Report from S. Ramesh and A. Bense about project 10: Certifying Some Distributional Robustness with Principled Adversarial Training

Ramesh Shyam and Bense Alexandre
Department of Mathematics
Swiss Federal Institute of Technology in Zürich
Rämistrasse 101 CH-8092
shramesh@student.ethz.ch and abense@student.ethz.ch

1 Motivation for the problem

In this report we will explain and dissect the article: Sinha, A, Namkoong, H, Volpi, R. and Duchi, J. Certifying Some Distributional Robustness with Principled Adversarial Training. (2017) *arXiv*.

1.1 Robustness

Neural networks have been the leading methods of machine learning in almost all fields for a few years. They allow to reach unprecedented accuracy with many types of data. We can mention the detection of flower or animal species on images, facial or vocal recognition, text translation or the creation of works in the mind of an artist. Neural networks therefore seemed flawless at first glance. Yet some scientific works have questioned the robustness of neural networks like in this article [1] where the authors have developed an algorithm named DeepFool to perturb and deceive image classification algorithms. We can look at an example shown in the Figure 1. Therefore, Neural Networks are actually lacking of robustness.

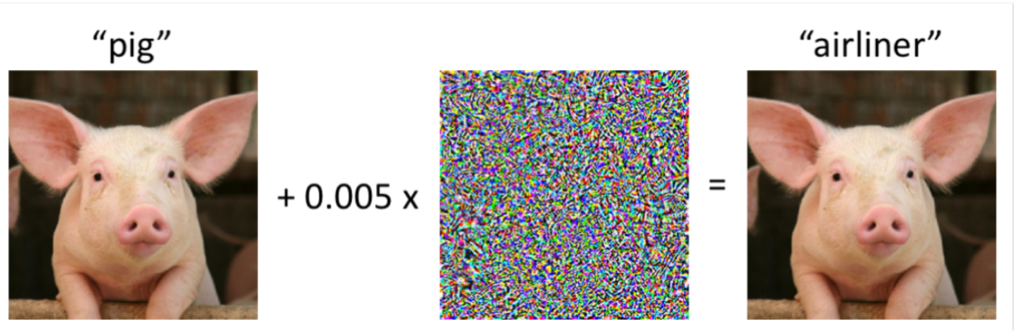


Figure 1: Example of misclassication.

The aim of the article under discussion is to tackle this issue and the goal of this report is to explain, detail and deconstruct the results. To do so, a classical machine learning setup is considered with data Z, a loss function l, a set of parameters of our model θ living in the set Θ and data-generating distribution P_0 living in a class of distributions $\mathcal{P}_0$. The problem can thus be mathematically understood as finding the solution to the problem:

$$\min_{\theta \in \Theta} \mathbb{E}_{P_0}[l(\theta; Z)] \tag{1}$$

Therefore, to counteract the robustness problems some researchers thought of adding small perturbations to the training base like in this paper [2]. The new problem to be solved then becomes:

$$\min_{\theta \in \Theta} \mathbb{E}_{P_0}(\sup_{u \in U}[l(\theta; Z+u)]) \tag{2}$$

for some perturbations u in a uncertainty set U. Yet, this problem is hard to investigate especially because of the supremum function in the expectancy. Thus, the authors have handled the adversaries using distributional robustness rather than pointwise robustness as described above. Indeed, distributional robustness is more easily solvable and different theoretical results can be deduced, which is a real advantage compared to the problem (2).

To justify even further the choice of the paper to focus on distributionally robustness instead of adversarial perturbations, we can look at research works published after 2017 like detailed in this paper [2], which have proved that the original problem (2) had some weaknesses. Indeed, the type of perturbation is decisive and in truth does not improve general robustness. For example, the authors of [2] proved that the perturbations l_1 and l_∞ are mutually exclusive. Therefore, it seems impossible to significantly improve the robustness against different types of perturbations. Moreover, even if these methods improve a certain robustness, global or individual, they always do so at the expense of accuracy. It would thus seem that there is a strong trade-off between robustness to different type of perturbations and accuracy. Therefore, using distributional robustness rather than pointwise robustness could be easier to investigate and at the same time handle better the robustness issue by not caring about the type of perturbation.

Concretely the distributional robust definition of the problem becomes:

$$\min_{\theta \in \Theta} \sup_{W_c(P,P_0) \leq \rho} \mathbb{E}_P[l(\theta; Z)] \tag{3}$$

with W_c the Wasserstein distance. Before going any further we would first like to introduce wasserstein distance and how it is relevant to our setting.

1.2 Wasserstein Distance

To measure the distance between 2 distributions P and Q, the usual metric used is Wasserstein Distance aka Earthmover's distance. Atleast in the discrete setting this can be imagined to be the minimum cost of moving probability mass from the support of P to Q such that after the movement the new distribution obtained is Q.

The formal definition is as follows: For P, Q probability measures in $\mathcal{Z}$, $\Pi(P,Q)$ be the set of all joint distribution over (P,Q) such that for all, $M \in \Pi(P,Q)$ $\int_p dM(p,A) = Q(A)$ and $\int_q dM(A,q) = P(A)$ for any Borel set A. Then the Wasserstein Distance between P and Q is defined as follows:

$$\inf_{M \in \Pi(P,Q)} \mathop{\mathbb{E}}_{Z,Z' \sim M} [c(Z,Z')] \tag{4}$$

Here $c(.,.)$ is a non-negative, lower semi-continuous function satisfying $c(z,z) = 0$ and called the cost.

1.2.1 Relation using Proposition-1

Now that we have defined the distributionally robust problem and Wasserstein distance, we would like to show how this is related to the pointwise (actual) robust adversarial problem. For this we use the Proposition-1 of the paper.

$$\begin{align} \sup_{W_c(P,P_0) \leq \rho} \mathbb{E}_P[l(\theta; Z)] &= \inf_{\gamma \geq 0}\{\gamma\rho + \mathop{\mathbb{E}}_{P_0}[\phi_\gamma(\theta; Z)]\} \tag{5} \\ &\leq \gamma\rho + \mathop{\mathbb{E}}_{P_0}[\phi_\gamma(\theta; Z)] \tag{6} \\ & \tag{7} \end{align}$$

Now ignoring the constant $\gamma\rho$, we focus on the other term. We first define ϕ_γ

$$\phi_\gamma(\theta; z) = \sup_{z_0 \in Z}\{l(\theta; z) - \gamma c(z_0, z)\} \tag{8}$$

This can be visualized as the adversarial perturbation defined above (2) with just the extra constraint of being close to z_0. This implies that the worst case loss for distributions close to the true distribution can be bounded by this adversarial problem $\mathbb{E}_{P_0}[\phi_\gamma(\theta; Z)]$.

1.3 Modified Problem

From now on, we denote this adversarial problem as the modified problem:

$$\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{P_0} \left[\phi_\gamma(\theta; Z)\right] \tag{9}$$

and focus on solving the modified problem as this can bound our original distributional robust problem as seen on figure 2.

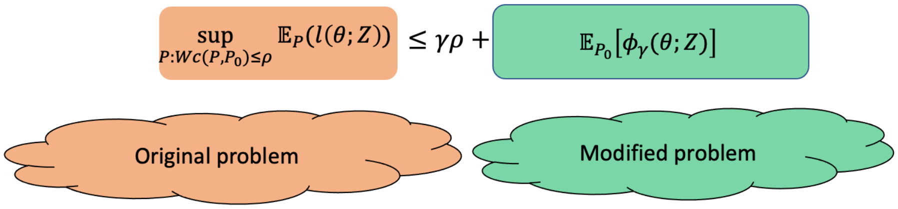

$\sup_{P: W_c(P,P_0) \leq \rho} \mathbb{E}_P(l(\theta; Z)) \leq \gamma\rho +$
$\mathbb{E}_{P_0}[\phi_\gamma(\theta; Z)]$
Original problem
Modified problem

Figure 2: Inequality between original and modified problem.

We now note that this modified problem is very similar to the usual risk minimization setup with just $l(\theta; z)$ replaced with $\phi_\gamma(\theta; z)$. Hence we would like our $\phi_\gamma(\theta; z)$ to be easily computable.

1.4 Computability of Robust Surrogate $\phi_\gamma(\theta; z)$

In order to make sure that the $\phi_\gamma(\theta; z)$ is easily computable the authors make the following assumptions on the cost function and the loss function:

Assumption A: $c(.,.)$ is continuous and $c(., z)$ is 1-strongly convex w.r.t some norm $||.||$.

Assumption B: $l(.,.)$ is smooth *i.e*,

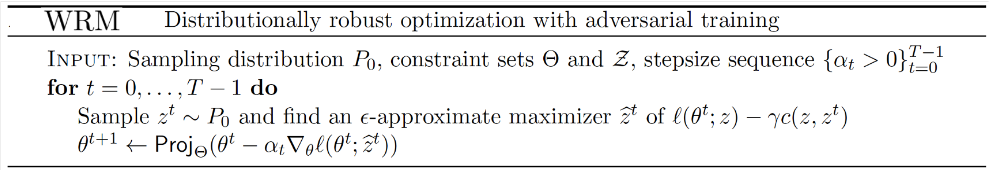
$$\begin{array}{ll} ||\nabla_\theta l(\theta; z) - \nabla_\theta l(\theta'; z)||_* \leq L_{\theta\theta}||\theta - \theta'|| & ||\nabla_z l(\theta; z) - \nabla_z l(\theta; z')||_* \leq L_{zz}||z - z'|| \\ ||\nabla_\theta l(\theta; z) - \nabla_\theta l(\theta; z')||_* \leq L_{\theta z}||z - z'|| & ||\nabla_z l(\theta; z) - \nabla_z l(\theta'; z)||_* \leq L_{z\theta}||\theta - \theta'|| \end{array}$$

with $||.||_*$ the dual norm. Using these assumptions it can be shown that for γ big enough, $l(\theta; z_0) - \gamma c(z_0, z)$ is 1-strongly concave w.r.t z_0 and hence the $\sup_{z_0}\{l(\theta; z_0) - \gamma c(z_0, z)\}$ is easily computable.

2 Results

2.1 Algorithm

To solve the modified problem (9) the authors have chosen the classical method of stochastic gradient descent, which is very popular in the field of machine learning. To do so, we can easily notice that the gradient of ϕ_γ according to θ is the gradient of l for the maximizer of $l(\theta; z) - \gamma c(z,.)$. We then end up with the following algorithm 3 which remains extremely similar to a risk minimization. This algorithm is therefore called Wasserstein Risk Minimization (WRM). Moreover, under assumptions



```
WRM    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 $\epsilon$-approximate maximizer $\widehat{z}^t$ of $\ell(\theta^t; z) - \gamma c(z, z^t)$
    $\theta^{t+1} \leftarrow \mathsf{Proj}_\Theta(\theta^t - \alpha_t \nabla_\theta \ell(\theta^t; \widehat{z}^t))$
```

Figure 3: WRM Algorithm.

A and B we can prove the theorem (10) which states the convergence of our algorithm WRM. The proof is based on the manipulation of classical inequalities with T the number of iterations:

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}_\theta(||\nabla(\mathbb{E}_{P_0}[\phi_\gamma(\theta,Z)])||)-\frac{4L_{\theta z}^2}{\gamma-Lzz}\leq O(\frac{1}{\sqrt{T}}) \tag{10}$$

Furthermore, this theorem is converging in expectancy quite quickly with a speed in $\frac{1}{\sqrt{T}}$. Thus, the authors have proven so far that their method gives quickly the right solution computationally. Now it need to be checked if the algorithm is indeed improving the robustness.

2.2 Robustness

We have already shown that the modified problem's population loss bounds the distributional robust problem's population loss (2). But we want to know whether the modified problem's empirical loss also bounds the distributional robust problem's population loss. Theorem-3 in the paper talks about this and proves the following:

$$\sup_{W_c(P,P_0)\leq\rho}\mathbb{E}_P[l(\theta;Z)]\leq\gamma\rho+\mathop{\mathbb{E}}_{\hat{P_n}}[\phi_\gamma(\theta;Z)]+\epsilon_n(t) \tag{11}$$

where $\epsilon_n(t):=\gamma b_1\sqrt{\frac{M_l}{n}}\int_0^1\sqrt{logN(\mathcal{F},M_l,\epsilon,||.||_{L^\infty(Z)})}d\epsilon+b_2M_l\sqrt{\frac{t}{n}}$ Now we try to dive deeper and understand how they have arrived at this result.
Let $\mathcal{F}=\{l(\theta;.):\forall\theta\in\Theta\}$ and $M_l=\sup_{\theta,z}l(\theta;z)$. We now note that this result has been proved by using Uniform law to relate the population loss and empirical loss of the modified problem. We know that the error term in uniform law is the rademacher complexity and is bounded by dudley's entropy integral bound. But the covering number used in this bound is w.r.t function space of ϕ_γ which is not $\mathcal{F}$. Hence we believe that they have used Rademacher contraction to use the covering number w.r.t $\mathcal{F}$ leading to the extra γ term outside the integral. But we were unable to prove that the mapping $l(\theta;z)\rightarrow\phi_\gamma(\theta;z)$ is a Lipschitz function.

2.3 Robustness for $\rho=\hat{\rho_n}(\theta)$

Now we would like to show how they have obtained the second result in Robustness Theorem in a detailed manner which is missing in the paper which will also be useful for us in the experiments. We first define the following as mentioned in the paper:

$$T_\gamma(\theta;z_0) := \arg\max_z\{l(\theta;z)-\gamma c(z,z_0)\} \tag{12}$$

$$P_n^*(\theta) := \arg\max_P\{\mathop{\mathbb{E}}_P[l(\theta;Z)]-\gamma W_c(P,\hat{P_n})\} \tag{13}$$

$$\hat{\rho_n}(\theta) := W_c(P_n^*(\theta),\hat{P}_n) \tag{14}$$

Based on these definitions, we note that $P_n^*(\theta)=\frac{1}{n}\sum_{i=1}^{n}\delta_{T_\gamma(\theta;Z_i)}$ and $\hat{\rho_n}(\theta)=\mathbb{E}_{\hat{P_n}}[c(T_\gamma(\theta;Z),Z)]$
Substituting $\rho=\hat{\rho_n}(\theta)$ in (11) we get

$$\sup_{W_c(P,P_0)\leq\hat{\rho}_n(\theta)}\mathbb{E}_P[l(\theta;Z)]\leq\gamma\hat{\rho}_n(\theta)+\mathop{\mathbb{E}}_{\hat{P_n}}[\phi_\gamma(\theta;Z)]+\epsilon_n(t) \tag{15}$$

Now from proposition-1 of the paper(5) W.K.T

$$\sup_{W_c(P,\hat{P}_n)\leq\rho}\mathbb{E}_P[l(\theta;Z)]=\inf_{\gamma\geq0}\{\gamma\rho+\mathop{\mathbb{E}}_{\hat{P_n}}[\phi_\gamma(\theta;Z)]\}$$

Now we differentiate the term inside inf to attain the minimum. We note that the optimum is attained at $\hat{\rho}_n(\theta)$. Hence using this in (15)

$$\sup_{W_c(P,P_0)\leq\hat{\rho}_n(\theta)}\mathbb{E}_P[l(\theta;Z)]\leq\sup_{W_c(P,\hat{P}_n)\leq\hat{\rho}_n(\theta)}\mathbb{E}_P[l(\theta;Z)]+\epsilon_n(t) \tag{16}$$

2.4 Generalization

In this section we will describe the Theorem-4 from the article which deals with generalization. The motivation for generalization is actually quite unclear in the paper and we will explain here how we understood the reason and the stakes of this result. To do so we have to look again at the quantities introduced previously and especially at $P_n^*(\theta)$ and $\hat{\rho}_n(\theta)$. We can interpret $P_n^*(\theta)$ as being the worst-case distribution and therefore $\hat{\rho}_n(\theta)$ is quantifying the level of robustness that we can achieve with our method. We then introduce the equivalent population quantites $P_0^*(\theta) := \arg\max_P\{\mathbb{E}_P\left[l(\theta; Z)\right] - \gamma W_c(P, \hat{P_0})\}$ and $\rho_0(\theta) := W_c(P_0^*(\theta), P_0)$.

Now if we focus on the equation (16) we can notice that for $\hat{\rho}_n(\theta)$, the empirical solution $\sup_{W_c(P,\hat{P}_n)\leq\hat{\rho}_n(\theta)} \mathbb{E}_P[l(\theta; Z)]$ approximates well $\sup_{W_c(P,P_0)\leq\hat{\rho}_n(\theta)} \mathbb{E}_P[l(\theta; Z)]$ and thus generalizes well. Yet, this generalization is actually not rigorous because these two quantities are defined for $\hat{\rho}_n(\theta)$ and not $\rho_0(\theta)$. It is therefore essential to investigate if the empirical level of robustness $\hat{\rho}_n(\theta)$ is generalizing well. This is the aim of the next result:

If the data is bounded, *i.e* $\exists M_z$ such that $||Z|| \leq M_z$ and c is L_c-Lipschitz. Then with probability higher than $1 - e^{-t}$:

$$\sup_{\theta\in\Theta} |\hat{\rho}_n(\theta) - \rho_0(\theta)| \leq 4L_c M_z \sqrt{\frac{1}{n}(t + \log N(\Theta, \frac{(\gamma - L_{zz})_+ t}{4L_c L_{z\theta}}, ||.||))} \tag{17}$$

It is important to notice for the improvement part that this result (17) was proved using Hoeffding's inequality.
This result proves that for a lot of possible covering numbers $\hat{\rho}_n(\theta)$ converges uniformly to $\rho_0(\theta)$ and hence generalizes.

2.5 Application to Neural Networks

All the previous results have been proved using mostly the assumptions A and B. We will justify in this paragraph how these assumptions are reasonable for Neural Networks.
First, assumption A is true for a lot of different costs and especially for the l_2 one. Second, assumption B for smoothness of the loss is verified for Neural Networks with softmax final loss and Lipschitz activations, which is true for a lot of classical functions like ELU, average pooling and sigmoid.
The proof of the smoothness of the loss is done using induction and basic calculus methods. If the reader wishes to go through the calculations, it is important to note that the middle line of factorization of equation (34) in the article is false. The first line and the result are true because of a telescopic sum manipulation.

3 Experiments

Now we are interested in checking the effectiveness of the algorithm and the robustness certificate practically. Before we describe the experiments in the paper, it is important to give a quick review of the other classical/contemprory methods of tackling adversaries.

3.1 Solving Inner Maximization Problem

If one examines the WRM algorithm in Figure (3), we could summarize the essence of the algorithm as follows:

```
for all data points in the sample do
    Step-1: Find an adversarial point close to the data point such that the loss is maximized
    Step-2: Perform gradient descent for $\theta$ w.r.t the newly obtained adversarial point
end
```

From our original definition of the adversarial problem

$$\min_{\theta\in\Theta} \mathbb{E}_{P_0} \sup_{u\in U}[l(\theta; Z + u)] \tag{18}$$

we note that the step-1 of the algorithm solves the inner maximization and step-2 solves the outer minimization. In general this is the essence of many adversarial algorithms such as FGSM, PGM, etc. wherein the Step-1 of the algorithm is achieved through various means.

3.1.1 FGSM

In FGSM the inner maximization problem is solved by adversarially perturbing the sample point using the gradient of the loss function as follows:

$$x \leftarrow x + \epsilon sign(\nabla_x l(\theta, x, y)) \tag{19}$$

3.1.2 PGM

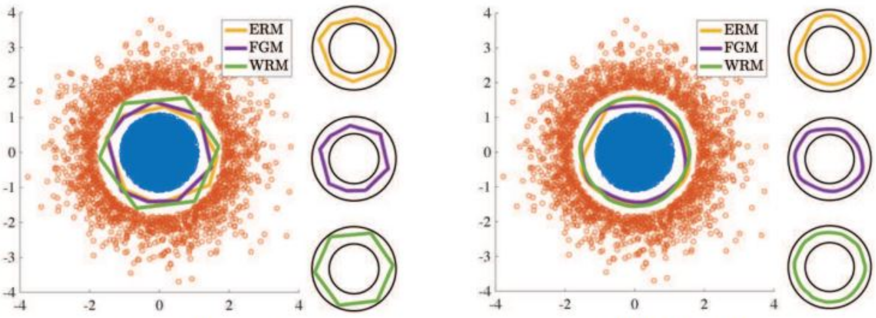
Unlike FGSM, this is a iterative process and at each step we project back into a ball around the original sample point

$$x_i^{t+1} \leftarrow \Pi(x_i^t + \alpha_t \Delta x_i^t(\theta)) \tag{20}$$

where Π denotes the projection onto $\{x : ||x - x_i||_p \leq \epsilon\}$ for $p = 2, \infty$ and $\Delta x_i^t(\theta) = \underset{||\eta||_p \leq \epsilon}{\arg\max}\{\nabla l(\theta; x_i^t, y)^T \eta\}$ Now we introduce the experimental setup and analyze how all these methods perform.

3.2 Experiment-1 (Synthetic Data)

The authors sampled data from Normal distribution i.e, $X \sim N(0_2, I_2)$ and decided labels based on radius as follows: $Y = sign(||x|| - \sqrt{2})$. Then this data was trained by a neural network with 2 hidden layers of size 4 and 2. They also removed data points between radii $\frac{1}{\sqrt{2}}$ and $\sqrt{2}$ so as to have clear separation of data. We now analyze the classification boundaries learnt by each algorithm from the above figure. We notice(Fig: 4) that the boundaries learnt from WRM are more uniform (especially for ReLU) w.r.t radii compared to the boundaries learnt by ERM and PGM. Since 70 % of the data lie in the blue circle, the adversaries mainly consist of blue points outside the boundary. Hence the symmetric WRM boundary lying farthest from the centre provides a robustness guarantee against adversarial perturbations.



(a) ReLU model

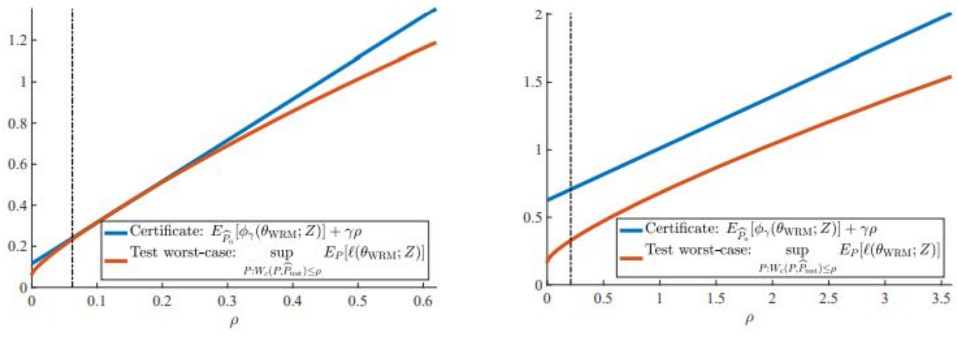
(b) ELU model

Figure 1. Experimental results on synthetic data. Training data are shown in blue and red. Classification boundaries are shown in yellow, purple, and green for ERM, FGM, and WRM respectively. The boundaries are shown with the training data as well as separately with the true class boundaries.

Figure 4: Experiment-1

From Theorem-3, it was proved that there exists a robustness certificate for the worst case population loss for the distributionally robust problem using the empirical loss of the modified problem.

$$\sup_{W_c(P,P_0)\leq\rho} \mathbb{E}_P[l(\theta; Z)] \leq \gamma\rho + \mathbb{E}_{\hat{P}_n}[\phi_\gamma(\theta; Z)] + \epsilon_n(t)$$



(a) Synthetic data

(b) MNIST

Figure 2. Empirical comparison between certificate of robustness (11) (blue) and test worst-case performance (red) for experiments with (a) synthetic data and (b) MNIST. We omit the certificate's error term $\epsilon_n(t)$. The vertical bar indicates the achieved level of robustness on the training set $\widehat{\rho}_n(\theta_{\text{WRM}})$.

Figure 5: Experiment-1

The graph (Fig:5) plots the test worst-case and the certificate (ignoring $\epsilon_n(t)$) for various ρ. We now want to describe how the test-worst case is calculated as it is not clearly described in the paper.

3.3 Calculating $\sup\limits_{W_c(P, \hat{P}_{test}) \leq \rho} \mathbb{E}_P[l(\theta; Z)]$

We first note that $\sup_{W_c(P, \hat{P}_{test}) \leq \rho} \mathbb{E}_P[l(\theta; Z)]$ is intractable for arbitrary ρ from Proposition-1.(5). Hence we first fix a γ and calculate $P^*_{test}(\theta) = \arg\max_P \{\mathbb{E}_P\,[l(\theta; z)] - \gamma W_c(P, \hat{P}_{test})\}$ and $\hat{\rho}_{test}(\theta) = W_c(P^*_{test}(\theta), \hat{P}_{test})$. We now note that for $\rho = \hat{\rho}_{test}(\theta)$

$$\begin{aligned} \sup_{W_c(P, \hat{P}_{test}) \leq \hat{\rho}_{test}(\theta)} \mathbb{E}_P[l(\theta; Z)] &= \inf_{\gamma \geq 0} \{\gamma \hat{\rho}_{test}(\theta) + \mathop{\mathbb{E}}_{\hat{P}_{test}}[\phi_\gamma(\theta; Z)]\} \\ &= \gamma \hat{\rho}_{test}(\theta) + \mathop{\mathbb{E}}_{\hat{P}_{test}}[\phi_\gamma(\theta; Z)] \quad \text{(as the minimum occurs at } \rho = \hat{\rho}_{test}(\theta)) \end{aligned}$$

We repeat this for various γ and form the graph of ρ vs $\sup\limits_{W_c(P, \hat{P}_{test}) \leq \rho} \mathbb{E}_P[l(\theta; Z)]$. And finally we observe that the certificate always upper bounds this worst case test loss for all ρ.

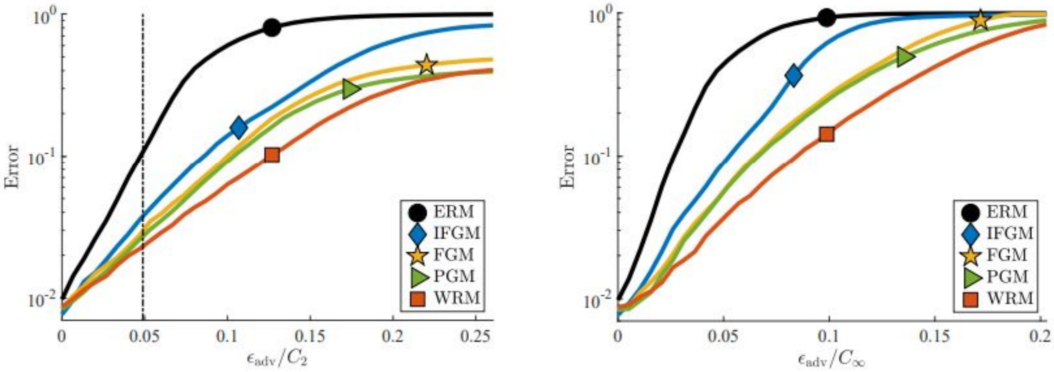
3.4 Experiment-2

For this experiment, the authors have created a Github repository https://github.com/duchi-lab/certifiable-distributional-robustness. We tried to run their code but had to use different packages and there was no 'requirements.txt' file. We can provide the details of the packages if the reader wants to rerun their code. In this experiment the authors use MNIST data and train a neural network with ELU activations. The values of the constants (γ, ρ, ϵ) are all clearly established in the paper. And the experimental setup is also clearly stated. We note from the graph (figure 6) that WRM performs the best in terms of error rate against all contemporary methods against PGM attacks of norm 2 and ∞.

In general,small values of ∇l is desired as it leads to more robust parameters w.r.t (adversarial)perturbations in data leading to stability of the model. We now want to focus on the relation between $\hat{\rho}$ and $\nabla l(\theta)$ which is not explained in detail in the paper.

3.5 ρ vs ∇l

We first recall that $\hat{\rho}_{test}(\theta) = \mathbb{E}_{\hat{P}_{test}}[c(T_\gamma(\theta; Z), Z)]$. The authors state that for small values of $\hat{\rho}_{test}(\theta)$, $\nabla_z l(\theta; z)$ is also small around the nominal input. We illustrate this with the figure 7



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

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

Figure 6: Experiment-2: PGM Attacks

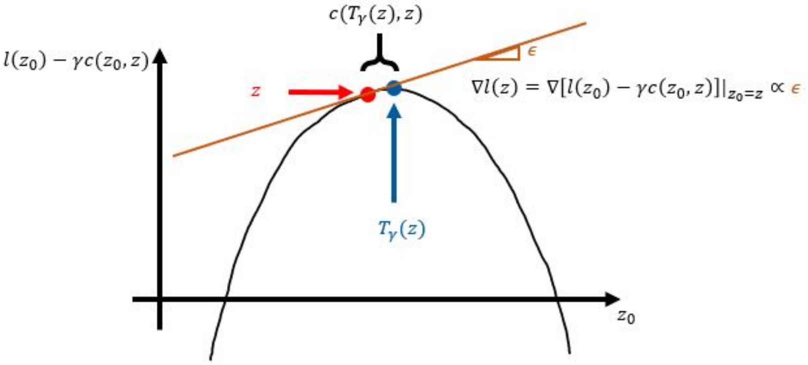

$c(T_\gamma(z), z)$
$l(z_0) - \gamma c(z_0, z)$
ϵ
z
$\nabla l(z) = \nabla[l(z_0) - \gamma c(z_0, z)]|_{z_0=z} \propto \epsilon$
$T_\gamma(z)$
z_0

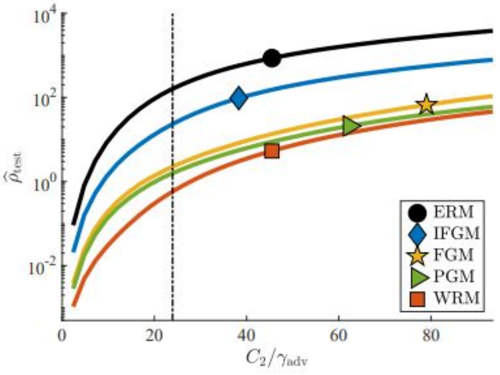
Figure 7: Experiment-2: Intuition

Here 'z' is sampled from $\hat{P}_{test}$ and we plot $l(z_0) - \gamma c(z_0, z)$ against z_0. We know that the maximum will be attained at $z_0 = T_\gamma(z)$ as defined in (12). As $\hat{\rho}_{test}(\theta)$ is the empirical mean of $c(T_\gamma(\theta; z), z)$, intuitively when $\hat{\rho}_{test}(\theta)$ is small, $c(T_\gamma(\theta; z), z)$ is also small. This implies z lies close to $T_\gamma(z; \theta)$ and the gradient of the function at z $\nabla_{z_0}(l(z_0) - \gamma c(z_0), z)|_{z_0=z}$ is also small as illustrated in the figure. But we know that $\nabla_z l(\theta; z) = \nabla_{z_0}(l(z_0) - \gamma c(z_0), z)|_{z_0=z}$ as cost function $c(z_0, z)$ attains its minimum value(0) at $z_0 = z$. Hence $\nabla_z l(\theta; z)$ is small for small values of $\hat{\rho}_{test}(\theta)$

We now plot $\hat{\rho}_{test}(\theta)$ for various values of γ(figure 8). For other methods $\hat{\rho}_{test}(\theta)$ is considered as the mean distance to the perturbed distribution. We notice that WRM has the smalledst $\hat{\rho}_{test}(\theta)$ for all values of γ leading to a more stable model compared to other methods. We finish this section by discussing one other experiment which is very intuitive in this setup(Appendix).

4 Discussion

This article details a procedure to improve the robustness of machine learning methods. This one has many advantages. It is simply a risk minimzatiom solved using stochastic gradient descent as detailed 3 and is therefore extremely simple to implement. Moreover, the authors proved theoretically and experimentally that this method improves the robustness against different attacks, which the classical adversarial perturbation methods has difficulties to achieve. Finally, this method is perfectly applicable to neural networks.



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

Figure 8: Experiment-2: $\hat{\rho}_{test}$

However, some weaknesses can be observed. First, γ must be sufficiently large to guarantee the existence of ϕ_γ. Thus, by duality, the method of the paper is robust only to small perturbations. Second, the generalization guarantees (17) that allow us to apply the robustness result (11) to our empirical algorithm use $||.||_\infty$-covering number which can be very large for neural networks.

5 Improvement

5.1 Dudley's bound for Theorem 4

In this paragraph we will derive two improvements of the bounds from the equation (17). Under the first assumption of (17) we have: c which is L_c-Lipschitz on the almost surely bounded (by M) domain Z. Thus, c is a.s. bounded by L_cM. Then using that $\hat{\rho}_n(\theta) = \mathbb{E}_{\hat{P}_n}[c(T_\gamma(\theta; Z), Z)]$ and $\rho_0(\theta) = \mathbb{E}_{P_0}[c(T_\gamma(\theta; Z), Z)]$ and applying uniform law we obtain $\forall t \in \mathbb{R}$:

$$\mathbb{P}(\sup_{\theta\in\Theta} |\hat{\rho}_n(\theta) - \rho_0(\theta)| \geq t + R_n(\mathcal{H})) \leq e^{-\frac{nt^2}{2L_c^2M^2}} \tag{21}$$

First Improvement: Now we can bound the Rademacher complexity using Dudley's entropy integral bound:

$$R_n(\mathcal{H}) \leq \frac{2}{\sqrt{n}} \inf_{\delta\geq 0} [\delta\sqrt{n} + 8 \int_{\frac{\delta}{4}}^{B} \sqrt{\log N(\Theta, u, ||.||)} du]$$

with $B = \sup_{\theta,\theta'} ||\theta - \theta'|| \leq diam(\Theta) =: D$.

Finally we obtain with $t = L_cM\sqrt{\frac{2\tau}{n}}$ and for probability greater than $1 - e^{-\tau}$ that:

$$\sup_{\theta\in\Theta} |\hat{\rho}_n(\theta) - \rho_0(\theta)| \leq \frac{2}{\sqrt{n}} \inf_{\delta\geq 0} [\delta\sqrt{n} + 8 \int_{\frac{\delta}{4}}^{B} \sqrt{\log N(\Theta, u, ||.||)} du] + L_cM\sqrt{\frac{2\tau}{n}} \tag{22}$$

Comparison with bound from the article: To compare this bound with the one given from the article we can study $\Theta \subset \mathbb{R}^d, \log N(\Theta, u, ||.||) \leq d\log(1 + \frac{D}{u})$ and look at both bounds. First for the Dudley's bound we have:

$$\begin{aligned}
\int_{\frac{\delta}{4}}^{B} \sqrt{\log N(\Theta, u, ||.||)} du &= \int_{\frac{\delta}{4}}^{B} \sqrt{d\log(1 + \frac{D}{u})} du \\
&\leq \int_{\frac{\delta}{4}}^{B} \sqrt{\frac{dD}{u}} du \text{ using concavity of the logarithm} \\
&\leq 2D\sqrt{d} - \sqrt{dD\delta} \text{ using that B is bounded by D}
\end{aligned}$$

Then we derive that:

$$\inf_{\delta\geq 0} f(\delta) = \delta\sqrt{n} + 8(2D\sqrt{d} - \sqrt{dD\delta}) = 16D\sqrt{d} - \frac{16dD}{\sqrt{n}}$$

Therefore,

$$R_n(\mathcal{H}) \leq \frac{2}{\sqrt{n}} \inf_{\delta\geq 0}[\delta\sqrt{n} + 8\int_{\frac{\delta}{4}}^{B} \sqrt{\log N(\Theta, u, ||.||)}du] \leq \frac{32}{\sqrt{n}}(D\sqrt{d} - \frac{dD}{\sqrt{n}}) \tag{23}$$

Finally, with probability greater than $1 - e^{-\tau}$

$$\sup_{\theta\in\Theta} |\hat{\rho}_n(\theta) - \rho_0(\theta)| \leq \frac{32D\sqrt{d}}{\sqrt{n}} + L_c M\sqrt{\frac{2\tau}{n}} := B_1 \tag{24}$$

If we use the bound from the article we have:

$$\sup_{\theta\in\Theta} |\hat{\rho}_n(\theta) - \rho_0(\theta)| \leq 4L_c M\sqrt{\frac{1}{n}(\tau + \log N(\Theta, \frac{[\gamma - L_{zz}]_+\tau}{4L_c L_{z\theta}}, ||.||))} \tag{25}$$

For $\log N(\Theta, u, ||.||) \leq d\log(1 + \frac{D}{u})$,

$$4L_c M\sqrt{\frac{1}{n}(\tau + \log N(\Theta, \frac{[\gamma - L_{zz}]_+\tau}{4L_c L_{z\theta}}, ||.||))} \leq 4L_c M\sqrt{\frac{1}{n}(\tau + d\log(1 + \frac{4DL_c L_{z\theta}}{[\gamma - L_{zz}]_+\tau}))} := B_2 \tag{26}$$

We can now compare B_1 our new bound derived from Dudley's entropy with B_2 the bound coming from the article. We can first notice that B_1 and B_2 are both behaving in $O(\frac{1}{\sqrt{n}})$.
Moreover we can see that for example with $\tau = 10$ (to have the inequality with high probability), $D = 0.1$ (small set), $\frac{[\gamma - L_{zz}]_+\tau}{4L_c L_{z\theta}} = 0.1$ (need to be little for the covering number), $L_c = M = 1$ and $d = 1$:

$$B_1 = 7.7 \leq B_2 = 13.1 \tag{27}$$

Therefore for little sets the Dudley's entropy integral bound could give tighter results.

Second improvement: We can derive a similar and tighter result as the previous paragraph if we are interested in classification problem. This is totally reasonable because to prove the smoothness assumption for Neural Networks we assumed that the final loss was the softmax function. Thus we can assume function taking values in $[0, 1]$ and containing 0. Then we can use the article [3]:

$$R_n(\Theta) \leq \inf_{\delta\geq 0}(\frac{4\delta}{\sqrt{n}} + \frac{12}{n}\int_{\delta}^{\sqrt{n}} \sqrt{\log N(\Theta, u, ||.||)}du) \tag{28}$$

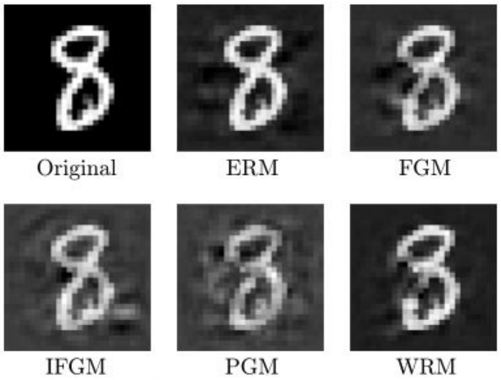
With the previous covering number $\log N(\Theta, u, ||.||) \leq d\log(1 + \frac{D}{u})$ we obtain:

$$\begin{aligned} R_n(\Theta) &\leq \inf_{\delta\geq 0}(\frac{4\delta}{\sqrt{n}} + \frac{12}{n}\int_{\delta}^{\sqrt{n}} \sqrt{\log N(\Theta, u, ||.||)}du) \text{ using concavity of the logarithm} \\ &\leq \frac{24}{n}[\sqrt{dD\sqrt{n}}] - \frac{36dD}{n\sqrt{n}} \\ &= O(n^{\frac{-3}{4}}) \end{aligned}$$

Finally we find a bound behaving in $L_c M\sqrt{\frac{2\tau}{n}} + O(n^{\frac{-3}{4}}) \approx L_c M\sqrt{\frac{2\tau}{n}}$ independent on the covering number, which completely cancels the last defect described in the discussion paragraph 4 when n is large enough.

Appendix

5.2 Experiment



(b) Perturbations on a test datapoint

Figure 9: Experiment-2: Decreasing γ

In this figure (9), the input is being adversarially perturbed w.r.t different methods until the model misclassifies it. In WRM this means decreasing γ. We notice that the perturbation misclassified by WRM is confusing for humans too.

5.3 Improvement

If we look again at the bound B_1 in equation 24 against B_2 in equation 26 we can notice that B_1 behaves in $O(D)$ and B_2 in $O(\sqrt{D})$. This the main reason why B_1 is smaller than B_2 only for small sets *i.e* with small D. We are convinced that we could actually improve B_1 if we used Jensen's inequality in the Dudley's bound instead of the concavity inequality $\log(1+x) \leq x$. As the $\inf_{\delta \geq 0}$ is not tractable, we could perhaps approximate this infimum experimentally to derive a new bound B_1, which should behave less harshly with D.

References

[1] Moosavi-Dezfooli, S-M, Fawzi, A and Frossard, P. (2016) DeepFool: a simple and accurate method to fool deep neural networks. *arXiv*.

[2] Tramèr, F, Boneh, D. (2019) Adversarial Training and Robustness for Multiple Perturbations *arXiv*.

[3] Bartlett, P, Foster, D.J. and Telgarsky, M. (2017) Spectrally-normalized margin bounds for neural networks. *arXiv*.