10^4$ for validation and $10^4$ for test. We used batch-size of 128, $\gamma = 0.04C_2$ ($C_2 = \mathbb{E}_{\hat{P}_n}[\|X\|_2] = 9.21$), learning rate of 0.0001 and 15 iterations with stepsize of $0.04/\sqrt{t}$. All methods achieved at least 99% accuracy on test set after 30 epochs. As learned from experience, sequences of stepsize and number of epochs are important hyper-parameters that decide the robustness achieved ($\hat{\rho}_n$). Since no information is provided for these hyper-parameters, we didn't obtain the exact same results shown in the paper.

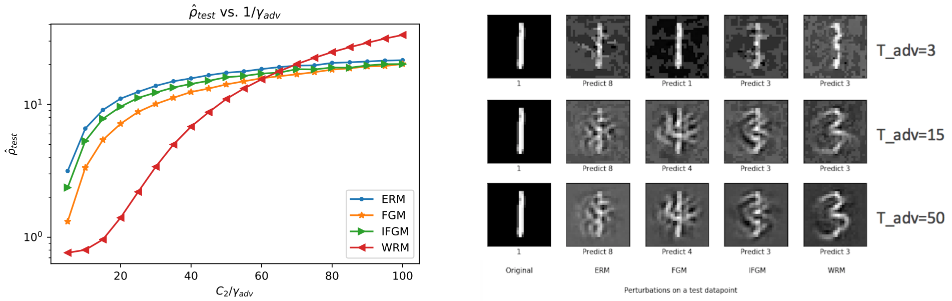


(a) Test error vs. $\epsilon_{\text{adv}}$ for $\|\cdot\|_2$ attack

(b) Test error vs. $\epsilon_{\text{adv}}$ for $\|\cdot\|_\infty$ attack

Figure 3: Test classification error vs. the adversarial perturbation level $\epsilon_{\text{adv}}$ under 2-norm and $\infty$-norm attack. The vertical bar indicates the achieved level of robustness on the training set $\sqrt{\hat{\rho}_n(\theta_{\text{WRM}})}$.

Figure 3 indicates the test error under different level of attacks. With larger perturbation (larger $\epsilon_{\text{adv}}$), test error increases. Among all methods, WRM has the smallest error which indicates better robustness. The models are trained with 2-norm attacks but can generalize to $\infty$-norm attacks and can also generalize to test set. This figure shows that WRM outperforms other adversarial techniques. However, we expect some analytical analysis about why WRM beats other methods.



(a) $\hat{\rho}_{\text{test}}$ vs. $1/\gamma_{\text{adv}}$

(b) Perturbations on a test datapoint

Figure 4: Stability of the loss surface. In (a), we show the average distance of the perturbed distribution $\hat{\rho}_{\text{test}}$ for a given $1/\gamma_{\text{adv}}$. In (b), we visualize the smallest WRM perturbation to make a model misclassify a datapoint.

Figure 4 reproduced the experiment on stability of loss surface. Different from the paper, we observed exploding perturbation with small $\gamma$ for WRM in figure 4(a). This may be reasonable since WRM is only for small to moderate perturbation (large $\gamma$) and the performance under larger perturbation is unclear. As explained by the author, small values of $\hat{\rho}_{\text{test}}$ correspond to small magnitudes of $\nabla_z l(\theta; z)$ which ensures the stability of the model. Figure 4(b) provides a qualitative analysis. We tried to find the smallest perturbation necessary to make a model misclassify for digits 0 to 9 and picked digit 1 in the figure. We increased the stepsize to 0.5 to accelerate the search. With number of iterations of 15, ERM predicts 8, FGM predicts 4, IFGM and WRM predict 3. For human eyes, the picture of WRM looks like 3 the most. It indicates that WRM is able to improve interpretability when reducing magnitude of gradients. Tried with different values of number of iterations, we found this parameter should be large enough to guarantee an optimal $\rho_{\text{test}}$ and thus a correct level of perturbation given a $\gamma$. The misclassification is not easily produced for ERM, FGM and IFGM and for digits 0, 3 and 6. This may be caused by the difficulty to find the gradient ascent direction.

## 3 Discussion

In general, the author clearly expresses his motivation of all experiments and successfully shows advantages of the proposed approach (certificated robustness, better performance) with either quantitative or qualitative results. We would like to share some thoughts and comments here after reproducing part of the paper.

- Experiment with ReLU and ELU demonstrates the importance of smoothness as the precondition of the proposed approach. We can obviously see the impact of ReLU activations for the synthetic data but we expect quantitative results to be more convinced, for examples comparison of test error in MNIST. Also, we are curious if other structures of neural network will impact the smoothness like the pooling layer or the ResNet. This helps us to know the proper settings to use this algorithm.
- The certificate of robustness is one of the most important advantages of the approach and the author succeeded to show it is indeed an upper bound of the worst-case perturbation. However, we didn't perfectly reproduce the same results. With more training, the upper bound is sometimes slightly below the worst-case perturbation for MNIST. We may miss some critical hyper-parameters or training techniques. The author should mention necessary information for reproduction.
- Comparison with other adversarial techniques like FGM, IFGM, PGM shows that WRM has better performance (lower test error under attack, more stability). We wonder why the author choose these methods as baselines and since the proposed approach is distributionally based, if other methods using similar techniques (e.g. defensive distillation [CW17]) should be compared. In addition, as learned from experience and also from other references [MMS$^+$17], stepsize is an important hyper-parameter, it's better to add more experiments to confirm if current conclusion holds for all values of this parameter.

- Even though benchmark with MNIST classifier is included in the paper, we would like to see the results on larger datasets like ImageNet or CIFAR10.
- There misses some analysis of the disadvantages of this algorithm. One limitation as emphasized by the author is that the algorithm applies only to small and moderate perturbation. Extra experiments with larger perturbation can be found in the appendix, but we expect some rigorous analysis. Additionally, what is a "small" perturbation is not clear. There is no explicit measure about this quantity thus we don't know up to when the proposed approach will outperform the other methods.

## References

[CW17] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[GSS15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[KGB16] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016.

[MMS$^+$17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[SND17] A. Sinha, H. Namkoong, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *ArXiv e-prints*, October 2017.