Achieving robustness in classification using optimal transport with hinge regularization

Mathieu Serrurier* Université Paul Sabatier
Franck Mamalet† IRT Saint-Exupery
Alberto González-Sanz Université Paul Sabatier
Thibaut Boissin IRT Saint-Exupery
Jean-Michel Loubes Université Paul Sabatier
Eustasio del Barrio Universidad de Valladolid

Abstract

Adversarial examples have pointed out Deep Neural Network's vulnerability to small local noise. It has been shown that constraining their Lipschitz constant should enhance robustness, but make them harder to learn with classical loss functions. We propose a new framework for binary classification, based on optimal transport, which integrates this Lipschitz constraint as a theoretical requirement. We propose to learn 1-Lipschitz networks using a new loss that is an hinge regularized version of the Kantorovich-Rubinstein dual formulation for the Wasserstein distance estimation. This loss function has a direct interpretation in terms of adversarial robustness together with certifiable robustness bound. We also prove that this hinge regularized version is still the dual formulation of an optimal transportation problem, and has a solution. We also establish several geometrical properties of this optimal solution, and extend the approach to multi-class problems. Experiments show that the proposed approach provides the expected guarantees in terms of robustness without any significant accuracy drop. The adversarial examples, on the proposed models, visibly and meaningfully change the input providing an explanation for the classification.

1. Introduction

The important progress in deep learning has led to a massive interest for these approaches in industry. However, when applying machine learning to critical tasks such has in the transportation or the medical domain, empirical and theoretical guarantees are required. Unfortunately, it has been shown that neural networks are weak to adversarial attacks: a carefully chosen small shift to the input, usually indistinguishable from noise, can change the class prediction [30]. This sensitivity to adversarial attacks is mainly due to the Lipschitz constant of a neural network which can be arbitrarily high when unconstrained. Most of white-box attacks (where the full model is available) take advantage of it to build adversarial examples by using gradient descent with respect to the input variables. *FGSM* [13] performs only one step of gradient descent when other approaches such as *PGD* [20, 6] find the optimal adversarial example iteratively. In black-box scenarios, gradients or logits of the model are not available. In such case, attacks start from large perturbations and then reducing it step by step (see for instance, boundary attacks [5] and pointwise attacks [28]).

There are three major types of strategy to address the issue of adversarial attacks. Agnostic defenses are independent of the model and consist of altering the input or the prediction. For instance, Cohen et al. obtain a provable certificate with respect to l_2 norm by using Gaussian random smoothing [8]. DEFENSE-GAN [27] uses a GAN to transform the input into the closest non-adversarial one at inference time. The second group of strategies relies on saddle point optimization by adding a penalty term measuring the empirical weakness against adversarial example during the learning process [20]. The last type of approaches focus on the Lipschitz constant of the network. It has been proven that bounding the Lipschitz constant of a neural network provides certifiable robustness guarantees against local adversarial attacks [15, 23], improves generalizations [29] and the interpretability of the model [31]. This constraint can be achieved layer by layer by using spectral normalization [7] or non-

*corresp. author: mathieu.serrurier@irit.fr
†corresp. author: franck.mamalet@irt-saintexupery.com

expansive layer [24]. In [18], Li et al. go beyond the Lipschitz constant bounding by requiring layers to be gradient norm preserving. Combined with hinge loss, it allows them to achieve stronger robustness certification. However, the main limitation of this approach relies in the link between the hinge margin parameter and the robustness of the network.

In this paper we propose a new classification framework based on optimal transport that integrates the Lipschitz constant and the gradient norm preserving constraint as a theoretical requirement. To the best of our knowledge, very few researches investigate the link between binary classification and optimal transport (in [10], Frogner et al. use Wasserstein loss to improve multilabel classification). In Wasserstein GAN [2], the k-Lipschitz networks used to measure the distance between two distributions act like a discriminator, in analogy with the initial GAN algorithm [12]. The Wasserstein distance is approximated using a loss based on the Kantorovich-Rubinstein dual formulation and a k-Lipschitz network constrained by weight clipping. However, as we will demonstrate, a vanilla classifier based on the Kantorovich-Rubinstein loss is suboptimal, even on toy datasets.

We propose a binary classifier based on a regularized version of Kantorovich-Rubinstein formulation using a hinge loss term. We show that it remains the dual of an optimal transport problem, and we prove that the optimal solution of the problem exists and makes no error when the classes are well separated. With this new optimization problem, we guarantee to have an accurate classifier with a loss that is defined on and takes advantage of 1-Lipschitz function. As in [18], we bound the Lipschitz constant of the linear layers by Björck normalization and use norm preserving activation functions [1]. However, the optimal transport interpretation of the problem makes the bridge between these constraints and the loss function. When solving this optimal transport problem, attacking a prediction corresponds to travel along the transport plan up to the decision frontier. The output of the optimal network is linked to the length of this path, which is maximized during the optimization process of the proposed loss.

The paper, and the contributions, are structured as follows. In Section 2, we recall the definition of Wasserstein distance and the dual optimization problem associated. We present the interesting properties of a classifier based on this approach, illustrate that it leads to a suboptimal classifier even on a toy dataset. Section 3 describes the proposed binary classifier, based on a regularized version of the Kantorovich-Rubinstein loss with a hinge regularization. We show that the primal of this classification problem is a new optimal transport problem and we demonstrate different mathematical properties of our approach. Section 4 is devoted to the way of constraining the classifier to be 1-Lipschitz and how to generalize the approach to multi classification problems. Section 5 presents the results of experiments on MNIST, Cifar10 and CelebA datasets, measuring and comparing the results of different approaches in terms of accuracy and robustness. Last, we demonstrate that with our approach, building an adversarial example requires explicitly changing the example to an in-between two-classes image, which correspond to a point halfway on the transport plan. Proofs, computations details and additional experiments are reported in the appendix.

2. Wasserstein distance and Kantorovich-Rubinstein classifier

In this paper we only consider the Wasserstein-1 distance, also called Earth-mover, and noted $\mathcal{W}$ for $\mathcal{W}_1$. The 1-Wasserstein distance between two probability distributions μ and ν in Ω, and its dual formulation by Kantorovich-Rubinstein duality [32], is defined as the solution of:

$$\mathcal{W}(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathop{\mathbb{E}}_{x,z\sim\pi} \parallel \mathbf{x}-\mathbf{z} \parallel \tag{1a}$$

$$= \sup_{f \in Lip_1(\Omega)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} [f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\nu} [f(\mathbf{x})] \tag{1b}$$

where $\Pi(\mu,\nu)$ is the set of all probability measures on $\Omega\times\Omega$ with marginals μ and ν and $Lip_1(\Omega)$ denotes the space of 1-Lipschitz functions over Ω. Although, the infimum in Eq. (1a) is not tractable in general, the dual problem can be estimated through the optimization of a regularized neural network. This approach has been introduced in WGAN [2] where $Lip_1(\Omega)$ is approximated by the set of neural networks with bounded weights (better approximations of $Lip_1(\Omega)$ will be discussed in Section 4).

We consider a binary classification problem on feature vector space $X \subset \Omega$ and labels $Y = \{-1, 1\}$. We name $P_+ = \mathbb{P}(X|Y = 1)$ and $P_- = \mathbb{P}(X|Y = -1)$, the conditional distributions with respect to Y. We note $p = P(Y = 1)$ and $1 - p = P(Y = -1)$ the apriori class distribution. The classification problem is balanced when $p = \frac{1}{2}$.

In WGAN, [2] proposed to use the learned neural network (denoted $\hat{f}$ in the following), by maximizing the Eq. (1b), as a discriminator between fake and real images, in analogy with GAN [12]. To build a classifier based on $\hat{f}$, one can simply note that if f^* is an optimal solution of Eq. (1b), then $f^*+C, C \in \mathbb{R}$, is also optimal.

Centering the function f^* (resp. $\hat{f}$), Eq. (2), enables classification according to the sign of $f_c^*(x)$ (resp. $\hat{f}_c$ for the empirical solution).

$$f_c^*(\mathbf{x}) = f^*(\mathbf{x}) - \frac{1}{2}\left(\underset{\mathbf{z}\sim P_+}{\mathbb{E}}[f^*(\mathbf{z})] + \underset{\mathbf{z}\sim P_-}{\mathbb{E}}[f^*(\mathbf{z})]\right). \quad (2)$$

Such a classifier would exhibit good properties in terms of robustness for two main reasons: First, it has been shown in [32] that the function f^* is directly related to the cost of transportation between two points linked by the transportation plan as follows:

$$\mathbb{P}_{\mathbf{x},\mathbf{z}\sim\pi^*}(f^*(\mathbf{x}) - f^*(\mathbf{z}) = ||\mathbf{x}-\mathbf{z}||) = 1. \quad (3)$$

Second, it was shown in [14, 1], that this optimal solution also induces a property stronger than 1-Lipschitz:

$$||\nabla f^*|| = 1 \text{ almost surely on the support of } \pi^*. \quad (4)$$

However, applying this vanilla classifier (Eq. (2)) to a toy dataset such as the two-moons problem, leads to a poor accuracy. Indeed, Figures 1a and 1b present respectively the distribution of the values of $\hat{f}_c(x)$ conditionally to the classes and the level map of $\hat{f}_c$. We can observe that, even if the classes are easily separable, the distributions of the values of $\hat{f}_c$ conditionally to the class overlap. Thus, the 0-level threshold on $\hat{f}_c$ does not correspond to the optimal separator (even if it is better than random). Intuitively, $\hat{f}_c$ maximizes the difference of the expectancy of the image of the two distributions but do not try to minimize their overlap (Fig. 1a).

3. Hinge regularized Kantorovich-Rubinstein classifier

3.1. Definitions and primal transportation problem

In order to improve the classification abilities of the classifier based on Wasserstein distance, we propose a Kantorovich-Rubinstein optimization problem regularized by an hinge loss :

$$\begin{aligned}\sup_{f\in Lip_1(\Omega)} &-\mathcal{L}_\lambda^{hKR}(f) = \\ &\inf_{f\in Lip_1(\Omega)} \underset{\mathbf{x}\sim P_-}{\mathbb{E}}[f(\mathbf{x})] - \underset{\mathbf{x}\sim P_+}{\mathbb{E}}[f(\mathbf{x})] \\ &+ \lambda \underset{\mathbf{x}}{\mathbb{E}}(1 - Yf(\mathbf{x}))_+\end{aligned} \quad (5)$$

where $(1 - \mathbf{y}f(\mathbf{x}))_+$ stands for the hinge loss $max(0, 1-\mathbf{y}f(\mathbf{x}))$ and $\lambda >= 0$. We name $\mathcal{L}_\lambda^{hKR}$ the hinge-KR loss. The goal is then to minimize this loss with an 1-Lipschitz neural network. When $\lambda = 0$, this corresponds to the Kantorovich-Rubinstein dual optimization problem. Intuitively, the 1-Lipschitz function f^* optimal with respect to Eq. (5) is the one that both separates the examples with a margin and spreads as much as possible the image of the distributions. When using only an hinge loss (as in [18] for instance), the examples outside the margin are no more taken into consideration. If the margin is increased to cover all the examples and if the class are equally distributed, the hinge loss becomes equivalent to the Kantorovich Rubinstein loss and then leads to a weak classifier.

In the following, we introduce Theorems that prove the existence of such an optimal function f^* and important properties of this function. Demonstrations of these theorems are in Appendix B.

Theorem 1 (Solution existence)**.** *For each $\lambda > 0$ there exists at least a solution f^* to the problem*

$$f^* := f_\lambda^* \in \arg\min_{f\in Lip_1(\Omega)} \mathcal{L}_\lambda^{hKR}(f).$$

Moreover, let ψ be an optimal transport potential for the transport problem from P_+ to P_-, f^* satisfies that

$$||f^*||_\infty \leq M := 1 + \text{diam}(\Omega) + \frac{L_1(\psi)}{\inf(p, 1-p)}. \quad (6)$$

The next theorem establishes that the Kantorovich-Rubinstein optimization problem with hinge regularization is still a transportation problem with relaxed constraints on the joint measure (which is no longer a joint probability measure).

Theorem 2 (Duality)**.** *Set $P_+, P_- \in \mathcal{P}(\Omega)$ and $\lambda > 0$, then the following equality holds*

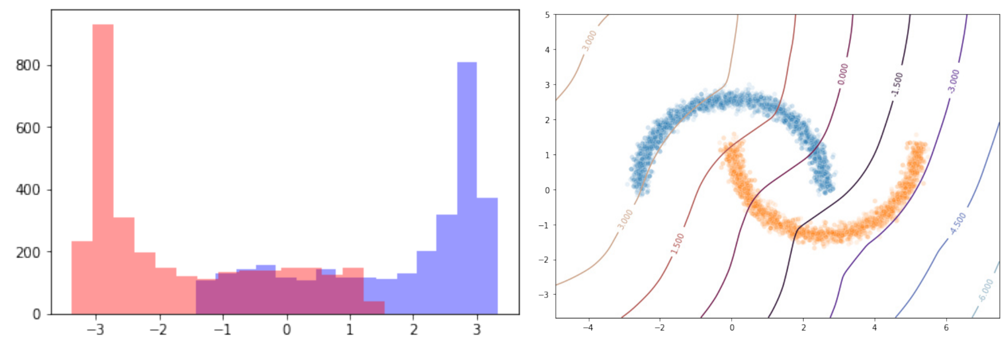
$$\begin{aligned}\sup_{f\in Lip_1(\Omega)} -\mathcal{L}_\lambda^{hKR}(f) = \inf_{\pi\in\Pi_\lambda^p(P_+,P_-)} \int_{\Omega\times\Omega} |\boldsymbol{x}-\boldsymbol{z}| d\pi \\ + \pi_{\boldsymbol{x}}(\Omega) + \pi_{\boldsymbol{z}}(\Omega) - 1\end{aligned} \quad (7)$$

Where $\Pi_\lambda^p(P_+, P_-)$ is the set consisting of positive measures $\pi \in \mathcal{M}_+(\Omega\times\Omega)$ which are absolutely continuous with respect to the joint measure $dP_+ \times dP_-$ and $\frac{d\pi_x}{dP_+} \in [p, p(1+\lambda)]$, $\frac{d\pi_z}{dP_-} \in [1-p, (1-p)(1+\lambda)]$.

3.2. Classification and geometrical properties

We note $\hat{f}$ the solution obtained by minimizing $\mathcal{L}_\lambda^{hKR}$ on a set of labeled examples and f^* the solution of Eq. (5). We don't assume that the solution found is optimal (i.e. $\hat{f} \neq f^*$) but we assume that $\hat{f}$ is 1-Lipschitz. Given a function f, a classifier based on $sign(f)$ and an example x, an adversarial example is defined as follows:

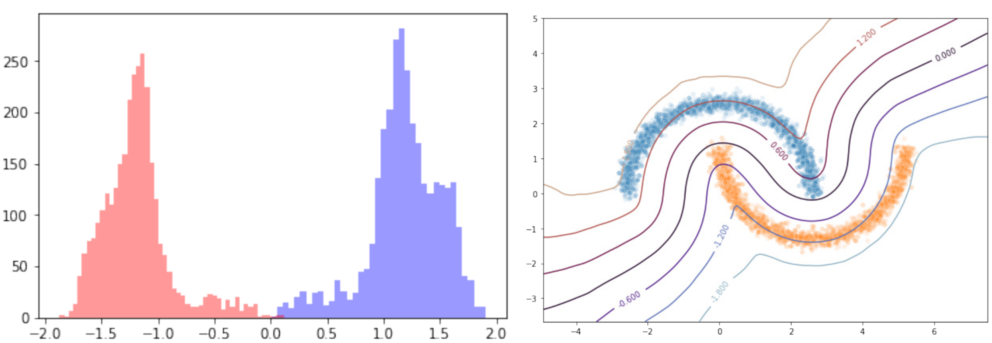
$$adv(f, \mathbf{x}) = \underset{\mathbf{z}\in\Omega | sign(f(\mathbf{z})) = -sign(f(\mathbf{x}))}{argmin} \parallel \mathbf{x} - \mathbf{z} \parallel. \quad (8)$$



(a) Distribution of $\hat{f}_c$ conditionally to the classes

(b) Level map of $\hat{f}_c$

Figure 1: Wasserstein classification (Eq. (2)) on the two moons.



(a) Distribution of $\hat{f}$ conditionally to the classes.

(b) Level map of $\hat{f}$

Figure 2: Hinge regularized Kantorovich-Rubinstein (hinge-KR) classification on the two moons problem

According to the 1-Lipschitz property of $\hat{f}$ we have

$$|\hat{f}(\mathbf{x})| \leq |\hat{f}(\mathbf{x}) - \hat{f}(adv(\hat{f}, \mathbf{x}))| \leq \| \mathbf{x} - adv(\hat{f}, \mathbf{x}) \| \ . \quad (9)$$

So $|\hat{f}(\mathbf{x})|$ is a lower bound of the distance of x to the separating boundary defined by $\hat{f}$ and thus a lower bound to the robustness to l_2 adversarial attacks. Thus, by minimizing $\mathbb{E}\left((1 - \mathbf{y}f(\mathbf{x}))_+\right)$, we maximize the accuracy of the classifier and by maximizing the discrepancy of the image of P_+ and P_- with respect to f we maximize the robustness with respect to adversarial attack. The proposition below establishes that the gradient of the optimal function with respect to Eq. (5) has norm 1 almost surely, as for the unregularized case (Eq. (4)).

Proposition 1. *Let π be the optimal measure of the dual version (7) of the hinge regularized optimal transport problem. Suppose that it is absolutely continuous with respect to Lebesgue measure. Then there exists at least a solution f^* of (7) such that $||\nabla f^*|| = 1$ almost surely.*

Furthermore, empirical results suggest that given $\mathbf{x}$, the image $tr_{f^*}(\mathbf{x})$ of $\mathbf{x}$ by transport plan and $adv(\mathbf{x})$ are in the same direction with respect to $\mathbf{x}$ and the direction is $-\nabla_x f^*(\mathbf{x})$. Combining this direction with the Eq. (9), we will show empirically (sect. 5) that

$$adv(\mathbf{x}) \approx x - c_x * f^*(\mathbf{x}) * \nabla_x f^*(\mathbf{x})$$

and

$$tr_{f^*}(\mathbf{x}) \approx x - c'_x * f^*(\mathbf{x}) * \nabla_x f^*(\mathbf{x})$$

with $1 \leq c_x \leq c'_x \in \mathbb{R}$. It turns out that this corresponds to the definition of FGSM attacks [13]. This suggests that in our frameworks, adversarial attacks amount to travel along the transportation path from the example to its transportation image.

The next proposition shows that, if the classes are well separated, maximizing the hinge-KR loss leads to a perfect classifier.

Proposition 2 (Separable classes)**.** *Set $P_+, P_- \in \mathcal{P}(\Omega)$ such that $P(Y = +1) = P(Y = -1)$ and $\lambda \geq 0$ and suppose that there exists $\epsilon > 0$ such that*

$$|\boldsymbol{x} - \boldsymbol{z}| > 2\epsilon \quad dP_+ \times dP_- \ \textit{almost surely} \qquad (10)$$

Then for each

$$f_\lambda \in \arg\sup_{f \in Lip_{1/\epsilon}(\Omega)} \int_\Omega f(dP_+ - dP_-) - \lambda\left(\int_\Omega (1-f)_+ dP_+ + \int_\Omega (1+f)_+ dP_-\right),$$

it is satisfied that $L_1(f_\lambda) = 0$. *Furthermore if* $\epsilon \geq 1$ *then* f_λ *is an optimal transport potential from* P_+ *to* P_- *for the cost* $|\boldsymbol{x} - \boldsymbol{z}|$.

We show in Fig. 2, on the two moons problem, that in contrast to the vanilla classifier based on Wasserstein (Eq. (2)), the proposed approach enables non overlapping distributions of $\hat{f}$ conditionally to the classes (Fig. 2a). In the same way, the 0-level cut of $\hat{f}$ (Fig. 2b) is a nearly optimal classification boundary. Moreover, the level cut of $\hat{f}$, on the support of the distributions, is close to the distance to this classification boundary.

4. Architecture

4.1. 1-Lipschitz gradient norm preserving network

In order to build a deep learning classifier based on the hinge-KR optimization problem (Eq. (5)), we have to constrain the Lipschitz constant of the neural network to be equal to 1. It is known that evaluating it exactly is a NP-hard problem [33]. The simplest way to constraint a network to be 1-Lipschitz is to impose this 1-Lipschitz property to each layer. For dense layers, the initial version of WGAN [2] consisted of clipping the weights of the layers. However, this is a very crude way to upper-bound Lipschitz constant. Normalizing by the Frobenius norm has also been proposed in [26]. In this paper, we use spectral normalization as proposed in [21], since the spectral norm is equal to the Lipschitz constant of the layer. At the inference step, we normalize the weights of each layer by the spectral norm of the matrix. This spectral norm is computed by iteratively evaluating the largest singular value with the power iteration algorithm [11]. This is done during the forward step and taken into account for the gradient computation. In the case of 2D-convolutional layers, normalizing by the spectral norm of convolution kernels is not enough and a supplementary multiplicative constant Λ is required (the regularization is then done by dividing W by $\Lambda||W||$). We propose, for zero padding, a tighter estimation of Λ than the one proposed in [7], computing the average duplication factor of non zero padded values in the feature map:

$$\Lambda = \sqrt{\frac{(k.w - \bar{k}.(\bar{k}+1)).(k.h - \bar{k}.(\bar{k}+1))}{h.w}} \tag{11}$$

for a kernel size equals to $k = 2 * \bar{k} + 1$. Even if this constant doesn't provide a strict upper bound of the Lipschitz constant (for instance, when the higher values are located in the center of the picture), it behaves very well empirically. Convolution with stride, pooling layers, detailed explanations and demonstrations are discussed in Appendix C.3.

As shown in Property 1, the optimal function f^* with respect to Eq. (5), verifies $||\nabla f^*|| = 1$ almost surely (*gradient norm preserving* (GNP) architecture [1]). We apply the approach described in [1], based on the use of specific activation functions and a process of normalization of the weights. Two norm preserving activation functions are proposed: i) **GroupSort2** : sorting the vector by pairs, ii) **FullSort** : sorting the full vector. These functions are vector-wise rather than element-wise. We also use the P-norm pooling [4], with $P = 2$ which is a norm-preserving average pooling. Concerning linear functions, a weight matrix W is gradient norm preserving if and only if all the singular values of W are equals to 1. In [1], the authors propose to use the Björck orthonormalization algorithm [3]. This algorithm is fully differential and, as for spectral normalization, is applied during the forward inference, and taken into account for back-propagation (see Appendix C.4 for details). We don't consider BCOP [18], which performs slightly better that Björck for convolution but at an higher computation cost. We developed a full tensorflow [9] implementation in an opensource library, called *DEEL.LIP*[1], that enables training of k-Lipschitz neural networks, and exporting them as standard layers for inference.

4.2. Multi-class hinge-KR classifier

To adapt our approach to the multi-class case, we propose to use q binary one-vs-all classifiers, where q is the number of classes. The set of labels is now $Y = \{C_1, \ldots, C_q\}$. We name $P_k = \mathbb{P}(X|Y = C_k)$ and $\neg P_k = \mathbb{P}(X|Y \neq C_k)$ the conditional distributions with respect to Y. We obtain the following optimization problem :

$$\sup_{f_1,\ldots,f_q \in Lip_1(\Omega)} -\mathcal{L}_\lambda^{hKR}(f_1,\ldots,f_q)$$

where

$$\mathcal{L}_\lambda^{hKR}(f_1,\ldots,f_q) = \sum_{k=1}^{q}\left[\underset{\mathbf{x}\sim\neg P_k}{\mathbb{E}}[f_k(\mathbf{x})] - \underset{\mathbf{x}\sim P_k}{\mathbb{E}}[f_k(\mathbf{x})] + \lambda\underset{\mathbf{x}}{\mathbb{E}}\left(m - (2 * \underset{y=C_k}{\mathbb{1}} - 1).f_k(x)\right)_+\right]. \tag{12}$$

[1] https://github.com/deel-ai/deel-lip

The global architecture is the same as the binary one except that the last layer has q outputs. For this last layer, each weight corresponding to each output neuron are normalized independently creating q 1-Lipschitz functions with gradient norm preserving properties. With this architecture, the optimal transport interpretation is still valid. The class predicted corresponds to the argmax of the classifier functions. With this approach, the provable robustness lower-bound is a half of the difference between the max and the second max values of $\{f_1(x), \dots, f_q(x)\}$.

In [18], Li et al. use classical multi-class hinge loss based on un-centered margin and apply constraint on the entire last layers rather than doing it independently. It allows to a lower provable adversarial robustness bound than ours. However, the $f_1, \dots, f_q$ obtained in this setting are no more 1-Lipschitz one by one, making the adversarial robustness bounds not comparable directly (see Appendix C.5).

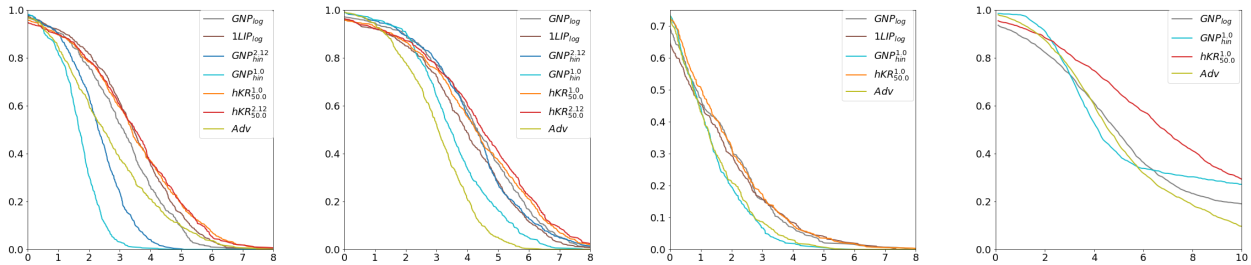
5. Experiments

In the experiment, we compare five approaches: i) Adv for Adversarial learning as in [20] , ii) $1LIP_{log}$ for log-entropy classifier with Björck orthonormalization and ReLU activation functions similar to Parseval networks [7], iii) GNP^m_{hin} for gradient norm preserving classifiers based on hinge loss with margin m as done in [18] iv) GNP_{log} for gradient norm preserving classifiers based on log entropy loss and v) hKR^m_α for gradient norm preserving classifiers based on the proposed hinge-KR with margin m and coefficient α. Note that, to the best of our knowledge, the GNP_{log} hasn't been applied yet for adversarial defenses. To have a fair comparison, all the classifiers share the same dense or convolutional architectures except for the weight normalization and the activation functions. We set α to 50 and margin m to 1 except for MNIST where we also consider $m = 2.12$ to be comparable with the experiments in [18]. For the GNP classifiers, we apply Björck orthonormalization (15 steps with p=1) after the spectral normalization (this improves the convergence of the Björck algorithm). We use fullsort activation function for dense layers, and GroupSort2 for the convolutional ones. Appendix D.1 provides the full description of the architecture and the optimization process.

We consider three classification problems, two multiclass problems (MNIST [17] and CIFAR-10 [16]) and a binary one (eyeglasses detection in Celeba-A dataset [19]). For MNIST and CIFAR-10, we use standard configurations with 10 classes and no data augmentation, 50000 examples in the training set and 10000 examples in the test set. In the binary problem, we consider the separation between people with or without eyeglasses on the CelebA dataset with 128x128x3 centered images. This is an unbalanced classification problem with 16914 examples in the training set (38% with eyeglasses) and 16914 examples in the test set (38% with eyeglasses).

Figure 3 presents the robustness against l_2 attacks with *DeepFool* attack [22] on the different datasets. We use this type of attack since none of the tested approaches is specifically built to resist to it. We can observe that the hKR^m_α approach is at the top of robustness on all the dataset and systematically better than GNP_{log} on all datasets and GNP^m_{hin} on the multiclass problem for the same value of m. On the CelebA dataset, GNP^m_{hin} and the adversarial approach Adv start with a better accuracy, but don't resist to large attacks. GNP_{log} and $1LIP$ have good performances on MNIST and CIFAR but their performances decrease when the models are deeper as for CelebA dataset ($1LIP$ was unable to converge on CelebA). To compare the different defense methods, we also apply a combination of the state-of-the art attacks $FGSM$ [13], PGD (l_2PGD) [20] and Carlini and Wagner (l_2CW) [6] on 500 images of the test set. All attacks are performed with the foolbox library [25]. For each value of ϵ, we run all the attacks and consider it as a success if at least one of them has succeeded. The results are presented in Figure 4 and Table 1 details the robustness values for CelebA. The results confirm and amplify the ones obtained with deepfool attacks. hKR^m_α obtain the highest level of robustness in all the situations for low and high values of ϵ. Adv has the highest robustness w.r.t. $FGSM$ attacks, since it was designed against to, but is a bit less resistant w.r.t. l_2PGD and weak w.r.t. l_2CW. hKR^m_α is especially strong w.r.t. Carlini and Wagner attack even with high values of ϵ. The discrepancy between the robustness of the approaches increase when models become deeper. On the CelebA dataset, GNP_{log} have difficulties to resist to higher noise. As expected, hKR^m_α seems to take most advantage of the gradient norm preserving architecture and obtains robustness against large attacks with an acceptable decrease of accuracy without noise.

Moreover, Table 1 shows that, for the proposed solution, accuracies w.r.t. ϵ are similar regardless the attacks. This suggests that optimal attacks are the same, and they are in the direction of the gradient of the classifier. This confirms the intuition pointed out in section 3.2 that optimal attack consists on traveling along the optimal transport map. In Figure 5, we compare adversarial images obtained with l_2 deepfool for the different models. The first row corresponds to the initial image. The following rows except the last one, are pictures obtained after attacks for the differ-



(a) MNIST dense (b) MNIST CNN (c) CIFAR (d) CelebA

Figure 3: Accuracy (Y-axis) w.r.t. of l_2 norm of fooling noise with deepfool attack (X-axis) on 2000 images of the test set

ent models. We design the attacks to push the image just beyond the classification boundary (50/50). For the Madry et al. approach (Adv) the noise is barely detectable. The noise is more visible with the gradient preserving approaches GNP_{log} and GNP^m_{hin} but it is still hard to interpret and sometimes meaningless. In contrast to the other approaches, the noise required to change the output class for the hKR classifier is highly structured, and interpretable. For people without glasses, the attack adds noise around the eyes and at the top of the nose. For people with eyeglasses, the attack tends to erase the glasses around the eyes and at the top of the nose.

The last row corresponds to the scheme proposed in Section 3.2 $att(\hat{f}, \mathbf{x}) \approx x - 2 * \hat{f}(\mathbf{x}) * \nabla_x \hat{f}(\mathbf{x})$ where $\hat{f}$ is the hKR^1_{50} model. Indeed, if $\nabla_x \hat{f}(\mathbf{x}) = 1$ and $\hat{f}(\mathbf{x})$ is the distance between the $\mathbf{x}$ and its image with respect to the transport, we expect to have $att(\hat{f}, \mathbf{x}) = tr(\hat{f}, \mathbf{x})$. The shifts are similar to the attacks and even more meaningful. This confirms that attacks against hKR can be interpreted as a transportation from a class to the other one and then requires to explicitly change the transport image in the opposite class. This suggests that attacks can explain classification for the hKR model. Moreover, the gradient of this model allows to build this explanation without relying on time-consuming algorithms.

	ϵ	Adv	GNP_{hin}	GNP_{log}	hKR_{50}
Base	0	**98.04**	97.2	94.51	96.07
$FGSM$	2	**90.52**	86.79	81.40	87.44
	5	**82.96**	37.70	43.90	64.01
l_2GPD	2	81.98	88.65	84.17	**88.75**
	5	61.35	34.90	48.02	**69.79**
l_2CW	2	32.14	80.80	79.02	**88.62**
	5	13.17	29.46	32.81	**56.44**

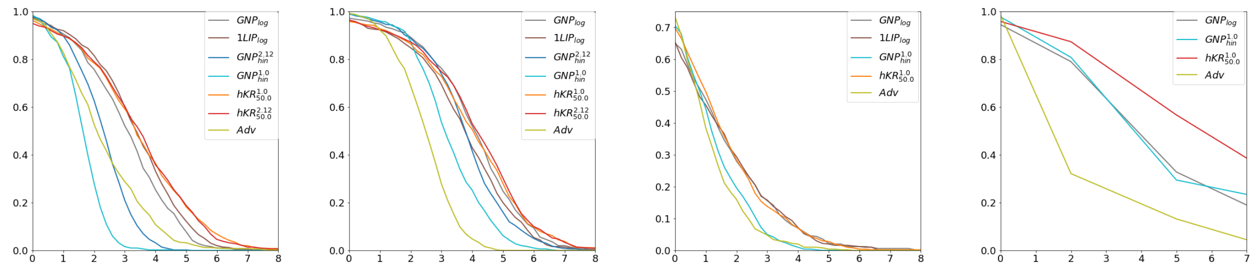
Table 1: Robustness w.r.t. various attacks on CelebA dataset

6. Conclusion and future works

This paper presents a novel classification framework and the associated deep learning process. Besides the interpretation of the classification task as a regularized optimal transport problem, we demonstrate that this new formalization has some valuable properties about error bounds and structural robustness regarding adversarial attacks. We also propose a systematic approach to ensure the 1-Lipschitz constraint of a neural network. This includes a state-of-the-art regularization algorithm and more precise constant evaluation for convolutional and pooling layers. Even if this regularization process can increase the computation time during learning (up to three times slower), it doesn't impact inference. We developed an open source python library based on tensorflow for 1-Lipschitz layers and gradient preserving activation and pooling functions. This makes the approach very easy to implement and to use.

The experiment emphasizes the theoretical results and confirms that the classifier has good and predictable robustness to adversarial attacks with an acceptable cost on accuracy. We also show that our classifier forces adversarial attacks to explicitly modify the input. This suggest that our models can use adversarial attacks for explaining a prediction as it is done with counterfactual explanation in logic [34].

In conclusion, we believe that this classification framework based on optimal transport is of great interest for critical problems since it provides both empirical and theoretical guarantees.



(a) MNIST dense (b) MNIST CNN (c) CIFAR (d) CelebA

Figure 4: Accuracy (Y-axis) w.r.t. of l_2 norm of $FGSM$, l_2PGD, deepfool and l_2 Carlini and Wagner combined attacks on 500 images of the test set

Figure 5: Adversarial examples on CelebA dataset. The first row is the input image.

References

[1] Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301, Long Beach, California, USA, June 2019. PMLR. 2, 3, 5

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, August 2017. PMLR. 2, 5

[3] Åke Björck and C. Bowie. An Iterative Algorithm for Computing the Best Estimate of an Orthogonal Matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, June 1971. 5

[4] Y. Lan Boureau, Jean Ponce, and Yann Lecun. A theoretical analysis of feature pooling in visual recognition. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, ICML 2010 - Proceedings, 27th International Conference on Machine Learning, pages 111–118, September 2010. 5

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1, 6

[7] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval Networks: Improving Robustness to Adversarial Examples. *arXiv:1704.08847 [cs, stat]*, April 2017. 1, 5, 6

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of Machine Learning Research*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1

[9] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[10] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2053–2061. MIT Press, 2015. 2

[11] Gene H. Golub and Henk A. van der Vorst. Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123(1):35–65, 2000. Numerical Analysis 2000. Vol. III: Linear Algebra. 5

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572 [stat.ML]*, December 2014. 1, 4, 6

[14] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. 3

[15] Matthias Hein and Maksym Andriushchenko. Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation. *arXiv:1705.08475 [cs, stat]*, May 2017. 1

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 6

[17] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. 6

[18] Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B. Grosse, and Jörn-Henrik

Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *arXiv:1911.00937*, April 2019. 2, 3, 5, 6

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, 2017. 1, 6

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ArXiv*, abs/1802.05957, 2018. 5

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. *arXiv:1511.04599 [cs]*, November 2015. 6

[23] Hajime Ono, Tsubasa Takahashi, and Kazuya Kakizaki. Lightweight Lipschitz Margin Training for Certified Defense against Adversarial Examples. *arXiv:1811.08080 [cs, stat]*, November 2018. 1

[24] Haifeng Qian and Mark N. Wegman. L2-nonexpansive neural networks. In *International Conference on Learning Representations*, 2019. 2

[25] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 6

[26] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *CoRR*, abs/1602.07868, 2016. 5

[27] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 1

[28] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. May 2018. 1

[29] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, August 2017. 1

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[31] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2018. 1

[32] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. 2, 3

[33] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3835–3844. Curran Associates, Inc., 2018. 5

[34] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. 7