Theoretical evidence for adversarial robustness through randomization

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, Jamal Atif

Theoretical evidence for adversarial robustness through randomization

Rafael Pinot[1,2] **Laurent Meunier**[1,3] **Alexandre Araujo**[1,4]
Hisashi Kashima[5,6] **Florian Yger**[1] **Cédric Gouy-Pailler**[2] **Jamal Atif**[1]

[1]Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France
[2]Institut LIST, CEA, Université Paris-Saclay [3]Facebook AI Research, Paris, France
[4]Wavestone, Paris, France [5]Kyoto University, Kyoto, Japan [6]RIKEN Center for AIP, Japan

Abstract

This paper investigates the theory of robustness against adversarial attacks. It focuses on the family of randomization techniques that consist in injecting noise in the network at inference time. These techniques have proven effective in many contexts, but lack theoretical arguments. We close this gap by presenting a theoretical analysis of these approaches, hence explaining why they perform well in practice. More precisely, we make two new contributions. The first one relates the randomization rate to robustness to adversarial attacks. This result applies for the general family of exponential distributions, and thus extends and unifies the previous approaches. The second contribution consists in devising a new upper bound on the adversarial generalization gap of randomized neural networks. We support our theoretical claims with a set of experiments.

1 Introduction

Adversarial attacks are some of the most puzzling and burning issues in modern machine learning. An adversarial attack refers to a small, imperceptible change of an input maliciously designed to fool the result of a machine learning algorithm. Since the seminal work of [46] exhibiting this intriguing phenomenon in the context of deep learning, a wealth of results have been published on designing attacks [21, 37, 36, 27, 7, 35] and defenses [21, 38, 23, 33, 42, 31]), or on trying to understand the very nature of this phenomenon [18, 43, 16, 17]. Most methods remain unsuccessful to defend against powerful adversaries [7, 32, 2]. Among the defense strategies, randomization has proven effective in some contexts. It consists in injecting random noise (both during training and inference phases) inside the network architecture, *i.e.* at a given layer of the network. Noise can be drawn either from Gaussian [30, 28, 40], Laplace [28], Uniform [49], or Multinomial [13] distributions. Remarkably, most of the considered distributions belong to the Exponential family. Albeit these significant efforts, several theoretical questions remain unanswered. Among these, we tackle the following, for which we provide principled and theoretically-founded answers:

Q1: To what extent does a noise drawn from the Exponential family preserve robustness (in a sense to be defined) to adversarial attacks?

A1: We introduce a definition of robustness to adversarial attacks that is suitable to the randomization defense mechanism. As this mechanism can be described as a non-deterministic querying process, called probabilistic mapping in the sequel, we propose a formal definition of robustness relying on a metric/divergence between probability measures. A key question arises then about the appropriate metric/divergence for our context. This requires tools for comparing divergences w.r.t. the introduced robustness definition. Renyi divergence turned out to be a measure of choice, since it satisfies most of the desired properties (coherence, strength, and computational tractability). Finally, thanks to the existing links between the Renyi divergence and the Exponential family, we were able to prove

Preprint. Under review.

that methods based on noise injection from the Exponential family ensures robustness to adversarial examples (cf Theorem 1) .

Q2: Can we guarantee a good accuracy under attack for classifiers defended with this kind of noise?

A2: We present an upper bound on the drop of accuracy (under attack) of the methods defended with noise drawn from the Exponential family (cf. Theorem 2). Then, we illustrate this result by training different randomized models with Laplace and Gaussian distributions on CIFAR10/CIFAR100. These experiments highlight the trade-off between accuracy and robustness that depends on the amount of noise one injects in the network. Our theoretical and experimental conclusion is that randomized defenses are competitive (with the current state-of-the-art [32]) given the intensity of noise injected in the network.

Outline of the paper: We present in Section 2 the related work on randomized defenses to adversarial examples. Section 3 introduces the definition of robustness relying on a metric/divergence between probability measures, and discusses the key role of the Renyi divergence. We state in Section 4 our main results on the robustness and accuracy of Exponential family-based defenses. Section 5 presents extensive experiments supporting our theoretical findings. Section 6 provides concluding remarks.

2 Related works

Injecting noise into algorithms to improve their robustness has been used for ages in detection and signal processing tasks [51, 8, 34]. It has also been extensively studied in several machine learning and optimization fields, *e.g.* robust optimization [5] and data augmentation techniques [39]. Recently, noise injection techniques have been adopted by the adversarial defense community, especially for neural networks, with very promising results. Randomization techniques are generally oriented towards one of the following objectives: experimental robustness or provable robustness.

Experimental robustness: The first technique explicitly using randomization at inference time as a defense appeared during the 2017 NIPS defense challenge [49]. This method uniformly samples over geometric transformations of the image to select a substitute image to feed the network. Then [13] proposed to use stochastic activation pruning based on a multinomial distribution for adversarial defense. Several papers [30, 40] propose to inject Gaussian noise directly on the activation of selected layers both at training and inference time. While these works hypothesize that noise injection makes the network robust to adversarial perturbations, they do not provide any formal justification on the nature of the noise they use or on the loss of accuracy/robustness of the network.

Provable robustness: In [28], the authors proposed a randomization method by exploiting the link between differential privacy [15] and adversarial robustness. Their framework, called "randomized smoothing" [1], inherits some theoretical results from the differential privacy community allowing them to evaluate the level of accuracy under attack of their method. Initial results from [28] have been refined in [29], and [10]. Our work belongs to this line of research. However, our framework does not treat exactly the same class of defenses. Notably, we provide theoretical arguments supporting the defense strategy based on randomization techniques relying on the exponential family, and derive a new bound on the adversarial generalization gap, which completes the results obtained so far on certified robustness. Furthermore, our focus is on the network randomized by noise injection, "randomized smoothing" instead uses this network to create a *new* classifier robust to attacks.

Since the initial discovery of adversarial examples, a wealth of non randomized defense approaches have also been proposed, inspired by various machine learning domains such as adversarial training [21, 31], image reconstruction [33, 42] or robust learning [21, 31]. Even if these methods have their own merits, a thorough evaluation made by [2] shows that most defenses can be easily broken with known powerful attacks [31, 7, 9]. Adversarial training, which consists in training a model directly on adversarial examples, came out as the best defense in average. Defense based on randomization could be overcome by the Expectation Over Transformation technique proposed by [3] which consists in taking the expectation over the network to craft the perturbation. In this paper, to ensure that our results are not biased by obfuscated gradients, we follow the principles of [2, 6] and evaluate our randomized networks with this technique. We show that randomized defenses are still competitive given the intensity of noise injected in the network.

[1] Name introduced in [10] which came later than [28].

3 General definitions of risk and robustness

3.1 Risk, robustness and probabilistic mappings

Let us consider two spaces $\mathcal{X}$ (with norm $\|.\|_{\mathcal{X}}$), and $\mathcal{Y}$. We consider the classification task that seeks a hypothesis (classifier) $h : \mathcal{X} \to \mathcal{Y}$ minimizing the risk of h w.r.t. some ground-truth distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The risk of h w.r.t $\mathcal{D}$ is defined as

$$\mathrm{Risk}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left(h(x) \neq y\right)\right].$$

Given a classifier $h : \mathcal{X} \to \mathcal{Y}$, and some input $x \in \mathcal{X}$ with true label $y_{true} \in \mathcal{Y}$, to generate an adversarial example, the adversary seeks a τ such that $h(x+\tau) \neq y_{true}$, with some budget α over the perturbation (*i.e* with $\|\tau\|_{\mathcal{X}} \leq \alpha$). α represents the maximum amount of perturbation one can add to x without being spotted (the perturbation remains humanly imperceptible). The overall goal of the adversary is to find a perturbation crafting strategy that both maximizes the risk of h, and keeps the values of $\|\tau\|_{\mathcal{X}}$ small. To measure this risk "under attack" we define the notion of adversarial α-radius risk of h w.r.t. $\mathcal{D}$ as follows

$$\mathrm{Risk}_{\alpha}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\|\tau\|_{\mathcal{X}}\leq\alpha} \mathbb{1}\left(h(x+\tau) \neq y\right)\right] .$$

In practice, the adversary does not have any access to the ground-truth distribution. The literature proposed several surrogate versions of $\mathrm{Risk}_{\alpha}(h)$ (see [14] for more details) to overcome this issue. We focus our analysis on the one used in *e.g* [46], or [16] denoted α-radius prediction-change risk of h w.r.t. $\mathcal{D}_{\mathcal{X}}$ (marginal of $\mathcal{D}$ for $\mathcal{X}$), and defined as

$$\text{PC-Risk}_{\alpha}(h) := \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\exists\tau \in \mathrm{B}(\alpha) \text{ s.t. } h(x+\tau) \neq h(x)\right]$$

where for any $\alpha \geq 0$, $\quad \mathrm{B}(\alpha) := \{\tau \in \mathcal{X} \text{ s.t. } \|\tau\|_{\mathcal{X}} \leq \alpha\}$.

As we will inject some noise in our classifier in order to defend against adversarial attacks, we need to introduce the notion of "probabilistic mapping". Let $\mathcal{Y}$ be the output space, and $\mathcal{F}_{\mathcal{Y}}$ a σ-$algebra$ over $\mathcal{Y}$. Let us also denote $\mathcal{P}(\mathcal{Y})$ the set of probability measures over $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$.

Definition 1 (Probabilistic mapping)**.** *Let $\mathcal{X}$ be an arbitrary space, and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ a measurable space. A* probabilistic mapping *from $\mathcal{X}$ to $\mathcal{Y}$ is a mapping $\mathrm{M} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. To obtain a numerical output out of this* probabilistic mapping*, one needs to sample y according to $\mathrm{M}(x)$.*

This definition does not depend on the nature of $\mathcal{Y}$ as long as $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ is measurable. In that sense, $\mathcal{Y}$ could be either the label space or any intermediate space corresponding to the output of an arbitrary hidden layer of a neural network. Moreover, any mapping can be considered as a probabilistic mapping, whether it explicitly injects noise (as in [28, 40, 13]) or not. In fact, any deterministic mapping can be considered as a probabilistic mapping, since it can be characterized by a Dirac measure. Accordingly, the definition of a probabilistic mapping is fully general and equally treats networks with or without noise injection. There exists no definition of robustness against adversarial attacks that comply with the notion of probabilistic mappings. We settle that by generalizing the notion of prediction-change risk initially introduced in [14] for deterministic classifiers. Let M be a probabilistic mapping from $\mathcal{X}$ to $\mathcal{Y}$, and $d_{\mathcal{P}(\mathcal{Y})}$ some metric/divergence on $\mathcal{P}(\mathcal{Y})$. We define the (α, ϵ)-radius prediction-change risk of M w.r.t. $\mathcal{D}_{\mathcal{X}}$ and $d_{\mathcal{P}(\mathcal{Y})}$ as

$$\text{PC-Risk}_{\alpha}(\mathrm{M}, \epsilon) := \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\exists\tau \in B(\alpha) \text{ s.t. } d_{\mathcal{P}(\mathcal{Y})}(\mathrm{M}(x+\tau), \mathrm{M}(x)) > \epsilon\right] .$$

These three generalized notions allow us to analyze noise injection defense mechanisms (Theorems 1, and 2). We can also define adversarial robustness (and later adversarial gap) thanks to these notions.

Definition 2 (Adversarial robustness)**.** *Let $d_{\mathcal{P}(\mathcal{Y})}$ be a metric/divergence on $\mathcal{P}(\mathcal{Y})$. The probabilistic mapping M is said to be $d_{\mathcal{P}(\mathcal{Y})}$-$(\alpha, \epsilon, \gamma)$ robust if $\text{PC-Risk}_{\alpha}(\mathrm{M}, \epsilon) \leq \gamma$.*

It is difficult in general to show that a classifier is $d_{\mathcal{P}(\mathcal{Y})}$-$(\alpha, \epsilon, \gamma)$ robust. However, we can derive some bounds for particular divergences that will ensure robustness up to a certain level (Theorem 1). It is worth noting that our definition of robustness depends on the considered metric/divergence between probability measures. Lemma 1 gives some insights on the monotony of the robustness according to the parameters, and the probability metric/divergence at hand.

Lemma 1. *Let* M *be a probabilistic mapping, and let* d_1 *and* d_2 *be two metrics on* $\mathcal{P}(\mathcal{Y})$*. If there exists a non decreasing function* $\phi : \mathbb{R} \to \mathbb{R}$ *such that* $\forall \mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$*,* $d_1(\mu_1, \mu_2) \leq \phi(d_2(\mu_1, \mu_2))$*, then the following assertion holds:*

$$\mathrm{M} \textit{ is } d_2\text{-}(\alpha, \epsilon, \gamma)\textit{-robust} \implies \mathrm{M} \textit{ is } d_1\text{-}(\alpha, \phi(\epsilon), \gamma)\textit{-robust}$$

As suggested in Definition 2 and Lemma 1, any given choice of metric/divergence will instantiate a particular notion of adversarial robustness and it should be carefully selected.

3.2 On the choice of the metric/divergence for robustness

The aforementioned formulation naturally raises the question of the choice of the metric used to defend against adversarial attacks. The main notions that govern the selection of an appropriate metric/divergence are *coherence*, *strength*, and *computational tractability*. A metric/divergence is said to be coherent if it naturally fits the task at hand (*e.g.* classification tasks are intrinsically linked to discrete/trivial metrics, conversely to regression tasks). The strength of a metric/divergence refers to its ability to cover (dominate) a wide class of others in the sense of Lemma 1. In the following, we will focus on both the total variation metric and the Renyi divergence, that we consider as respectively the most coherent with the classification task using probabilistic mappings, and the strongest divergence. We first discuss how total variation metric is *coherent* with randomized classifiers but suffers from computational issues. Hopefully, the Renyi divergence provides good guarantees about adversarial robustness, enjoys nice *computational properties*, in particular when considering Exponential family distributions, and is *strong* enough to dominate a wide range of metrics/divergences including total variation.

Let μ_1 and μ_2 be two measures in $\mathcal{P}(\mathcal{Y})$, both dominated by a third measure ν. The trivial distance $d_T(\mu_1, \mu_2) := \mathbb{1}(\mu_1 \neq \mu_2)$ is the simplest distance one can define between μ_1 and μ_2. In the deterministic case, it is straightforward to compute (since the numerical output of the algorithm characterizes its associated measure), but this is not the case in general. In fact one might not have access to the true distribution of the mapping, but just to the numerical outputs. Therefore, one needs to consider more sophisticated metrics/divergences, such as the total variation distance $d_{TV}(\mu_1, \mu_2) := \sup_{Y \in \mathcal{F}_\mathcal{Y}} |\mu_1(Y) - \mu_2(Y)|$. The total variation distance is one of the most broadly used probability metrics. It admits several very simple interpretations, and is a very useful tool in many mathematical fields such as probability theory, Bayesian statistics, coupling or transportation theory. In transportation theory, it can be rewritten as the solution of the Monge-Kantorovich problem with the cost function $c(y_1, y_2) = \mathbb{1}(y_1 \neq y_2)$: $\inf \int_{\mathcal{Y}^2} \mathbb{1}(y_1 \neq y_2)\, d\pi(y_1, y_2)$, where the infimum is taken over all joint probability measures π on $(\mathcal{Y} \times \mathcal{Y}, \mathcal{F}_\mathcal{Y} \otimes \mathcal{F}_\mathcal{Y})$ with marginals μ_1 and μ_2. According to this interpretation, it seems quite natural to consider the total variation distance as a relaxation of the trivial distance on $[0, 1]$ (see [48] for details). In the deterministic case, the total variation and the trivial distance coincides. In general, the total variation allows a finer analysis of the probabilistic mappings than the trivial distance. But it suffers from a high computational complexity. In the following of the paper we will show how to ensure robustness regarding TV distance.

Finally, denoting by g_1 and g_2 the respective probability distributions w.r.t. ν, the Renyi divergence of order λ [41] writes as $d_{R,\lambda}(\mu_1, \mu_2) := \frac{1}{\lambda - 1} \log \int_\mathcal{Y} g_2(y) \left(\frac{g_1(y)}{g_2(y)}\right)^\lambda d\nu(y)$. The Renyi divergence is a generalized measure defined on the interval $(1, \infty)$, where it equals the Kullback-Leibler divergence when $\lambda \to 1$ (that will be denoted d_{KL}), and the maximum divergence when $\lambda \to \infty$. It also has the very special property of being non decreasing w.r.t. λ. This divergence is very common in machine learning, especially in its Kullback-Leibler form as it is widely used as the loss function (cross entropy) of classification algorithms. It enjoys the desired properties since it bounds the TV distance, and is tractable. Furthermore, Proposition 1 proves that Renyi-robustness implies TV-robustness, making it a suitable surrogate for the trivial distance.

Proposition 1 (Renyi-robustness implies TV-robustness)**.** *Let* M *be a probabilistic mapping, then* $\forall \lambda \geq 1$*:*

$$\mathrm{M} \textit{ is } d_{R,\lambda}\text{-}(\alpha, \epsilon, \gamma)\textit{-robust} \implies \mathrm{M} \textit{ is } d_{TV}\text{-}(\alpha, \epsilon', \gamma)\textit{-robust}$$

$$\text{with } \epsilon' = \min\left(\frac{3}{2}\left(\sqrt{1 + \frac{4\epsilon}{9}} - 1\right)^{1/2}, \frac{\exp(\epsilon + 1) - 1}{\exp(\epsilon + 1) + 1}\right) .$$

A crucial property of Renyi-robustness is the *Data processing inequality*. It is a well-known inequality from information theory which states that *"post-processing cannot increase information"* [11, 4]. In our case, if we consider a Renyi-robust probabilistic mapping, composing it with a deterministic mapping maintains Renyi-robustness with the same level.

Proposition 2 (Data processing inequality)**.** *Let us consider a probabilistic mapping* $\mathrm{M} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. *Let us also denote* $\rho : \mathcal{Y} \to \mathcal{Y}'$ *a deterministic function. If* $U \sim \mathrm{M}(x)$ *then the probability measure* $M'(x)$ *s.t* $\rho(U) \sim M'(x)$ *defines a probabilistic mapping* $M' : \mathcal{X} \to \mathcal{P}(\mathcal{Y}')$.

For any $\lambda > 1$ *if* M *is* $d_{R,\lambda}$*-*$(\alpha, \epsilon, \gamma)$ *robust then* M' *is also is* $d_{R,\lambda}$*-*$(\alpha, \epsilon, \gamma)$ *robust.*

Data processing inequality will allow us later to inject some additive noise in any layer of a neural network and to ensure Renyi-robustness.

4 Defense mechanisms based on Exponential family noise injection

4.1 Robustness through Exponential family noise injection

For now, the question of which class of noise to add is treated *ad hoc*. We choose here to investigate one particular class of noise closely linked to the Renyi divergence, namely Exponential family distributions, and demonstrate their interest. Let us first recall what the Exponential family is.

Definition 3 (Exponential family)**.** *Let* Θ *be an open convex set of* $\mathbb{R}^n$*, and* $\theta \in \Theta$*. Let* ν *be a measure dominated by* μ *(either by the Lebesgue or counting measure), it is said to be part of the* Exponential family *of parameter* θ *(denoted* $E_F(\theta, t, k)$*) if it has the following probability density function*

$$p_F(z, \theta) = \exp\left\{\langle t(z), \theta\rangle - u(\theta) + k(z)\right\}$$

where $t(z)$ *is a sufficient statistic,* k *a carrier measure (either for a Lebesgue or a counting measure) and* $u(\theta) = \log \int_z \exp\left\{< t(z), \theta > + k(z)\right\} dz$.

To show the robustness of randomized networks with noise injected from the Exponential family, one needs to define the notion of sensitivity for a given deterministic function:

Definition 4 (Sensitivity of a function)**.** *For any* $\alpha \geq 0$ *and for any* $||.||_A$ *and* $||.||_B$ *two norms, the* α*-sensitivity of* f *w.r.t.* $||.||_A$ *and* $||.||_B$ *is defined as*

$$\Delta_\alpha^{A,B}(f) := \sup_{x,y \in \mathcal{X}, ||x-y||_A \leq \alpha} ||f(x) - f(y)||_B \ .$$

Let us consider an n-layer feedforward neural network $\mathcal{N}(.) = \phi^n \circ ... \circ \phi^1(.)$. For any $i \in [n]$, we define $\mathcal{N}_{|i}(.) = \phi^i \circ ... \circ \phi^1(.)$ the neural network truncated at layer i. Theorem 1 shows that, injecting noise drawn from an Exponential family distribution ensures robustness to adversarial example attacks in the sense of Definition 2.

Theorem 1 (Exponential family ensures robustness)**.** *Let us denote* $\mathcal{N}_X^i(.) = \phi^n \circ ... \circ \phi^{i+1}(\mathcal{N}_{|i}(.) + X)$ *with* X *a random variable. Let us also consider two arbitrary norms* $||.||_A$ *and* $||.||_B$ *respectively on* $\mathcal{X}$ *and on the output space of* $\mathcal{N}_X^i$.

- *If* $X \sim E_F(\theta, t, k)$ *where* t *and* k *have non-decreasing modulus of continuity* ω_t *and* ω_k*. Then for any* $\alpha \geq 0$*,* $\mathcal{N}_X^i(.)$ *defines a probabilistic mapping that is* $d_{R,\lambda}$*-*(α, ϵ) *robust with* $\epsilon = ||\theta||_2 \omega_t^{B,2}(\Delta_\alpha^{A,B}(\phi)) + \omega_k^{B,1}(\Delta_\alpha^{A,B}(\phi))$ *where* $||.||_2$ *is the norm corresponding to the scalar product in the definition of the exponential family density function and* $||.||_1$ *is the absolute value on* $\mathbb{R}$*. The notion of continuity modulus is defined in the supplementary material.*
- *If* X *is a centered Gaussian random variable with a non degenerated matrix parameter* Σ*. Then for any* $\alpha \geq 0$*,* $\mathcal{N}_X^i(.)$ *defines a probabilistic mapping that is* $d_{R,\lambda}$*-*(α, ϵ) *robust with* $\epsilon = \frac{\lambda \Delta_\alpha^{A,2}(\phi)^2}{2\sigma_{min}(\Sigma)}$ *where* $||.||_2$ *is the canonical Euclidean norm on* $\mathbb{R}^n$.

In simpler words, the previous theorem ensures stability in the neural network when injecting noise w.r.t. the distribution of the output. Intuitively, if two inputs are close w.r.t. $\|.\|_A$, the output distributions of the network will be close in the sense of Renyi divergence. It is well known that

in the case of deterministic neural networks, the Lipschitz constant becomes bigger as the number of layers increases [22]. By injecting noise at layer i, the notion of robustness only depends on the sensitivity of the first i layers of the network and not the following ones. In that sense, randomization provides a more precise control on the "continuity" of the neural network. In the next section, we show that thanks to the notion of robustness w.r.t. probabilistic mappings, one can bound the loss of accuracy of a randomized neural network when it is attacked.

4.2 Bound on the generalization gap under attack

The notions of risk and adversarial risk can easily be generalized to encompass probabilistic mappings.

Definition 5 (Risks for probabilistic mappings)**.** *Let* M *be a probabilistic mapping from* $\mathcal{X}$ *to* $\mathcal{Y}$*, the risk and the* α*-radius adversarial risk of* M *w.r.t.* $\mathcal{D}$ *are defined as*

$$\mathrm{Risk}(\mathrm{M}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{E}_{y'\sim\mathrm{M}(x)}\left[\mathbb{1}\left(y' \neq y\right)\right]\right]$$

$$\mathrm{Risk}_{\alpha}(\mathrm{M}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\|\tau\|_{\mathcal{X}}\leq\alpha} \mathbb{E}_{y'\sim\mathrm{M}(x+\tau)}\left[\mathbb{1}\left(y' \neq y\right)\right]\right] \quad .$$

The definition of adversarial risk for a probabilistic mapping can be matched with the concept of Expectation over Transformation (EoT) attacks [2]. Indeed, EoT attacks aim at computing the best opponent in expectation for a given random transformation. In the adversarial risk definition, the adversary chooses the perturbation which has the greatest probability to fool the model, which is a stronger objective than the EoT objective. Theorem 2 provides a bound on the gap between the adversarial risk and the regular risk:

Theorem 2 (Adversarial generalization gap bound in the randomized setting)**.** *Let* M *be the probabilistic mapping at hand. Let us suppose that* M *is* $d_{R,\lambda}$*-*(α, ϵ) *robust for some* $\lambda \geq 1$ *then:*

$$|\,\mathrm{Risk}_{\alpha}(\mathrm{M}) - \mathrm{Risk}(\mathrm{M})| \leq 1 - e^{-\epsilon}\mathbb{E}_x\left[e^{-H(\mathrm{M}(x))}\right]$$

where H *is the Shannon entropy* $H(p) = -\sum_i p_i \log(p_i)$.

This theorem gives a control on the loss of accuracy under attack w.r.t. the robustness parameter ϵ and the entropy of the predictor. It provides a tradeoff between the quantity of noise added in the network and the accuracy under attack. Intuitively, when the noise increases, for any input, the output distribution tends towards the uniform distribution, then, $\epsilon \to 0$ and $H(\mathrm{M}(x)) \to \log(K)$, and the risk and the adversarial risk both tends to $\frac{1}{K}$ where K is the number of classes in the classification problem. On the opposite, if no noise is injected, for any input, the output distribution is a Dirac distribution, then, if the prediction for the adversarial example is not the same as for the regular one, $\epsilon \to \infty$ and $H(\mathrm{M}(x)) \to 0$. Hence, the noise needs to be designed both to preserve accuracy and robustness to adversarial attacks. In the Section 5, we give an illustration of this bound when M is a neural network with noise injection at input level as presented in Theorem 1.

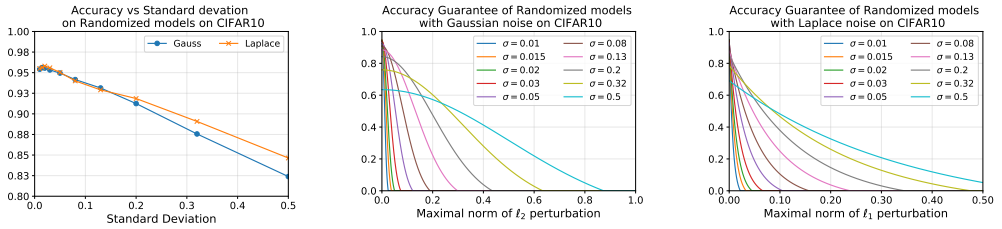
5 Numerical experiments

To illustrate our theoretical findings, we train randomized neural networks with a simple method which consists in injecting a noise drawn from an Exponential family distribution in the image during training and inference. This section aims to answer **Q2** stated in the introduction, by tackling the following sub-questions:

- **Q2.1:** How does the randomization impact the accuracy of the network? And, how does the theoretical trade-off between accuracy and robustness apply in practice?
- **Q2.2:** What is the accuracy under attack of randomized neural networks against powerful iterative attacks? And how does randomized neural networks compare to state-of-the-art defenses given the intensity of the injected noise?

5.1 Experimental setup

We present our results and analysis on CIFAR-10, CIFAR-100 [26] and ImageNet datasets [12]. For CIFAR-10 and CIFAR-100 [26], we used a Wide ResNet architecture [50] which is a variant of the



(a) (b) (c)

Figure 1: (a) Impact of the standard deviation of the injected noise on accuracy in a randomized model on CIFAR-10 dataset with a Wide ResNet architecture. (b) and (c) illustration of the guaranteed accuracy of different randomized models with Gaussian (b) and Laplace (c) noises given the norm of the adversarial perturbation.

ResNet model from [24]. We use 28 layers with a widen factor of 10. We train all networks for 200 epochs, a batch size of 400, dropout 0.3 and Leaky Relu activation with a slope on $\mathbb{R}^-$ of 0.1. We minimize the Cross Entropy Loss with Momentum 0.9 and use a piecewise constant learning rate of 0.1, 0.02, 0.004 and 0.00008 after respectively 7500, 15000 and 20000 steps. The networks achieve for CIFAR10 and 100 a TOP-1 accuracy of 95.8% and 79.1% respectively on test images. For ImageNet [12], we use an Inception ResNet v2 [45] which is the sate of the art architecture for this dataset and achieve a TOP-1 accuracy of 80%. For the training of ImageNet, we use the same hyper parameters setting as the original implementation. We train the network for 120 epochs with a batch size of 256, dropout 0.8 and Relu as activation function. All evaluations were done with a single crop on the non-blacklisted subset of the validation set.

To transform these classical networks to probabilistic mappings, we inject noise drawn from Laplace and Gaussian distributions, each with various standard deviations. While the noise could theoretically be injected anywhere in the network, we inject the noise on the image for simplicity. More experiments with noise injected in the first layer of the network are presented in the supplementary material. To evaluate our models under attack, we use three powerful iterative attacks with different norms: *ElasticNet* attack (EAD) [9] with ℓ_1 distortion, *Carlini&Wagner* attack (C&W) [7] with ℓ_2 distortion and *Projected Gradient Descent* attack (PGD) [31] with ℓ_∞ distortion. All standard deviations and attack intensities are in between -1 and 1. Precise descriptions of our numerical experiments and of the attacks used for evaluation are deferred to the supplementary material.

Attacks against randomized defenses: It has been pointed out by [3, 6] that in a white box setting, an attacker with a complete knowledge of the system will know the distribution of the noise injected in the network. As such, to create a stronger adversarial example, the attacker can take the expectation of the loss or the logits of the randomized network during the computation of the attack. This technique is called Expectation Over Transformation (EoT) and we use a Monte Carlo method with 80 simulations to approximate the best perturbation for a randomized network.

5.2 Experimental results

Trade-off between accuracy and intensity of noise (Q2.1): When injecting noise as a defense mechanism, regardless of the distribution it is drawn from, we observe (as in Figure 1(a)) that the accuracy decreases when the noise intensity grows. In that sense, noise needs to be calibrated to preserve both accuracy and robustness against adversarial attacks, i.e. it needs to be large enough to preserve robustness and small enough to preserve accuracy. Figure 1(a) shows the loss of accuracy on CIFAR10 from 0.95 to 0.82 (respectively 0.95 to 0.84) with noise drawn from a Gaussian distribution (respectively Laplace) with a standard deviation from 0.01 to 0.5. Figure 1(b) and 1(c) illustrate the theoretical lower bound on accuracy under attack of Theorem 2 for different distributions and standard deviations. The term in entropy of Theorem 2 has been estimated using a Monte Carlo method with 10^4 simulations. The trade-off between accuracy and robustness from Theorem 2 thus appears w.r.t the noise intensity. With small noises, the accuracy is high, but the guaranteed accuracy drops fast w.r.t the magnitude of the adversarial perturbation. Conversely, with bigger noises, the accuracy is lower but decreases slowly w.r.t the magnitude of the adversarial perturbation. These Figures also show that Theorem 2 gives strong accuracy guarantees against small adversarial perturbations. Next

Table 1: Accuracy under attack on the CIFAR-10 dataset with a randomized Wide ResNet architecture. We compare the accuracy on natural images and under attack with different noise over 3 iterative attacks (the number of steps is next to the name) made with 80 Monte Carlo simulations to compute EoT attacks. The first line is the baseline, no noise has been injected.

Distribution	**Sd**	**Natural**	ℓ_1 **– EAD 60**	ℓ_2 **– C&W 60**	ℓ_∞ **– PGD 20**
-	-	0.958	0.035	0.034	0.384
Normal	0.01	0.954	0.193	0.294	0.408
	0.50	0.824	0.448	0.523	0.587
Laplace	0.01	0.955	0.208	0.313	0.389
	0.50	0.846	0.464	0.494	0.589

Table 2: Accuracy under attack of randomized neural network with different distributions and standard deviations versus adversarial training by Madry et al. [31]. The PGD attack has been made with 20 step, an epsilon of 0.06 and a step size of 0.006 (input space between -1 and $+1$). The Carlini&Wagner attack uses 30 steps, 9 binary search steps and a 0.01 learning rate. The first line refers to the baseline without attack.

Attack	**Steps**	**Madry et al.** [31]	**Normal 0.32**	**Laplace 0.32**	**Normal 0.5**	**Laplace 0.5**
-	-	0.873	0.876	0.891	0.824	0.846
ℓ_∞ – PGD	20	0.456	0.566	0.576	0.587	0.589
ℓ_2 – C&W	30	0.468	0.512	0.502	0.489	0.479

paragraph shows that in practice, randomized networks achieve much higher accuracy under attack than the theoretical bound, and keep this accuracy against much larger perturbations.

Performance of randomized networks under attacks and comparison to state of the art (Q2.2): While Figure 1(b) and 1(c) illustrated a theoretical robustness against growing adversarial perturbations, Table 1 illustrates this trade-off experimentally. It compares the accuracy obtained under attack by a deterministic network with the one obtained by randomized networks with Gaussian and Laplace noises both with low (0.01) and high (0.5) standard deviations. Randomized networks with a small noise lead to no loss in accuracy with a small robustness while high noise leads to a higher robustness at the expense of loss of accuracy (~ 11 points).

Finally, Table 2 compares the accuracy and the accuracy under attack of randomized networks with Gaussian and Laplace distributions for different standard deviations against adversarial training from Madry et al. [31]. We observe that the accuracy on natural images of both noise injection methods are similar to the one from [31]. Moreover, both methods are more robust than adversarial training to PGD and C&W attacks. As with all the experiments, to construct an EoT attack, we use 80 Monte Carlo simulations at every step of PGD and C&W attacks. These experiments show that randomized defenses can be competitive given the intensity of noise injected in the network. Note that these experiments have been led with EoT of size 80. For much bigger sizes of EoT these results would be mitigated. Nevertheless, the accuracy would never drop under the bounds illustrated in Figure 5.2, since Theorem 2 gives a bound that on the worst case attack strategy (including EoT).

6 Conclusion and future works

This paper brings new contributions to the field of provable defenses to adversarial attacks. Principled answers have been provided to key questions on the interest of randomization techniques, and on their loss of accuracy under attack. The obtained bounds have been illustrated in practice by conducting thorough experiments on baseline datasets such as CIFAR and ImageNet. We show in particular that a simple method based on injecting noise drawn from the Exponential family is competitive compared to baseline approaches while leading to provable guarantees. Future work will focus on investigating other noise distributions belonging or not to the Exponential family, combining randomization with more sophisticated defenses and on devising new tight bounds on the adversarial generalization gap.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[4] N. J. Beaudry and R. Renner. An intuitive proof of the data processing inequality. *Quantum Info. Comput.*, 12(5-6):432–441, May 2012.

[5] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[6] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, and A. Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[8] F. Chapeau-Blondeau and D. Rousseau. Noise-enhanced performance for an optimal bayesian estimator. *IEEE Transactions on Signal Processing*, 52(5):1327–1334, 2004.

[9] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[10] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019.

[11] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[13] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

[14] D. Diochnos, S. Mahloujifar, and M. Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, pages 10380–10389, 2018.

[15] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[16] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1186–1195. Curran Associates, Inc., 2018.

[17] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

[18] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto. Empirical study of the topology and geometry of deep networks. In *IEEE CVPR*, 2018.

[19] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

[20] G. L. Gilardoni. On pinsker's and vajda's type inequalities for csiszár's f-divergences. *IEEE Transactions on Information Theory*, 56(11):5377–5386, Nov 2010.

[21] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[22] H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

[23] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] P. J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[27] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[28] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 727–743, 2018.

[29] B. Li, C. Chen, W. Wang, and L. Carin. Second-order adversarial attack and certifiable robustness. *CoRR*, abs/1809.03113, 2018.

[30] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. Towards robust neural networks via random self-ensemble. In *European Conference on Computer Vision*, pages 381–397. Springer, 2018.

[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[33] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147. ACM, 2017.

[34] S. Mitaim and B. Kosko. Adaptive stochastic resonance. *Proceedings of the IEEE*, 86(11):2152–2183, 1998.

[35] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94. Ieee, 2017.

[36] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[37] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[38] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[39] L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[40] A. S. Rakin, Z. He, and D. Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv preprint arXiv:1811.09310*, 2018.

[41] A. Rényi. On measures of entropy and information. Technical report, HUNGARIAN ACADEMY OF SCIENCES Budapest Hungary, 1961.

[42] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[43] C.-J. Simon-Gabriel, Y. Ollivier, B. Schölkopf, L. Bottou, and D. Lopez-Paz. Adversarial vulnerability of neural networks increases with input dimension. *arXiv preprint arXiv:1802.01421*, 2018.

[44] M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[46] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[47] I. Vajda. Note on discrimination information and variation. *IEEE Trans. Inform. Theory*, 16(6):771–773, Nov. 1970.

[48] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[49] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[50] S. Zagoruyko and N. Komodakis. Wide residual networks. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[51] S. Zozor and P.-O. Amblard. Stochastic resonance in discrete time nonlinear AR(1) models. *IEEE transactions on Signal Processing*, 47(1):108–122, 1999.

Supplementary Material

Table of Contents

A Notations and definitions

Let us consider an output space $\mathcal{Y}$, and $\mathcal{F}_\mathcal{Y}$ a $\sigma\text{-}algebra$ over $\mathcal{Y}$. We denote $\mathcal{P}(\mathcal{Y})$ the set of probability measures over $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$. Let $(\mathcal{Y}', \mathcal{F}_{\mathcal{Y}'})$ be a second measurable space, and ϕ a measurable mapping from $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ to $(\mathcal{Y}', \mathcal{F}_{\mathcal{Y}'})$. Finally Let us consider μ, ν two measures on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$.

Dominated measure: μ is said to be dominated by ν (denoted $\mu \ll \nu$) if for all $Y \in \mathcal{F}_\mathcal{Y},\ \nu(Y) = 0 \implies \mu(Y) = 0$. If μ is dominated by ν, there is a measurable function $h : \mathcal{Y} \to [0, +\infty)$ such that for all $Y \in \mathcal{F}_\mathcal{Y}$, $\mu(Y) = \int_Y h\, d\nu$. h is called the Radon-Nikodym derivative of μ w.r.t. ν and is denoted $\frac{d\mu}{d\nu}$.

Push-forward measure: the push-forward measure of ν by ϕ (denoted $\phi\#\nu$) is the measure on $(\mathcal{Y}', \mathcal{F}_{\mathcal{Y}'})$ such that $\forall Z \in \mathcal{F}_{\mathcal{Y}'},\ \phi\#\nu(Z) = \nu(\phi^{-1}(Z))$.

Convolution product: the convolution of ν with μ, denoted $\nu * \mu$ is the push-forward measure of $\nu \otimes \mu$ by the addition on $\mathcal{Y}$. Since the convolution between functions is defined accordingly, we use $*$ indifferently for measures and simple functions.

Modulus of continuity: Let us consider $f : (E, \|.\|_E) \to (F, \|.\|_F)$. f admits a non-decreasing modulus of continuity regarding $\|.\|_E$ and $\|.\|_F$ if there exists a non-decreasing function $\omega_f^{E,F} : \mathbb{R}^+ \to \mathbb{R}^+$ such as for all $x, y \in E$, $||f(y) - f(x)||_F \leq \omega_f^{E,F}(||x - y||_E)$.

B Main proofs

As a trade-off between completeness and conciseness, we delayed the proofs of our theorems to this section.

B.1 Proof of Lemma 1

Lemma 1. *Let* M *be a probabilistic mapping, and let* d_1 *and* d_2 *be two metrics on* $\mathcal{P}(\mathcal{Y})$*. If there exists a non decreasing function* $\phi : \mathbb{R} \to \mathbb{R}$ *such that* $\forall \mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$*,* $d_1(\mu_1, \mu_2) \leq \phi(d_2(\mu_1, \mu_2))$*, then the following holds:*

$$\mathrm{M} \textit{ is } d_2\textit{-}(\alpha, \epsilon, \gamma)\textit{-robust} \implies \mathrm{M} \textit{ is } d_1\textit{-}(\alpha, \phi(\epsilon), \gamma)\textit{-robust}$$

Proof. Let us consider a probabilistic mapping M, $x \sim \mathcal{D}$, and $\tau \in B(\alpha)$, one has $d_1(\mathrm{M}(x), \mathrm{M}(x+\tau)) \leq \phi(d_2(\mathrm{M}(x), \mathrm{M}(x+\tau))) \leq \phi(\epsilon)$. Hence $\mathbb{P}_{x\sim\mathcal{D}}\left[\forall \tau \in B(\alpha),\ d_1(\mathrm{M}(x+\tau), \mathrm{M}(x)) \leq \phi(\epsilon)\right] \leq 1 - \gamma$. By inverting the inequality, one gets the expected result. □

B.2 Proof of Proposition 1

Proposition 1 (Renyi-robustness implies TV-robustness)**.** *Let* M *be a probabilistic mapping, then for all* $\lambda \geq 1$*:*

$$\mathrm{M} \textit{ is } d_{R,\lambda}\textit{-}(\alpha, \epsilon, \gamma)\textit{-robust} \implies \mathrm{M} \textit{ is } d_{TV}\textit{-}(\alpha, \epsilon', \gamma)\textit{-robust}$$

$$\text{with } \epsilon' = \min\left(\frac{3}{2}\left(\sqrt{1 + \frac{4\epsilon}{9}} - 1\right)^{1/2}, \frac{\exp(\epsilon+1) - 1}{\exp(\epsilon+1) + 1}\right).$$

Proof. Given two probability measures μ_1 and μ_2 on $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, and $\lambda > 0$ one wants to find a bound on $d_{TV}(\mu_1, \mu_2)$ as a functional of $d_{R,\lambda}(\mu_1, \mu_2)$.

Thanks to [20], one has

$$d_{KL}(\mu_1, \mu_2) \geq 2 d_{TV}(\mu_1, \mu_2)^2 + \frac{4 d_{TV}(\mu_1, \mu_2)^4}{9}.$$

From which it follows that

$$d_{TV}(\mu_1, \mu_2)^2 \leq \frac{9}{4}\left(\sqrt{1 + \frac{4 d_{KL}(\mu_1, \mu_2)}{9}} - 1\right)$$

One thus finally gets:

$$d_{TV}(\mu_1, \mu_2) \leq \frac{3}{2}\left(\sqrt{1 + \frac{4 d_{KL}(\mu_1, \mu_2)}{9}} - 1\right)^{1/2}$$

Moreover, using inequality from [47], one gets:

$$d_{KL}(\mu_1, \mu_2) \geq \log\left(\frac{1 + d_{TV}(\mu_1, \mu_2)}{1 - d_{TV}(\mu_1, \mu_2)}\right) - \frac{2 d_{TV}(\mu_1, \mu_2)}{1 + d_{TV}(\mu_1, \mu_2)}$$

For the sake of simplicity, since the second part of the right hand side of the equation is non increasing given $d_{TV}(\mu_1, \mu_2)$, and since $0 \leq d_{TV}(\mu_1, \mu_2) \leq 1$ one gets:

$$d_{KL}(\mu_1, \mu_2) + 1 \geq \log\left(\frac{1 + d_{TV}(\mu_1, \mu_2)}{1 - d_{TV}(\mu_1, \mu_2)}\right)$$

Hence, one gets:

$$\frac{\exp(d_{KL}(\mu_1, \mu_2) + 1) - 1}{\exp(d_{KL}(\mu_1, \mu_2) + 1) + 1} \geq d_{TV}(\mu_1, \mu_2)$$

By combining the two results, one obtains:

$$d_{TV}(\mu_1, \mu_2) \leq \min\left(\frac{3}{2}\left(\sqrt{1 + \frac{4 d_{KL}(\mu_1, \mu_2)}{9}} - 1\right)^{1/2}, \frac{\exp(d_{KL}(\mu_1, \mu_2) + 1) - 1}{\exp(d_{KL}(\mu_1, \mu_2) + 1) + 1}\right).$$

To conclude for $\lambda > 1$ it suffices to use Lemma 1, and the monotony of Renyi divergence regarding λ. □

B.3 Proof of Proposition 2

Proposition 2 (Data processing inequality)**.** *Let us consider a probabilistic mapping* $\mathrm{M} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$*. Let us also denote* $\rho : \mathcal{Y} \to \mathcal{Y}'$ *a deterministic function. If* $U \sim \mathrm{M}(x)$ *then the probability measure* $M'(x)$ *s.t.* $\rho(U) \sim M'(x)$ *defines a probabilistic mapping* $M' : \mathcal{X} \to \mathcal{P}(\mathcal{Y}')$*.*

For any $\lambda > 1$ *if* M *is* $d_{R,\lambda}$*-*$(\alpha, \epsilon, \gamma)$ *robust then* M' *is also is* $d_{R,\lambda}$*-*$(\alpha, \epsilon, \gamma)$ *robust.*

Proof. Let us consider M a $d_{R,\lambda}$-$(\alpha, \epsilon, \gamma)$ robust algorithm. Let us also take $x \in \mathcal{X}$, and $\tau \in B(\alpha)$. Without loss of generality, we consider that $\mathrm{M}(x)$, and $\mathrm{M}(x+\tau)$ are dominated by the same measure μ. Finally let us take ρ a measurable mapping from $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ to $(\mathcal{Y}', \mathcal{F}_{\mathcal{Y}'})$. For the sake of readability we denote $\mathrm{M}(x) = \nu_1$ and $M(x+\tau) = \nu_2$ (therefore $M'(x) = \rho\#\nu_1$, and $M'(x+\tau) = \rho\#\nu_2$).

Since $\mu >> \nu_1, \nu_2$, one has $\rho\#\mu >> \rho\#\nu_1, \rho\#\nu_2$. Hence one has

$$d_{R,\lambda}(\rho\#\nu_1, \rho\#\nu_2) = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}} \left(\frac{d\rho\#\nu_1}{d\rho\#\mu}\right)^{\lambda} \left(\frac{d\rho\#\nu_2}{d\rho\#\mu}\right)^{1-\lambda} d\rho\#\mu = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}} \left(\frac{d\rho\#\nu_1}{d\rho\#\nu_2}\right)^{\lambda} d\rho\#\nu_2$$

Simply using the transfer theorem, one gets

$$d_{R,\lambda}(\rho\#\nu_1, \rho\#\nu_2) = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}'} \left(\frac{d\rho\#\nu_1}{d\rho\#\nu_2} \circ \rho\right)^{\lambda} d\nu_2$$

Since $\left(\frac{d\rho\#\nu_1}{d\rho\#\nu_2} \circ \rho\right) = \mathbb{E}\left(\frac{d\nu_1}{d\nu_2} \middle| \rho^{-1}(\mathcal{F}_{\mathcal{Y}})\right)$ one easily gets the following:

$$d_{R,\lambda}(\rho\#\nu_1, \rho\#\nu_2) = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}'} \left(\frac{d\rho\#\nu_1}{d\rho\#\nu_2} \circ \rho\right)^{\lambda} d\nu_2 = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}'} \mathbb{E}\left(\frac{d\nu_1}{d\nu_2} \middle| \rho^{-1}(\mathcal{F}_{\mathcal{Y}})\right)^{\lambda} d\nu_2$$

Finally, by using the Jensen inequality, and the property of the conditional expectation, one has

$$d_{R,\lambda}(\rho\#\nu_1, \rho\#\nu_2) \leq \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}'} \mathbb{E}\left(\frac{d\nu_1}{d\nu_2}^{\lambda} \middle| \rho^{-1}(\mathcal{F}_{\mathcal{Y}})\right) d\nu_2 = \frac{1}{\lambda - 1} \log \int_{\mathcal{Y}'} \frac{d\nu_1}{d\nu_2}^{\lambda} d\nu_2 = d_{R,\lambda}(\nu_1, \nu_2).$$

□

B.4 Proof of Theorem 1

Lemma 2. *Let* $\psi : \mathcal{X} \to \mathbb{R}^n$ *be a mapping. For any* $\alpha \geq 0$ *and for any norms* $||.||_A$ *and* $||.||_B$*, one can define* $\Delta_{\alpha}^{A,B}(\psi) := \sup_{x,y \in \mathcal{X}, ||x-y||_A \leq \alpha} ||\psi(x) - \psi(y)||_B$*. Let* X *be a random variable. We denote* $\mathrm{M}(x)$ *the probability measure of the random variable* $\psi(x) + X$*.*

- *If* $X \sim E_F(\theta, t, k)$ *where* t *and* k *have non-decreasing modulus of continuity* ω_t *and* ω_k*.*
  Then for any $\alpha \geq 0$*,* M *defines a probabilistic mapping that is* $d_{R,\lambda}$*-*(α, ϵ) *robust with* $\epsilon = ||\theta||_2 \omega_t^{B,2}(\Delta_{\alpha}^{A,B}(\phi)) + \omega_k^{B,1}(\Delta_{\alpha}^{A,B}(\phi))$ *where* $||.||_2$ *is the norm corresponding to the scalar product in the definition of the exponential family density function and* $||.||_1$ *is here the absolute value on* $\mathbb{R}$*.*
- *If* X *is a centered Gaussian random variable with a non degenerated matrix parameter* Σ*. Then for any* $\alpha \geq 0$*,* M *defines a probabilistic mapping that is* $d_{R,\lambda}$*-*(α, ϵ) *robust with* $\epsilon = \frac{\lambda \Delta_{\alpha}^{A,2}(\phi)^2}{2\sigma_{min}(\Sigma)}$ *where* $||.||_2$ *is the canonical Euclidean norm on* $\mathbb{R}^n$*.*

Proof. Let us consider M the probabilistic mapping constructed from noise injections respectively drawn from 1) an exponential family with non-decreasing modulus of continuity, or 2) a non degenerate Gaussian. Let us take $x \in \mathcal{X}$, and $\tau \in B(\alpha)$. Without loss of generality, we consider that $\mathrm{M}(x)$, and $\mathrm{M}(x+\tau)$ are dominated by the same measure μ. Let us also denote, p_F the Radon-Nikodym derivative of the noise drawn in 1) with respect to μ, p_G the Radon-Nikodym derivative of the noise drawn or in 2) with respect to μ and δ_a the Dirac function mapping any element to 1 if it equals a and to 0 otherwise.

1)

$$
\begin{aligned}
d_{R,\lambda}\left(\mathrm{M}(x),\mathrm{M}(x+\tau)\right) &= d_{R,\lambda}\left(\nu * \delta_{\psi(x)}, \nu * \delta_{\psi(x+\tau)}\right) \\
&\leq d_{R,\infty}\left(\nu * \delta_{\psi(x)}, \nu * \delta_{\psi(x+\tau)}\right) \\
&= \log \sup_{z\in\mathbb{R}^n} \frac{(p_F * \delta_{\psi(x)})(z)}{(p_F * \delta_{\psi(x+\tau)})(z)} \\
&= \log \sup_{z\in\mathbb{R}^n} \exp(< t(z-\psi(x)) - t(z-\psi(x+\tau)), \theta > \\
&+ k(z-\psi(x)) - k(z-\psi(x+\tau))) \\
&\leq \sup_{z\in\mathbb{R}^n} ||\theta||_2 ||t(z-\psi(x)) - t(z-\psi(x+\tau))||_2 + |k(z-\psi(x)) - k(z-\psi(x+\tau))| \\
&\leq ||\theta||_2 \omega_t^{B,2}(||\psi(x+\tau)-\psi(x)||_B) + \omega_k^{B,1}(||\psi(x+\tau)-\psi(x)||_B) \\
&\leq ||\theta||_2 \omega_t^{B,2}(\Delta_\alpha^{A,B}(\psi)) + \omega_k^{B,1}(\Delta_\alpha^{A,B}(\psi))
\end{aligned}
$$

2)

$$
d_{R,\lambda}\left(\mathrm{M}(x),\mathrm{M}(x+\tau)\right) = \frac{1}{\lambda-1}\log\int_{\mathbb{R}^n} \left(p_G * \delta_{\psi(x)}\right)^{\lambda} \times \left(p_G * \delta_{\psi(x+\tau)}\right)^{1-\lambda} d\mu
$$

$$
= \frac{1}{\lambda-1}\log\int_{\mathbb{R}^n} \frac{\exp\left\{-1/2\left(\lambda\left(z-\psi(x)\right)^\top \Sigma^{-1}\left(z-\psi(x)\right) + (1-\lambda)\left(z-\psi(x+\tau)\right)^\top \Sigma^{-1}\left(z-\psi(x+\tau)\right)\right)\right\}}{(2\pi)^{n/2}|\Sigma|^{1/2}} dz
$$

$$
= \frac{-\left(\lambda\psi(x)^\top\Sigma^{-1}\psi(x) + (1-\lambda)\psi(x+\tau)^\top\Sigma^{-1}\psi(x+\tau) - \left(\lambda\psi(x)+(1-\lambda)\psi(x+\tau)\right)^\top \Sigma^{-1}\left(\lambda\psi(x)+(1-\lambda)\psi(x+\tau)\right)\right)}{2\lambda-2}
$$

$$
= \frac{\lambda^2-\lambda}{2(\lambda-1)}(\psi(x)-\psi(x+\tau))^\top\Sigma^{-1}(\psi(x)-\psi(x+\tau))
$$

$$
\leq \frac{\lambda}{2}\sigma_{max}\left(\Sigma^{-1}\right)||(\psi(x)-\psi(x+\tau))||_2^2 \leq \frac{\lambda\Delta_\alpha^{A,2}(\psi)^2}{2\sigma_{min}(\Sigma)}.
$$

□

Theorem 1 (Exponential family ensures robustness)**.** *Let us denote $\mathcal{N}_X^i(.) = \phi^n \circ ... \circ \phi^{i+1}(\mathcal{N}_{|i}(.) + X)$ with X a random variable. Let us also consider $||.||_A$, and $||.||_B$ two arbitrary norms respectively on $\mathcal{X}$ and on the output space of $\mathcal{N}_X^i$.*

- *If $X \sim E_F(\theta, t, k)$ where t and k have non-decreasing modulus of continuity ω_t and ω_k. Then for any $\alpha \geq 0$, $\mathcal{N}_X^i(.)$ defines a probabilistic mapping that is $d_{R,\lambda}$-(α,ϵ) robust with $\epsilon = ||\theta||_2\omega_t^{B,2}(\Delta_\alpha^{A,B}(\phi)) + \omega_k^{B,1}(\Delta_\alpha^{A,B}(\phi))$ where $||.||_2$ is the norm corresponding to the scalar product in the definition of the exponential family density function and $||.||_1$ is here the absolute value on $\mathbb{R}$. The notion of continuity modulus is defined in the preamble of this supplementary material.*
- *If X is a centered Gaussian[2] random variable with a non degenerated matrix parameter Σ. Then for any $\alpha \geq 0$, $\mathcal{N}_X^i(.)$ defines a probabilistic mapping that is $d_{R,\lambda}$-(α,ϵ) robust with $\epsilon = \frac{\lambda\Delta_\alpha^{A,2}(\phi)^2}{2\sigma_{min}(\Sigma)}$ where $||.||_2$ is the canonical Euclidean norm on $\mathbb{R}^n$.*

Proof. This theorem is a direct consequence of Lemma 2 and Proposition 2. By applying Lemma 2 to $\psi = \mathcal{N}_{|i}$ and Proposition 2 to $\rho = \phi^n \circ ... \circ \phi^{i+1}$, we immediately get the result. □

B.5 Proof of Theorem 2

Theorem 2 (Adversarial generalization gap bound in the randomized setting)**.** *Let* M *be the probabilistic mapping at hand. Let suppose that* M *is $d_{R,\lambda}$-(α,ϵ) robust for some $\lambda \geq 1$ then:*

[2] Although the Gaussian distribution belongs to the exponential family, it does not satisfy the modulus of continuity constraint on t and its robustness has to be proved differently.

$$|\operatorname{Risk}_\alpha(\mathrm{M}) - \operatorname{Risk}(\mathrm{M})| \leq 1 - e^{-\epsilon}\mathbb{E}_x\left[e^{-H(\mathrm{M}(x))}\right]$$

where H is the Shannon entropy: $H(p) = -\sum_i p_i \log(p_i)$

Proof. Let M be a randomized network with a noise X injected at layer i. We have:

$$\begin{aligned}
|\operatorname{Risk}_\alpha(\mathrm{M}) - \operatorname{Risk}(\mathrm{M})| &= \left|\mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{E}_{y'\sim\mathrm{M}(x+\tau)}\left[\mathbb{1}\left(y_1 \neq y\right)\right] - \mathbb{E}_{y_2\sim\mathrm{M}(x)}\left[\mathbb{1}\left(y' \neq y\right)\right]\right]\right| \\
&= \left|\mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{E}_{y_1\sim\mathrm{M}(x+\tau),y_2\sim\mathrm{M}(x)}\left[\mathbb{1}\left(y_1 \neq y\right) - \mathbb{1}\left(y_2 \neq y\right)\right]\right]\right| \\
&\leq \mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{E}_{y_1\sim\mathrm{M}(x+\tau),y_2\sim\mathrm{M}(x)}\left[\left|\mathbb{1}\left(y_1 \neq y\right) - \mathbb{1}\left(y_2 \neq y\right)\right|\right]\right] \\
&\leq \mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{E}_{y_1\sim\mathrm{M}(x+\tau),y_2\sim\mathrm{M}(x)}\left[\mathbb{1}\left(y_1 \neq y_2\right)\right]\right] \\
&= \mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{P}_{y_1\sim\mathrm{M}(x+\tau),y_2\sim\mathrm{M}(x)}(y_1 \neq y_2)\right]
\end{aligned}$$

For two discrete random independent variables of law $P = (p_1, ..., p_K)$ and $Q = (q_1, ..., q_K)$, thanks to Jensen's inequality:

$$\mathbb{P}(P = Q) = \sum_{i=1}^{K} p_i q_i \geq \exp\left(\sum_{i=1}^{K} p_i \log q_i\right) = \exp\left(-d_{KL}(P, Q) - H(P)\right)$$

Then we have:

$$\begin{aligned}
\mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} \mathbb{P}_{y_1\sim\mathrm{M}(x+\tau),y_2\sim\mathrm{M}(x)}(y_1 \neq y_2)\right] &\leq \mathbb{E}_{(x,y)}\left[\sup_{\tau/||\tau||\leq\alpha} 1 - e^{-d_{KL}(\mathrm{M}(x),\mathrm{M}(x+\tau))-H(\mathrm{M}(x))}\right] \\
&\leq \mathbb{E}_{(x,y)}\left[1 - e^{-\epsilon-H(\mathrm{M}(x))}\right] \\
&= 1 - e^{-\epsilon}\mathbb{E}_x\left[e^{-H(\mathrm{M}(x))}\right]
\end{aligned}$$

□

C Additional results and discussions

In this section, we give some additional results on both the strength of the Renyi-divergence and a bound on the generalization gap for TV-distance.

C.1 About Renyi divergence

In the main submission, we chose to use the Renyi-robustness as the principled measure of robustness. Since Renyi-divergence is a good surrogate for the trivial distance (which is a generalization of the $0-1$-loss for probabilistic mappings), we supported this statement by showing that Renyi-divergence is stronger than TV-distance. In this section, we extend this result to most of the classical divergences used in Machine Learning and show that Renyi-divergence is stronger than all of them.

Let us consider an output space $\mathcal{Y}$, $\mathcal{F}_\mathcal{Y}$ a σ-$algebra$ over $\mathcal{Y}$, and μ_1, μ_2, ν three measures on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$, with μ_1, μ_2 in the set of probability measures over $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ denoted $\mathcal{P}(\mathcal{Y})$. One has $\nu >> \mu_1, \mu_2$ and one denotes g_1 and g_2 the Radon-Nikodym derivatives with respect to ν.

The Separation distance:

$$d_S(\mu_1, \mu_2) := \sup_{\{z\} \in \mathcal{F}_\mathcal{Y}} 1 - \frac{\mu_1(\{z\})}{\mu_2(\{z\})}.$$

The Hellinger distance:

$$d_H(\mu_1, \mu_2) := \left[\int_\mathcal{Y} \left(\sqrt{g_1} - \sqrt{g_2} \right)^2 d\nu \right]^{1/2}.$$

The Prokhorov metric:

$$d_P(\mu_1, \mu_2) := \inf \left\{ \zeta > 0 : \mu_1(B) \leq \mu_2(B^\zeta) + \zeta \text{ for all Borel sets } B \right\} \text{ where } B^\zeta = \{x : \inf_{y \in B} d_{\mathcal{Y}'}(x, y) \leq \zeta\}.$$

The Discrepancy metric:

$$d_D(\mu_1, \mu_2) := \sup_{\text{all closed balls B}} |\mu_1(B) - \mu_2(B)|.$$

Lemma 3. *Given two probability measures μ_1 and μ_2 on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ the Separation metric and the Renyi divergence satisfy the following relation: $d_S(\mu_1, \mu_2) \leq d_{R,\infty}(\mu_1, \mu_2)$*

Proof. The function $x :\to 1-x-|\ln(x)|$ is negative on $\mathbb{R}$, therefore for any $\{z\} \in \mathcal{Y}$ one has $1 - \frac{\mu_1(\{z\})}{\mu_2(\{z\})} \leq \left| \ln \frac{\mu_1(\{z\})}{\mu_2(\{z\})} \right|$, hence $\sup_{\{z\} \in \mathcal{F}_\mathcal{Y}} 1 - \frac{\mu_1(\{z\})}{\mu_2(\{z\})} \leq \sup_{\{z\} \in \mathcal{F}_\mathcal{Y}} \left| \ln \frac{\mu_1(\{z\})}{\mu_2(\{z\})} \right| \leq \sup_{Z \in \mathcal{F}_\mathcal{Y}} \left| \ln \frac{\mu_1(Z)}{\mu_2(Z)} \right| = d_{R,\infty}(\mu_1, \mu_2)$ □

Theorem 3. *Let* M *be the probabilistic mapping, then for all $\lambda > 1$ if* M *is $d_{R,\lambda}$-$(\alpha, \epsilon, \gamma)$-robust the following assertions holds:*

(1) M *is d_H-$(\alpha, \sqrt{\epsilon}, \gamma)$-robust.*

(2) M *is d_P-$(\alpha, \epsilon', \gamma)$-robust* ***and*** *$d_D$-$(\alpha, \epsilon', \gamma)$-robust, for $\epsilon' = \min \left(\frac{3}{2} \left(\sqrt{1 + \frac{4\epsilon}{9}} - 1 \right)^{1/2}, \frac{\exp(\epsilon+1)-1}{\exp(\epsilon+1)+1} \right)$.*

(3) M *is d_W-$(\alpha, \epsilon', \gamma)$-robust with $\epsilon' = \min \left(\frac{3}{2diam(\mathcal{Y})} \left(\sqrt{1 + \frac{4\epsilon}{9}} - 1 \right)^{1/2}, \frac{\exp(\epsilon+1)-1}{diam(\mathcal{Y})(\exp(\epsilon+1)+1)} \right)$.*

(4) if $\lambda = \infty$, M *is d_S-$(\alpha, \epsilon, \gamma)$-robust.*

Proof.

(1) The proof is a simple adaptation of Proposition 1 using the inequality $d_H(\mu_1, \mu_2)^2 \leq d_{KL}(\mu_1, \mu_2)$ [19] and Lemma 1.

(2) Using the inequalities $d_D(\mu_1, \mu_2) \leq d_{TV}(\mu_1, \mu_2)$ [19] and $d_P(\mu_1, \mu_2) \leq d_{TV}(\mu_1, \mu_2)$ [25], the proof is immediate, using Theorem 1 and Lemma 1.

(3) The proof is adapted from Proposition 1 using the inequality $d_W(\mu_1, \mu_2) \leq diam(\mathcal{Y}) d_{TV}(\mu_1, \mu_2)$ [19].

(4) The result is a straightforward application of Lemma 3, and Lemma 1.

□

C.2 Generalization gap with TV-robustness

In the main paper, we give a bound on the generalization gap based on the Renyi-robustness. We extend this result to the TV-robustness, highlighting the fact that generalization gap could be derived from any of the above divergences.

Theorem 4. *Let* M *be the probabilistic mapping at hand. Let us suppose that* M *is d_{TV}-(α, ϵ) robust then:*

$$|\operatorname{Risk}_\alpha(\mathrm{M}) - \operatorname{Risk}(\mathrm{M})| \leq 1 - (\mathbb{E}_x \left[e^{-H_c(\mathrm{M}(x))} \right] - \epsilon)$$

where H_c is the collision entropy: $H_c(p) = -\log(\sum_i p_i^2)$

Proof. For two discrete random independent variables of law $P = (p_1, ..., p_K)$ and $Q = (q_1, ..., q_K)$, thanks to Jensen's inequality:

$$\mathbb{P}(P = Q) = \sum_{i=1}^K p_i q_i = \sum_{i=1}^K p_i^2 - \sum p_i (p_i - q_i) \geq e^{-H_c(P)} - d_{TV}(P, Q)$$

because, for any $i \in [K]$, $p_i - q_i \leq d_{TV}(P, Q)$.

Then, the proof is a simple adaptation of the model of proof from Theorem 2. □

D Additional empirical evaluation

Due to space limitations, we had to defer the thorough description of our experimental setup and the results of some additional experiments.

D.1 Architectures & Hyper-parameters

We conduct experiments with 3 different dataset:

- CIFAR-10 and CIFAR-100 datasets, which are composed of 50K training samples, 10000 test samples and respectively 10 and 100 different classes. Images are trained and evaluated with a resolution of 32 by 32 pixels.
- ImageNet dataset, which is composed of ~ 1.2M training examples, $50K$ test samples and 1000 classes. Images are trained and evaluated with a resolution of 299 by 299 pixels.

For CIFAR-10 and CIFAR-100 [26], we used a Wide ResNet architecture [50] which is a variant of the ResNet model from [24]. We used 28 layers with a widen factor of 10. We trained all the networks for 200 epochs, a batch size of 400, dropout 0.3 and Leaky Relu activation with a slope on $\mathbb{R}^-$ of 0.1. We used the cross entropy loss with Momentum 0.9 and a piecewise constant learning rate of 0.1, 0.02, 0.004 and 0.00008 after respectively 7500, 15000 and 20000 steps. The networks achieve for CIFAR10 and 100 a TOP-1 accuracy of 95.8% and 79.1% respectively on test images.

For ImageNet [12], we used an Inception ResNet v2 [45] which is the sate of the art architecture for this dataset and achieved a TOP-1 accuracy of 80%. For the training of ImageNet, we used the same hyper parameters setting as the original implementation. We trained the network for 120 epochs with a batch size of 256, dropout 0.8, Relu as activation function. All evaluations were done with a single crop on the non-blacklisted subset of the validation set.

D.2 Evaluation under attack

We evaluate our models against the strongest possible attacks from the literature using different norms (ℓ_1, ℓ_2 and ℓ_∞) which are all optimization based attacks. On their guide to evaluate robustness, Carlini et al. [6] proposed the three following attacks for each norm:

ℓ_2 – Carlini & Wagner attack and ℓ_1 – ElasticNet attack The ℓ_2 Carlini & Wagner attack ($C\&W$) introduced in [7] is formulated as:

$$\min_{x+r\in\mathcal{X}} c \times ||r||_2 + g(x+r)$$

where g is a function such that $g(y) \geq 0$ iff $f(y) = l'$ with l' the target class. The authors listed some g functions. we choose the following one:

$$g(x) = \max(F_{k(x)}(x) - \max_{i\neq k(x)}(F_i(x)), -\kappa)$$

where F is the softmax function and κ a positive constant.

Instead of using box-constrained L-BFGS [46] as in the original attack, the authors use instead a new variable for $x + r$:

$$x + r = \frac{1}{2}(\tanh(w) + 1)$$

Then a binary search is performed to optimize the constant c and ADAM or SGD for computing an optimal solution.

ℓ_1 – ElasticNet attack is an adaptation of ℓ_2 C&W attack where the objective is adaptive to ℓ_1 perturbations:

$$\min_{x+r\in\mathcal{X}} c_1 \times \|r\|_1 + c_2 \times \|r\|_2 + g(x+r)$$

ℓ_∞ – PGD attack. The PGD attack proposed by [31] is a generalization of the iterative FGSM attack proposed in [27]. The goal of the adversary is to solve the following problem:

$$\underset{\|r\|_p\leq\epsilon}{\operatorname{argmax}} \mathcal{L}(F_\theta(x+r), y)$$

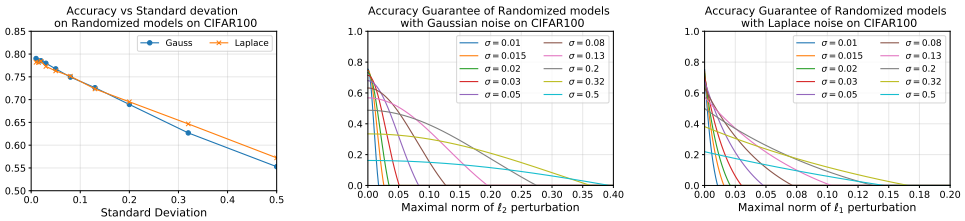
In practice, the authors proposed an iterative method to compute a solution:

$$x^{t+1} = P_{x \oplus r}(x^t + \alpha \operatorname{sign}(\nabla_x \mathcal{L}(F_\theta(x^t), y)))$$

Where $x \oplus r$ is the Minkowski sum between $\{x\}$ and $\{r \text{ s.t. } \|r\|_p \leq \epsilon\}$, α a gradient step size, P_S is the projection operator on S and x^0 is randomly chosen in $x \oplus r$.

D.3 Detailed results on CIFAR-10 and CIFAR-100

Figure 2: (a) Impact of the standard deviation of the injected noise on accuracy in a randomized model on CIFAR-100 dataset with a Wide ResNet architecture. (b) and (c) illustration of the guaranteed accuracy of different randomized models with Gaussian (b) and Laplace (c) noises given the norm of the adversarial perturbation.



(a) (b) (c)

Figure 3(a) presents the trade-off accuracy versus intensity of noise for the CIFAR-100 dataset. As for CIFAR-10, we observe that the accuracy decreases from 0.79 with a small noise (0.01) to ∼0.55 with a higher noise (0.5). The Figures 3(b) and 3(c) are coherent with the theoretical guarantee of accuracy (Theorem 2) that the model can achieve under attack with a given perturbation and noise.

Table 3 and 4 summarize the results on the accuracy and accuracy under attack of CIFAR-10 and CIFAR-100 datasets with a Randomized Wide ResNet architecture given the standard deviation of the injected noise and the number of iterations of the attack. For PGD, we use an epsilon max of 0.06 and a step size of 0.006 for an input space of between -1 and +1. We show that injecting noise empirically helps defending neural networks against adversarial attacks.

Table 3: Accuracy and Accuracy under attack of CIFAR-10 dataset

		ℓ_1 – EAD			ℓ_2 – C&W			ℓ_∞ – PGD		
	Natural	**20**	**50**	**60**	**20**	**50**	**60**	**10**	**15**	**20**
Normal (Sd)										
0.010	0.954		0.208	0.193	0.172	0.271	0.294	0.411	0.428	0.408
0.050	0.950	0.265	0.347	0.367	0.350	0.454	0.423	0.638	0.549	0.486
0.130	0.931	0.389	0.401	0.411	0.443	0.495	0.515	0.710	0.636	0.553
0.200	0.913	0.411	0.456		0.470	0.481	0.516	**0.724**	0.629	0.539
0.320	0.876	0.442	0.450	0.445	0.475	**0.522**	0.499	0.720	**0.641**	0.566
0.500	0.824	**0.453**	**0.513**	**0.448**	**0.503**	0.494	**0.523**	0.694	0.608	**0.587**
Laplace (Sd)										
0.010	0.955	0.167	0.190	0.208	0.184	0.279	0.313	0.474	0.423	0.389
0.050	0.950	0.326	0.315	0.355	0.387	0.458	0.448	0.630	0.534	0.515
0.130	0.929	0.388	0.426	0.435	0.461	**0.515**	0.493	0.688	0.599	0.538
0.200	0.919	0.417		**0.464**	0.484	0.481	0.501	0.730	0.600	0.569
0.320	0.891	**0.460**	0.443	0.448	0.472	0.499	**0.520**	**0.750**	**0.665**	0.576
0.500	0.846	0.454	**0.471**	**0.464**	**0.488**		0.494	0.721	0.650	**0.589**
Exponential (Sd)										
0.010	0.953	0.153	0.174		0.228	0.292	0.306	0.443	0.404	0.395
0.050	0.953	0.312	0.326	0.330	0.343	0.468	0.435	0.616	0.575	0.479
0.130	0.940	0.373	0.402	0.411	0.424		0.504	0.679	0.585	0.526
0.200	0.936	0.394		0.414	0.455	**0.510**	0.501	0.701	0.623	0.550
0.320	0.919	**0.429**	0.426	0.416	**0.494**	0.492	0.513	0.739	0.638	0.564
0.500	0.900	0.423	**0.454**	**0.470**	0.488	0.494	**0.516**	**0.752**	**0.699**	**0.594**

D.4 Large scale robustness

Adversarial training fails to generalize to higher dimensional datasets such as ImageNet. We conducted experiments with the large scale ImageNet dataset and compared our randomized neural network against large scale adversarial

Table 4: Accuracy and Accuracy under attack of CIFAR-100 dataset.

		ℓ_1 – EAD			ℓ_2 – C&W			ℓ_∞ – PGD		
	Natural	**20**	**50**	**60**	**20**	**50**	**60**	**10**	**15**	**20**
Normal (Sd)										
0.010	0.790	0.235	0.234	0.228	0.235	0.318	0.316	0.257	0.176	0.187
0.050	0.768	0.321	0.294	0.320	0.357	0.377	0.410	0.377	0.296	0.254
0.130	0.726	**0.357**	**0.371**	0.349	0.387	**0.427**	**0.428**	0.414	0.319	0.260
0.200	0.689	0.338	0.350	**0.384**	0.394	0.381		0.439	0.356	0.277
0.320	0.627	0.334	0.344	0.350	0.328	0.364	0.400	**0.441**	0.366	0.299
0.500	0.553	0.322	0.331	0.331	0.349	0.342	0.351	0.408	**0.374**	**0.308**
Laplace (Sd)										
0.010	0.782	0.199	0.227	0.243	0.225	0.311	0.321	0.236	0.190	0.177
0.050	0.763	0.326	0.317	0.331	0.354	0.377	0.409	0.368	0.319	0.256
0.130	0.723	0.337	0.357	0.344	**0.408**	0.414	0.408	0.420	0.346	0.293
0.200	0.695	**0.355**	0.349	**0.361**	0.393	0.405	0.393	0.445	0.340	0.303
0.320	0.647	0.324	**0.373**	0.357	0.388	0.387	0.373	**0.460**	0.381	0.303
0.500	0.572	0.310	0.308	0.323	0.358	0.351	0.361	0.425	**0.403**	**0.329**
Exponential (Sd)										
0.010	0.785	0.218	0.251	0.217	0.247	0.278	0.321	0.250	0.214	0.169
0.050	0.767	0.323	0.337	0.317	0.346	0.380	0.402	0.356	0.291	0.235
0.130	0.749	0.330		0.356	**0.403**	**0.444**	**0.421**	0.400	0.328	0.266
0.200	0.731	0.345	0.361	0.357	0.388	0.424	0.406	0.427	0.340	0.267
0.320	0.703	0.349	0.351	0.340	0.388	0.439	0.399	0.433	0.351	0.280
0.500	0.673	**0.387**	**0.378**	**0.378**	0.396	0.435		**0.485**	**0.370**	**0.322**

training proposed by Kurakin et al. [27]. One can observe from Table 5 that the model from Kurakin et al. is neither robust against recent ℓ_1 nor ℓ_2 iterative attacks such as EAD and C&W. Moreover, it offers a small robustness against ℓ_∞ PGD attack. Our randomized neural network with EoT attacks offers a small robustness on ℓ_1 and ℓ_2 attacks while being less robust against PGD.

Table 5: Accuracy under attack of the Adversarial model training by Kurakin et al. [27] and an Inception Resnet v2 model training with normal 0.1 noise injected in the image on the ImageNet dataset.

	Baseline	ℓ_1 **EAD 60**	ℓ_2 **C&W 60**	ℓ_∞ **PGD**
Kurakin et al. [27]	0.739	0.097	0.100	0.239
Normal 0.1	0.625	0.255	0.301	0.061

D.5 Experiments with noise on the first activation

The aim of the following experiments is empirically illustrate the *Data processing inequality* in Proposition 2.

Table 6 and 7 present the experiments conducted with the same set of parameters as the previous ones on CIFAR-10 and CIFAR-100, but with the noise injected in the first activation layer instead of directly in the image. We observe from Table 6 that we can inject more noise with a marginal loss on accuracy. The accuracy under attack is presented in Table 7 for a selection of models.

Table 6: Impact of the distribution and the intensity of the noise with randomized networks with noise injected on the first activation

Sd	**Normal**	**Sd**	**Laplace**	**Sd**	**Exponential**
0.01	0.956	0.01	0.955	0.01	0.953
0.23	0.943	0.05	0.947	0.08	0.943
0.45	0.935	0.10	0.933	0.15	0.938
0.68	0.926	0.15	0.916	0.23	0.925
0.90	0.916	0.20	0.911	0.30	0.919
1.00	0.916	0.25	0.897	0.38	0.903
1.34	0.906	0.30	0.889	0.45	0.897
1.55	0.900	0.35	0.882	0.53	0.886
1.77	0.893	0.40	0.867	0.60	0.885
2.00	0.886	0.45	0.855	0.68	0.875

Table 7: Accuracy and Accuracy under attack of selected models with noise on the first activation

				ℓ_1 – EAD			ℓ_2 – C&W			ℓ_∞ – PGD	
			Natural	20	50	60	20	50	60	10	20
Dataset	**Distribution**	**Sd**									
CIFAR10	Normal	1.55	0.900	0.441	0.440	0.413	0.477	0.482	0.484	0.683	0.526
	Laplace	0.25	0.897	0.388	0.436	**0.445**	0.481	0.506	0.491	0.664	0.493
	Exponential	0.38	**0.903**	**0.456**	**0.463**	0.438	**0.495**	**0.516**	**0.506**	**0.697**	**0.557**
CIFAR100	Normal	0.45	0.741	**0.362**	0.352	0.353	0.352	0.410	0.418	0.380	0.250
	Laplace	0.10	**0.742**	0.350	**0.367**	0.350	0.371	**0.419**		0.418	**0.264**
	Exponential	0.15	0.741	0.354	0.356	**0.373**	**0.394**	0.409	**0.420**	**0.430**	0.258

E Additional discussions on the experiments

For the sake of completeness and reproducibility, we give some additional insights on the noise injection scheme and comprehensive details on our numerical experiments.

E.1 On the need for injecting noise in the training phase

Robustness has always been thought as a property to be enforced at inference time and it is tempting to focus only on injecting noise at inference. However, simply doing so ruins the accuracy of the algorithm (as it becomes an instance of distribution shift [44]). Indeed, making the assumption that the training and test distributions matches, in practice, injecting some noise at inference would result in changing the test distribution.

Distribution shift occurs when the training distribution differs from the test distribution. This implies that the hypothesis minimizing the empirical risk is not consistent, i.e. it does not converge to the true model as the training size increases. A way to circumvent that is to ensure that training and test distributions matches using importance weighting (in the case of covariate-shift) or with noise injection in the training phases as well (in our case).

E.2 Reproducibility of the experiments

We emphasize that all experiments should be easily reproducible. All our experiments are developed with TensorFlow version 1.12 [1]. The code is available as supplemental material and will be open sourced upon acceptance of the paper. The archive contains a *readme* file containing a small documentation on how to run the experiments, a configuration file which defines the architecture and the hyper-parameters of the experiments, python scripts which generate a bash command to run the experiments. The code contains Randomized Wide Resnet used for CIFAR-10 and CIFAR100, Inception Resnet v2 used for ImageNet, PGD, EAD and C&W attacks used for evaluation under attack. We ran our experiments, on a cluster with computers each having 8 GPU Nvidia V100.