A PAC-Bayes Analysis of Adversarial Robustness

Paul Viallard, Guillaume Vidot, Amaury Habrard, Emilie Morvant

A PAC-Bayes Analysis of Adversarial Robustness

Paul Viallard[1]*, **Guillaume Vidot**[23]*, **Amaury Habrard**[1], **Emilie Morvant**[1]
[1] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
[2] Airbus Opération S.A.S
[3] University of Toulouse, Institut de Recherche en Informatique de Toulouse, France

Abstract

We propose the first general PAC-Bayesian generalization bounds for adversarial robustness, that estimate, at test time, how much a model will be invariant to imperceptible perturbations in the input. Instead of deriving a worst-case analysis of the risk of a hypothesis over all the possible perturbations, we leverage the PAC-Bayesian framework to bound the averaged risk on the perturbations for majority votes (over the whole class of hypotheses). Our theoretically founded analysis has the advantage to provide general bounds *(i)* that are valid for any kind of attacks (*i.e.*, the adversarial attacks), *(ii)* that are tight thanks to the PAC-Bayesian framework, *(iii)* that can be directly minimized during the learning phase to obtain a robust model on different attacks at test time.

1 Introduction

While machine learning algorithms are able to solve a huge variety of tasks, Szegedy et al. [2014] pointed out a crucial *weakness*: the possibility to generate samples similar to the originals (*i.e.*, with no or insignificant change recognizable by the human eyes) but with a different outcome from the algorithm. This phenomenon, known as "adversarial examples", contributes to the impossibility to ensure the safety of machine learning algorithms for safety-critical applications such as aeronautic functions (*e.g.*, vision-based navigation), autonomous driving, or medical diagnosis (see, *e.g.*, Huang et al. [2020]). Adversarial robustness is thus a critical issue in machine learning that studies the ability of a model to be robust or invariant to perturbations of its input. A perturbed input that fools the model is usually called an *adversarial example*. In other words, an adversarial example can be defined as an example that has been modified by an imperceptible noise (or that does not exceed a threshold) but which leads to a misclassification. One line of research is referred to as adversarial robustness verification [*e.g.*, Gehr et al., 2018, Huang et al., 2017, Singh et al., 2019, Tsuzuku et al., 2018], where the objective is to formally check whether the neighborhood of each sample does not contain any adversarial examples. This kind of method comes with some limitations such as scalability or overapproximation [Gehr et al., 2018, Katz et al., 2017, Singh et al., 2019]. In this paper we stand in another setting called adversarial attack/defense [*e.g.*, Papernot et al., 2016, Goodfellow et al., 2015, Madry et al., 2018, Carlini and Wagner, 2017, Zantedeschi et al., 2017, Kurakin et al., 2017]. An adversarial attack consists in finding perturbed examples that defeat machine learning algorithms while the adversarial defense techniques enhance their adversarial robustness to make the attacks useless. While a lot of methods exist, adversarial robustness suffers from a lack of general theoretical understandings (see Section 2.2).

To tackle this issue, we propose in this paper to formulate the adversarial robustness through the lens of a well-founded statistical machine learning theory called PAC-Bayes and introduced by Shawe-Taylor and Williamson [1997], McAllester [1998]. This theory has the advantage to provide tight

*Paul Viallard and Guillaume Vidot contributed equally to this work

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

generalization bounds in average over the set of hypotheses considered (leading to bounds for a weighted majority vote over this set), in contrast to other theories such as VC-dimension or Rademacher-based approaches that give worst-case analysis, *i.e.*, for all the hypotheses. We start by defining our setting called *adversarially robust PAC-Bayes*. The idea consists in considering an *averaged adversarial robustness risk* which corresponds to the probability that the model misclassifies a perturbed example (this can be seen as an averaged risk over the perturbations). This measure can be too optimistic and not enough informative since for each example we sample only one perturbation. Thus we also define an *averaged-max adversarial risk* as the probability that there exists at least one perturbation (taken in a set of sampled perturbations) that leads to a misclassification. These definitions, based on averaged quantities, have the advantage *(i)* of still being suitable for the PAC-Bayesian framework and majority vote classifiers and *(ii)* to be related to the classical adversarial robustness risk. Then, for each of our adversarial risks, we derive a PAC-Bayesian generalization bound that can are valid to any kind of attack. From an algorithmic point of view, these bounds can be directly minimized in order to learn a majority vote robust in average to attacks. We empirically illustrate that our framework is able to provide generalization guarantees with non-vacuous bounds for the adversarial risk while ensuring efficient protection to adversarial attacks.

Organization of the paper. Section 2 recalls basics on usual adversarial robustness. We state our new adversarial robustness PAC-Bayesian setting along with our theoretical results in Section 3, and we empirically show its soundness in Section 4. All the proofs of the results are deferred in Appendix.

2 Basics on adversarial robustness

2.1 General setting

We tackle binary classification tasks with the input space $X=\mathbb{R}^d$ and the output/label space $Y=\{-1,+1\}$. We assume that D is a fixed but unknown distribution on $X\times Y$. An example is denoted by $(x,y)\in X\times Y$. Let $S=\{(x_i,y_i)\}_{i=1}^m$ be the learning sample consisted of m examples *i.i.d.* from D; We denote the distribution of such m-sample by D^m. Let $\mathcal{H}$ be a set of real-valued functions from X to $[-1,+1]$ called voters or hypotheses. Usually, given a learning sample $S\sim D^m$, a learner aims at finding the best hypothesis h from $\mathcal{H}$ that commits as few errors as possible on unseen data from D. One wants to find $h\in\mathcal{H}$ that minimizes the true risk $R_D(h)$ on D defined as

$$R_D(h) = \mathop{\mathbb{E}}_{(x,y)\sim D} \ell\left(h,(x,y)\right), \tag{1}$$

where $\ell:\mathcal{H}\times X\times Y\to\mathbb{R}^+$ is the loss function. In practice since D is unknown we cannot compute $R_D(h)$, we usually deal with the empirical risk $R_S(h)$ estimated on S and defined as

$$R_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h,(x_i,y_i)).$$

From a classic ideal machine learning standpoint, we are able to learn a well-performing classifier with strong guarantees on unseen data, and even to measure how much the model will be able to generalize on D (*e.g.*, with generalization bounds).

However, in real-life applications at classification time, an imperceptible perturbation of the input (*e.g.*, due to a malicious attack or a noise) can have a bad influence on the classification performance on unseen data [Szegedy et al., 2014]: the usual guarantees do not stand anymore. Such imperceptible perturbation can be modeled by a (relatively small) noise in the input. Let $b>0$ and $\|\cdot\|$ be an arbitrary norm (the most used norms are the ℓ_1, ℓ_2 and ℓ_∞-norms), the set of possible noises B is defined by

$$B=\left\{\epsilon\in\mathbb{R}^d \;\middle|\; \|\epsilon\|\leq b\right\}.$$

The learner aims to find an *adversarial robust* classifier that is robust in average to all noises in B over $(x,y)\sim D$. More formally, one wants to minimize the adversarial robust true risk $R_D^{\texttt{ROB}}(h)$ defined as

$$R_D^{\texttt{ROB}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim D} \max_{\epsilon\in B} \ell\left(h,(x+\epsilon,y)\right). \tag{2}$$

Similarly as in the classic setting, since D is unknown, $R_D^{\texttt{ROB}}(h)$ cannot be directly computed, and then one usually deals with the empirical adversarial risk

$$R_S^{\texttt{ROB}}(h) = \frac{1}{m}\sum_{i=1}^{m} \max_{\epsilon\in B}\ell\left(h,(x_i+\epsilon,y_i)\right).$$

That being said, a learned classifier h should be robust to *adversarial attacks* that aim at finding an *adversarial example* $x+\epsilon^*(x,y)$ to fool h for given example (x, y), where $\epsilon^*(x,y)$ is defined as

$$\epsilon^*(x,y) \in \operatorname{argmax}_{\epsilon\in B} \ \ell(h, (x+\epsilon, y)). \tag{3}$$

In consequence, *adversarial defense* mechanisms often rely on the adversarial attacks by replacing the original examples with the adversarial ones during the learning phase; This procedure is called adversarial training. Even if there are other defenses, adversarial training appears to be one of the most efficient defense mechanisms [Ren et al., 2020].

2.2 Related works

Adversarial Attacks/Defenses. Numerous methods[2] exist to solve–or approximate–the optimization of Equation (3). Among them, the Fast Gradient Sign Method (FGSM Goodfellow et al. [2015]) is an attack consisting in generating a noise ϵ in the direction of the gradient of the loss function with respect to the input x. Kurakin et al. [2017] introduced IFGSM, an iterative version of FGSM: at each iteration, one repeats FGSM and adds to x a noise, that is the sign of the gradient of the loss with respect to x. Following the same principle as IFGSM, Madry et al. [2018] proposed a method based on Projected Gradient Descent (PGD) that includes a random initialization of x before the optimization. Another technique known as the *Carlini and Wagner Attack* [Carlini and Wagner, 2017] aims at finding adversarial examples $x+\epsilon^*(x,y)$ that are as close as possible to the original x, *i.e.*, they want an attack being the most imperceptible as possible. However, producing such imperceptible perturbation leads to a high-running time in practice. Contrary to the most popular techniques that look for a model with a low adversarial robust risk (Equation (2)), our work stands in another line of research where the idea is to relax this worst-case risk measure by considering an *averaged* adversarial robust risk over the noises instead of a $\max$-based formulation [see, *e.g.*, Zantedeschi et al., 2017, Hendrycks and Dietterich, 2019]. Our averaged formulation is introduced in the Section 3.

Generalization Bounds. Recently, few generalization bounds for adversarial robustness have been introduced [*e.g.* Khim and Loh, 2018, Yin et al., 2019, Montasser et al., 2019, 2020, Cohen et al., 2019, Salman et al., 2019]. Khim and Loh, and Yin et al.'s results are Rademacher complexity-based bounds. The former makes use of a surrogate of the adversarial risk; The latter provides bounds in the specific case of neural networks and linear classifiers, and involves an unavoidable polynomial dependence on the dimension of the input. Montasser et al. study robust PAC-learning for PAC-learnable classes with finite VC-dimension for unweighted majority votes that have been "robustified" with a boosting algorithm. However, their algorithm requires to consider all possible adversarial perturbations for each example which is intractable in practice, and their bound suffers also from a large constant as indicated at the end of the Montasser et al. [Theorem 3.1 2019]'s proof. Cohen et al. provide bounds that estimate what is the minimum noise to get an adversarial example (in the case of perturbations expressed as Gaussian noise) while our results give the probability to be fooled by an adversarial example. Salman et al. leverage Cohen et al.'s method and adversarial training in order to get tighter bounds. Moreover, Farnia et al. present margin-based bounds on the adversarial robust risk for specific neural networks and attacks (such as FGSM or PGD). While they made use of a classical PAC-Bayes bound, their result is not a PAC-Bayesian analysis and stands in the family of uniform-convergence bounds [see Nagarajan and Kolter, 2019, Ap. J for details]. In this paper, we provide PAC-Bayes bounds for general models expressed as majority votes, their bounds are thus not directly comparable to ours.

3 Adversarially robust PAC-Bayes

Although few theoretical results exist, the majority of works come either without theoretical guarantee or with very specific theoretical justifications. In the following, we aim at giving a different point of view on adversarial robustness based on the so-called PAC-Bayesian framework. By leveraging this framework, we derive a general generalization bound for adversarial robustness based on an averaged notion of risk that allows us to learn robust models at test time. We introduce below our new setting referred to as adversarially robust PAC-Bayes.

[2]The reader can refer to Ren et al. [2020] for a survey on adversarial attacks and defenses.

3.1 Adversarially robust majority vote

The PAC-Bayesian framework provides practical and theoretical tools to analyze majority vote classifiers. Assuming the voters set $\mathcal{H}$ and a learning sample S as defined in Section 2, our goal is not anymore to learn one classifier from $\mathcal{H}$ but to learn a well-performing weighted combination of the voters involved in $\mathcal{H}$, the weights being modeled by a distribution $\mathcal{Q}$ on $\mathcal{H}$. This distribution is called the posterior distribution and is learned from S given a prior distribution $\mathcal{P}$ on $\mathcal{H}$. The learned weighted combination is called a $\mathcal{Q}$-weighted majority vote and is defined by

$$\forall x \in X, \quad H_{\mathcal{Q}}(x) = \operatorname{sign}\left[\mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x)\right]. \tag{4}$$

In the rest of the paper, we consider the 0-1 loss function classically used for majority votes in PAC-Bayes and defined as $\ell(h,(x,y))=\mathbf{I}\left(h(x)\neq y\right)$ with $\mathbf{I}(a)=1$ if a is true, and 0 otherwise. In this context, the adversarial perturbation related to Equation (3) becomes

$$\epsilon^*(x,y) \in \operatorname{argmax}_{\epsilon\in B} \mathbf{I}(H_{\mathcal{Q}}(x+\epsilon)\neq y). \tag{5}$$

Optimizing this problem is intractable due to the non-convexity of $H_{\mathcal{Q}}$ induced by the sign function. Note that the adversarial attacks of the literature (like PGD or IFGSM) aim at finding the optimal perturbation $\epsilon^*(x,y)$, but, in practice one considers an approximation of this perturbation.

Hence, instead of searching for the noise that maximizes the chance of fooling the algorithm, we propose to model the perturbation according to an example-dependent distribution. First let us define $\omega_{(x,y)}$ a distribution, on the set of possible noises B, that is dependent on an example $(x,y)\in X\times Y$. Then we denote as $\mathbf{D}$ the distribution on $(X\times Y)\times B$ defined as $\mathbf{D}((x,y),\epsilon)=D(x,y)\cdot\omega_{(x,y)}(\epsilon)$ which further permits to generate *perturbed examples*. To estimate our risks (defined below) for a given example $(x_i,y_i)\sim D$, we consider a set of n perturbations sampled from $\omega_{(x_i,y_i)}$ denoted by $\boldsymbol{\mathcal{E}}_i=\{\epsilon_j^i\}_{j=1}^n$. Then we consider as a learning set the $m\times n$-sample $\mathbf{S}=\{((x_i,y_i),\boldsymbol{\mathcal{E}}_i)\}_{i=1}^m\in(X\times Y\times B^n)^m$. In other words, each $((x_i,y_i),\boldsymbol{\mathcal{E}}_i)\in\mathbf{S}$ is sampled from a distribution that we denote by $\mathbf{D}^n$ such that

$$\mathbf{D}^n((x_i,y_i),\boldsymbol{\mathcal{E}}_i) = D(x_i,y_i)\cdot\prod_{j=1}^n \omega_{(x_i,y_i)}(\epsilon_j^i).$$

Then, inspired by the works of Zantedeschi et al. [2017], Hendrycks and Dietterich [2019], we define our *robustness averaged adversarial risk* as follows.

Definition 1 (Averaged Adversarial Risk)**.** *For any distribution $\mathbf{D}$ on $(X\times Y)\times B$, for any distribution $\mathcal{Q}$ on $\mathcal{H}$, the averaged adversarial risk of $H_{\mathcal{Q}}$ is defined as*

$$\begin{aligned} R_{\mathbf{D}}(H_{\mathcal{Q}}) &= \Pr_{((x,y),\epsilon)\sim\mathbf{D}} (H_{\mathcal{Q}}(x+\epsilon)\neq y) \\ &= \mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}} \mathbf{I}(H_{\mathcal{Q}}(x+\epsilon)\neq y). \end{aligned}$$

The empirical averaged adversarial risk is computed on a $m\times n$-sample $\mathbf{S}=\{((x_i,y_i),\boldsymbol{\mathcal{E}}_i)\}_{i=1}^m$ is

$$R_{\mathbf{S}}(H_{\mathcal{Q}}) = \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \mathbf{I}(H_{\mathcal{Q}}(x_i+\epsilon_j^i)\neq y_i).$$

As we will show in Proposition 3, the risk $R_{\mathbf{D}}(H_{\mathcal{Q}})$ can considered optimistic regarding $\epsilon^*(x,y)$ of Equation (5). Indeed, instead of taking the ϵ maximizing the loss, a unique ϵ is drawn from a distribution. Hence, it can lead to a non-informative risk regarding the occurrence of adversarial examples. To overcome this, we propose an extension that we refer as *averaged-max adversarial risk*.

Definition 2 (Averaged-Max Adversarial Risk)**.** *For any distribution $\mathbf{D}$ on $(X\times Y)\times B$, for any distribution $\mathcal{Q}$ on $\mathcal{H}$, the averaged-max adversarial risk of $H_{\mathcal{Q}}$ is defined as*

$$A_{\mathbf{D}^n}(H_{\mathcal{Q}}) = \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^n} \big(\exists\,\epsilon\in\boldsymbol{\mathcal{E}}, H_{\mathcal{Q}}(x+\epsilon)\neq y\big).$$

The empirical averaged-max adversarial risk computed on a $m\times n$-sample $\mathbf{S}=\{((x_i,y_i),\boldsymbol{\mathcal{E}}_i)\}_{i=1}^m$ is

$$A_{\mathbf{S}}(H_{\mathcal{Q}}) = \frac{1}{m}\sum_{i=1}^m \max_{\epsilon\in\boldsymbol{\mathcal{E}}_i} \mathbf{I}(H_{\mathcal{Q}}(x_i+\epsilon)\neq y_i).$$

For an example $(x,y)\sim D$, instead of checking if one perturbed example $x+\epsilon$ is adversarial, we sample n perturbed examples $x+\epsilon_1,\ldots,x+\epsilon_n$ and we check if at least one example is adversarial.

3.2 Relations between the adversarial risks

Proposition 3 below shows the intrinsic relationships between the classical adversarial risk $R_D^{\texttt{ROB}}(H_\mathcal{Q})$ and our two relaxations $R_\mathbf{D}(H_\mathcal{Q})$ and $A_{\mathbf{D}^n}(H_\mathcal{Q})$. In particular, Proposition 3 shows that the larger n, the number of perturbed examples, the higher is the chance to get an adversarial example and then to be close to the adversarial risk $R_D^{\texttt{ROB}}(H_\mathcal{Q})$.

Proposition 3. *For any distribution* $\mathbf{D}$ *on* $(X\times Y)\times B$*, for any distribution* $\mathcal{Q}$ *on* $\mathcal{H}$*, for any* $(n, n') \in \mathbb{N}^2$*, with* $n \geq n' \geq 1$*, we have*

$$R_\mathbf{D}(H_\mathcal{Q}) \;\leq\; A_{\mathbf{D}^{n'}}(H_\mathcal{Q}) \;\leq\; A_{\mathbf{D}^n}(H_\mathcal{Q}) \;\leq\; R_D^{\texttt{ROB}}(H_\mathcal{Q}). \tag{6}$$

The left-hand side of Equation (6) confirms that the averaged adversarial risk $R_\mathbf{D}(H_\mathcal{Q})$ is optimistic regarding the classical $R_D^{\texttt{ROB}}(H_\mathcal{Q})$. Proposition 4 estimates how close $R_\mathbf{D}(H_\mathcal{Q})$ can be to $R_D^{\texttt{ROB}}(H_\mathcal{Q})$.

Proposition 4. *For any distribution* $\mathbf{D}$ *on* $(X \times Y) \times B$*, for any distribution* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$R_D^{\texttt{ROB}}(H_\mathcal{Q}) - \mathrm{TV}(\Pi\|\Delta) \;\leq\; R_\mathbf{D}(H_\mathcal{Q}),$$

where Δ *and* Π *are distributions on* $X\times Y$*, and* $\Delta(x', y')$*, respectively* $\Pi(x', y')$*, corresponds to the probability of drawing a perturbed example* $(x+\epsilon)$ *with* $((x, y), \epsilon)\sim\mathbf{D}$*, respectively an adversarial example* $(x+\epsilon^*(x,y), y)$ *with* $(x, y)\sim D$*. We have*

$$\Delta(x', y') = \Pr_{((x,y),\epsilon)\sim\mathbf{D}}\left[x+\epsilon=x', y=y'\right], \quad \textit{and} \quad \Pi(x', y') = \Pr_{(x,y)\sim D}\left[x+\epsilon^*(x, y)=x', y=y'\right], \tag{7}$$

and $\mathrm{TV}(\Pi\|\Delta) = \mathop{\mathbb{E}}_{(x',y')\sim\Delta} \frac{1}{2}\left|\frac{\Pi(x',y')}{\Delta(x',y')}-1\right|$*, is the Total Variation (TV) distance between* Π *and* Δ*.*

Note that $\epsilon^*(x,y)$ depends on $\mathcal{Q}$, and hence Π depends on $\mathcal{Q}$. From Equation (7), $R_D^{\texttt{ROB}}(H_\mathcal{Q})$ and $R_\mathbf{D}(H_\mathcal{Q})$ can be rewritten (see Lemmas 8 and 9 in Appendix B) respectively with Δ and Π as

$$R_\mathbf{D}(H_\mathcal{Q}) = \Pr_{(x',y')\sim\Delta}\left[H_\mathcal{Q}(x') \neq y'\right], \quad \text{and} \quad R_D^{\texttt{ROB}}(H_\mathcal{Q}) = \Pr_{(x',y')\sim\Pi}\left[H_\mathcal{Q}(x') \neq y'\right].$$

Finally, Propositions 3 and 4 relate the adversarial risk $R_\mathbf{D}(H_\mathcal{Q})$ to the "standard" adversarial risk $R_D^{\texttt{ROB}}(H_\mathcal{Q})$. Indeed, by merging the two propositions we obtain

$$R_D^{\texttt{ROB}}(H_\mathcal{Q}) - \mathrm{TV}(\Pi\|\Delta) \;\leq\; R_\mathbf{D}(H_\mathcal{Q}) \;\leq\; A_{\mathbf{D}^n}(H_\mathcal{Q}) \;\leq\; R_D^{\texttt{ROB}}(H_\mathcal{Q}). \tag{8}$$

Hence, the smaller the TV distance $\mathrm{TV}(\Pi\|\Delta)$, the closer the averaged adversarial risk $R_\mathbf{D}(H_\mathcal{Q})$ is from $R_D^{\texttt{ROB}}(H_\mathcal{Q})$ and the more probable an example $((x, y), \epsilon)$ sampled from $\mathbf{D}$ would be adversarial, *i.e.*, when our "averaged" adversarial example looks like a "specific" adversarial example. Moreover, Equation (8) justifies that the PAC-Bayesian point of view makes sense for adversarial learning with theoretical guarantees: the PAC-Bayesian guarantees we derive in the next section for our adversarial risks also give some guarantees on the "standard risk" $R_D^{\texttt{ROB}}(H_\mathcal{Q})$.

3.3 PAC-Bayesian bounds on the adversarially robust majority vote

First of all, since $R_\mathbf{D}(H_\mathcal{Q})$ and $A_{\mathbf{D}^n}(H_\mathcal{Q})$ risks are not differentiable due to the indicator function, we propose to use a common surrogate in PAC-Bayes (known as the Gibbs risk): instead of considering the risk of the $\mathcal{Q}$-weighted majority vote, we consider the expectation over $\mathcal{Q}$ of the individual risks of the voters involved in $\mathcal{H}$. In our case, we define the surrogates with the linear loss as

$$\overline{R_\mathbf{D}}(H_\mathcal{Q}) = \mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}} \frac{1}{2}\left[1-y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon)\right],$$

$$\text{and} \quad \overline{A_{\mathbf{D}^n}}(H_\mathcal{Q}) = \mathop{\mathbb{E}}_{((x,y),\boldsymbol{\epsilon})\sim\mathbf{D}^n} \frac{1}{2}\left[1-\min_{\epsilon\in\boldsymbol{\epsilon}}\left(y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon)\right)\right].$$

The next theorem relates these surrogates to our risks, implying that a generalization bound for $\overline{R_\mathbf{D}}(H_\mathcal{Q})$, *resp.* for $\overline{A_{\mathbf{D}^n}}(H_\mathcal{Q})$, leads to a generalization bound for $R_\mathbf{D}(H_\mathcal{Q})$, *resp.* $A_{\mathbf{D}^n}(H_\mathcal{Q})$.

Theorem 5. *For any distributions* $\mathbf{D}$ *on* $(X\times Y)\times B$ *and* $\mathcal{Q}$ *on* $\mathcal{H}$*, for any* $n>1$*, we have*

$$R_\mathbf{D}(H_\mathcal{Q}) \leq 2\overline{R_\mathbf{D}}(H_\mathcal{Q}), \qquad \textit{and} \qquad A_{\mathbf{D}^n}(H_\mathcal{Q}) \leq 2\overline{A_{\mathbf{D}^n}}(H_\mathcal{Q}).$$

Theorem 6 below presents our PAC-Bayesian generalization bounds for $\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}})$. Before that, it is important to mention that the empirical counterpart of $\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}})$ is computed on $\mathbf{S}$ which is composed of non identically independently distributed samples, meaning that a "classical" proof technique is not applicable. The trick here is to make use of a result of Ralaivola et al. [2010] that provides a *chromatic PAC-Bayes bound*, *i.e.*, a bound which supports non-independent data.

Theorem 6. *For any distribution* $\mathbf{D}$ *on* $(X\times Y)\times B$*, for any set of voters* $\mathcal{H}$*, for any prior* $\mathcal{P}$ *on* $\mathcal{H}$*, for any* n*, with probability at least* $1-\delta$ *over* $\mathbf{S}$*, for all posteriors* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$\mathrm{kl}(\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}})) \leq \frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{m+1}{\delta}\right], \tag{9}$$

$$\textit{and}\quad \overline{R_{\mathbf{D}}}(H_{\mathcal{Q}}) \leq \overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{m+1}{\delta}\right]}, \tag{10}$$

$$\textit{where}\quad \overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[1-y_i \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x_i+\epsilon_j^i)\right],$$

$\mathrm{kl}(a\|b)=a\ln\frac{a}{b}+(1-a)\ln\frac{1-a}{1-b}$*, and* $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})=\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\ln\frac{\mathcal{P}(h)}{\mathcal{Q}(h)}$ *the KL-divergence between* $\mathcal{P}$ *and* $\mathcal{Q}$*.*

Surprisingly, this theorem states bounds that do not depend on the number of perturbed examples n but only on the number of original examples m. The reason is that the n perturbed examples are inter-dependent (see the proof in Appendix). Note that Equation (9) is expressed as a Seeger [2002]'s bound and is tighter but less interpretable than Equation (10) expressed as a McAllester [1998]'s bound; These bounds involve the usual trade-off between the empirical risk $\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}})$ and $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$.

We now state a generalization bound for $\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}})$. Since this value involves a minimum term, we cannot use the same trick as for Theorem 6. To bypass this issue, we use the TV distance between two "artificial" distributions on $\boldsymbol{\mathcal{E}}_i$. Given $((x_i,y_i),\boldsymbol{\mathcal{E}}_i)\in\mathbf{S}$, let π_i be an arbitrary distribution on $\boldsymbol{\mathcal{E}}_i$, and given $h\in\mathcal{H}$, let ρ_i^h be a Dirac distribution on $\boldsymbol{\mathcal{E}}_i$ such that $\rho_i^h(\epsilon)=1$ if $\epsilon=\mathrm{argmax}_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\frac{1}{2}\left[1-y_ih(x_i+\epsilon)\right]$ (*i.e.*, if ϵ is maximizing the linear loss), and 0 otherwise.

Theorem 7. *For any distribution* $\mathbf{D}$ *on* $(X\times Y)\times B$*, for any set of voters* $\mathcal{H}$*, for any prior* $\mathcal{P}$ *on* $\mathcal{H}$*, for any* n*, with probability at least* $1-\delta$ *over* $\mathbf{S}$*, for all posteriors* $\mathcal{Q}$ *on* $\mathcal{H}$*, for all* $i\in\{1,\ldots,m\}$*, for all distributions* π_i *on* $\boldsymbol{\mathcal{E}}_i$ *independent from a voter* $h\in\mathcal{H}$*, we have*

$$\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}}) \leq \frac{1}{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\frac{1}{2}\left(1-y_ih(x_i+\epsilon)\right) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\tfrac{2\sqrt{m}}{\delta}\right]} \tag{11}$$

$$\leq \overline{A_{\mathbf{S}}}(H_{\mathcal{Q}}) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\mathrm{TV}(\rho_i^h\|\pi_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\tfrac{2\sqrt{m}}{\delta}\right]}, \tag{12}$$

where $\overline{A_{\mathbf{S}}}(H_{\mathcal{Q}}) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\Big[1-\min_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\Big(y_i\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}h(x_i+\epsilon)\Big)\Big]$*, and* $\mathrm{TV}(\rho\|\pi)=\mathop{\mathbb{E}}_{\epsilon\sim\pi}\frac{1}{2}\left|\left[\frac{\rho(\epsilon)}{\pi(\epsilon)}\right]-1\right|$*.*

To minimize the true average-max risk $\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}})$ from Equation (11), we have to minimize a trade-off between $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$ (*i.e.*, how much the posterior weights are close to the prior ones) and the empirical risk $\frac{1}{m}\mathbb{E}_{h\sim\mathcal{Q}}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\frac{1}{2}\left(1-y_ih(x_i+\epsilon)\right)$. However, to compute the empirical risk, the loss for each voter and each perturbation has to be calculated and can be time-consuming. With Equation (12), we propose an alternative, which can be efficiently optimized using $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{h\sim\mathcal{Q}}\mathrm{TV}(\rho_i^h\|\pi_i)$ and the empirical average-max risk $\overline{A_{\mathbf{S}}}(H_{\mathcal{Q}})$. Intuitively, Equation (12) can be seen as a trade-off between the empirical risk, which reflects the robustness of the majority vote, and two penalization terms: the KL term and the TV term. The KL-divergence $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$ controls how much the posterior $\mathcal{Q}$ can differ from the prior ones $\mathcal{P}$. While the TV term $\mathbb{E}_h\,\mathrm{TV}(\rho_i^h\|\pi_i)$ controls the diversity of the voters, *i.e.*, the ability of the voters to be fooled on the same adversarial example. From an algorithmic view, an interesting behavior is that the bound of Equation (12) stands for all distributions π_i on $\boldsymbol{\mathcal{E}}_i$. This suggests that given (x_i,y_i), we want to find π_i minimizing $\mathbb{E}_{h\sim\mathcal{Q}}\mathrm{TV}(\rho_i^h\|\pi_i)$. Ideally,

this term tends to 0 when π_i is close[3] to ρ_i^h and all voters have their loss maximized by the same perturbation $\epsilon \in \boldsymbol{\mathcal{E}}_i$.

To learn a well-performing majority vote, one solution is to minimize the right-hand side of the bounds, meaning that we would like to find a good trade-off between a low empirical risk $\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}})$ or $\overline{A_{\mathbf{S}}}(H_{\mathcal{Q}})$ and a low divergence between the prior weights and the learned posterior ones $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$.

4 Experimental evaluation on differentiable decision trees

In this section, we illustrate the soundness of our framework in the context of differentiable decision trees learning. First of all, we describe our learning procedure designed from our theoretical results.

4.1 From the bounds to an algorithm

We consider a finite voters set $\mathcal{H}$ consisting of differentiable decision trees [Kontschieder et al., 2016] where each $h \in \mathcal{H}$ is parametrized by a weight vector $\boldsymbol{w}^h$. Inspired by Masegosa et al. [2020], we learn the decision trees of $\mathcal{H}$ and a data-dependent prior distribution $\mathcal{P}$ from a first learning set $\mathcal{S}'$ (independent from $\mathcal{S}$); This is a common approach in PAC-Bayes [Parrado-Hernández et al., 2012, Lever et al., 2013, Dziugaite and Roy, 2018, Dziugaite et al., 2021]. Then, the posterior distribution is learned from the second learning set $\mathcal{S}$ by minimizing the bounds. This means we need to minimize the risk and the KL-divergence term. Our two-step learning procedure is summarized in Algorithm 1.

Step 1. Starting from an initial prior $\mathcal{P}_0$ and an initial set of voters $\mathcal{H}_0$, where each voter h is parametrized by a weight vector $\boldsymbol{w}_0^h$, the objective of this step is to construct the hypothesis set $\mathcal{H}$ and the prior distribution $\mathcal{P}$ to give as input to Step 2 for minimizing the bound. To do so, at each epoch t of the Step 1, we learn from $\mathcal{S}'$ an "intermediate" prior $\mathcal{P}_t$ on an "intermediate" hypothesis set $\mathcal{H}_t$ consisting of voters h parametrized by the weights $\boldsymbol{w}_t^h$; Note that the optimization in Line 9 is done with respect to $\boldsymbol{w}_t{=}\{\boldsymbol{w}_t^h\}_{h\in\mathcal{H}_t}$. At each iteration of the optimizer, from Lines 4 to 7, for each (x, y) of the current batch $\mathbb{S}'$, we attack the majority vote $H_{\mathcal{P}_t}$ to obtain a perturbed example $x{+}\epsilon$. Then, in Lines 8 and 9, we perform a forward pass in the majority vote with the perturbed examples and update the weights $\boldsymbol{w}_t$ and the prior $\mathcal{P}_t$ according to the linear loss. To sum up, from Lines 11 to 20 at the end of Step 1, the prior $\mathcal{P}$ and the hypothesis set $\mathcal{H}$ constructed for Step 2 are the ones associated to the best epoch $t^* \in \{1, \ldots, T'\}$ that permits to minimize $\overline{R_{\mathcal{S}_t}}(H_{\mathcal{P}_t})$, where $\mathcal{S}_t{=}\{\texttt{attack}(x, y) \,|\, (x, y) \in \mathcal{S}\}$ is the perturbed set obtained by attacking the majority vote $H_{\mathcal{P}_t}$.

Step 2. Starting from the prior $\mathcal{P}$ on $\mathcal{H}$ and the learning set $\mathcal{S}$, we perform the same process as in Step 1 except that the considered objective function corresponds to the desired bound to optimize (Line 30, denoted $\texttt{B}(\cdot)$). For the sake of readability, we deferred in Appendix G the definition of $\texttt{B}(\cdot)$ for Equations (9) and (12). Note that the "intermediate" priors do not depend on $\mathcal{S}$, since they are learned from $\mathcal{S}'$: the bounds are then valid.

4.2 Experiments[4]

In this section, we empirically illustrate that our PAC-Bayesian framework for adversarial robustness is able to provide generalization guarantees with non-vacuous bounds for the adversarial risk.

Setting. We stand in a white-box setting meaning that the attacker knows the voters set $\mathcal{H}$, the prior distribution $\mathcal{P}$, and the posterior one $\mathcal{Q}$. We empirically study 2 attacks with the ℓ_2-norm and ℓ_∞-norm: the Projected Gradient Descent (PGD, Madry et al. [2018]) and the iterative version of FGSM (IFGSM, Kurakin et al. [2017]). We fix the number of iterations at $k{=}20$ and the step size at $\frac{b}{k}$ for PGD and IFGSM (where $b{=}1$ for ℓ_2-norm and $b{=}0.1$ for ℓ_∞-norm). One specificity of our setting is that we deal with the perturbation distribution $\omega_{(x,y)}$. We propose PGD_U and IFGSM_U, two variants of PGD and IFGSM. To attack an example with PGD_U or IFGSM_U we proceed with the following steps: *(1)* We attack the prior majority vote $H_{\mathcal{P}}$ with the attack PGD or IFGSM: we will obtain a first perturbation ϵ' ; *(2)* We sample n uniform noises $\eta_1, \ldots, \eta_n$ between -10^{-2} and $+10^{-2}$; *(3)* We set

[3]Note that, since ρ_i^h is a Dirac distribution, we have $\mathbb{E}_h\, \mathrm{TV}(\rho_i^h\|\pi_i){=}\frac{1}{2}\left[1{-}\,\mathbb{E}_h\, \pi_i(\epsilon_h^*){+}\,\mathbb{E}_h \sum_{\epsilon\neq\epsilon_h^*} \pi_i(\epsilon)\right]$, with $\epsilon_h^* = \mathrm{argmax}_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\, \frac{1}{2}\big[1{-}y_i h(x_i{+}\epsilon)\big]$.

[4]The source code is available at `https://github.com/paulviallard/NeurIPS21-PB-Robustness`.

Algorithm 1 Average Adversarial Training with Guarantee

Require: $\mathcal{S}, \mathcal{S}'$: disjoint learning sets – T, T': number of epochs – $\mathcal{P}_0$: initial prior – $\mathcal{H}_0$ (with $\boldsymbol{w}_0$): initial hypothesis set – $\texttt{attack}()$: the attack function – $\texttt{B}(\cdot)$: the objective function associated to a bound

Step 1 – prior and voters' set construction

for t from 1 to T' **do**
  $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1}$ and $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1}$ ($\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1}$)
  for all batches $\mathbb{S}'$ (from $\mathcal{S}'$) **do**
    for all $(x, y) \in \mathbb{S}'$ **do**
      $(x{+}\epsilon, y) \leftarrow \texttt{attack}(x, y)$
      $\mathbb{S}' \leftarrow (\mathbb{S}' \setminus \{(x, y)\}) \cup \{(x{+}\epsilon, y)\}$
    end for
    Update $\mathcal{P}_t$ with $\nabla_{\mathcal{P}_t} \overline{R_{\mathbb{S}'}}(H_{\mathcal{P}_t})$
    Update $\boldsymbol{w}_t$ with $\nabla_{\boldsymbol{w}_t} \overline{R_{\mathbb{S}'}}(H_{\mathcal{P}_t})$
  end for
  $\mathcal{S}_t \leftarrow \emptyset$
  for all $(x, y) \in \mathcal{S}$ **do**
    $(x{+}\epsilon, y) \leftarrow \texttt{attack}(x, y)$
    $\mathcal{S}_t \leftarrow \mathcal{S}_t \cup \{(x{+}\epsilon, y)\}$
  end for
  $t^* \leftarrow \operatorname{argmin}_{t' \in \{1,\dots,t\}} \overline{R_{\mathcal{S}_{t'}}}(H_{\mathcal{P}_{t'}})$
  $\mathcal{P} \leftarrow \mathcal{P}_{t^*}$
  $\mathcal{H} \leftarrow \mathcal{H}_{t^*}$
end for
return $(\mathcal{P}, \mathcal{H})$

Step 2 – bound minimization

$(\mathcal{P}, \mathcal{H}) \leftarrow$ Output of **Step 1**
$\mathcal{Q}_0 \leftarrow \mathcal{P}$
for t from 1 to T **do**
  for all batches $\mathbb{S}$ (from $\mathcal{S}$) **do**
    $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1}$
    for all $(x, y) \in \mathbb{S}$ **do**
      $(x{+}\epsilon, y) \leftarrow \texttt{attack}(x, y)$
      $\mathbb{S} \leftarrow (\mathbb{S} \setminus \{(x, y)\}) \cup \{(x{+}\epsilon, y)\}$
    end for
    Update $\mathcal{Q}_t$ with $\nabla_{\mathcal{Q}_t} \texttt{B}_{\mathbb{S}}(H_{\mathcal{Q}_t})$
  end for
  $\mathcal{S}_t \leftarrow \emptyset$
  for all $(x, y) \in \mathcal{S}$ **do**
    $(x{+}\epsilon, y) \leftarrow \texttt{attack}(x, y)$
    $\mathcal{S}_t \leftarrow \mathcal{S}_t \cup \{(x{+}\epsilon, y)\}$
  end for
  $t^* \leftarrow \operatorname{argmin}_{t' \in \{1,\dots,t\}} \texttt{B}_{\mathcal{S}_{t'}}(H_{\mathcal{Q}_{t'}})$
  $\mathcal{Q} \leftarrow \mathcal{Q}_{t^*}$
end for
return $(\mathcal{Q}, \mathcal{H})$

the i-th perturbation as $\epsilon_i = \epsilon' + \eta_i$. Note that, for PGD_U and IFGSM_U, after one attack we end up with $n{=}100$ perturbed examples. We set $n{=}1$ when these attacks are used as a defense mechanism in Algorithm 1. Indeed since the adversarial training is iterative, we do not need to sample numerous perturbations for each example: we sample a new perturbation each time the example is forwarded through the decision trees. We also consider a naive defense referred to as UNIF that only adds a noise uniformly such that the ℓ_p-norm of the added noise is lower than b.
We study the following scenarios of defense/attack. These scenarios correspond to all the pairs (Defense, Attack) belonging to the set $\{\text{—}, \text{UNIF}, \text{PGD}, \text{IFGSM}\} \times \{\text{—}, \text{PGD}, \text{IFGSM}\}$ for the baseline, and $\{\text{—}, \text{UNIF}, \text{PGD}_\text{U}, \text{IFGSM}_\text{U}\} \times \{\text{—}, \text{PGD}_\text{U}, \text{IFGSM}_\text{U}\}$, where "—" means that we do not defend, *i.e.*, the attack returns the original example (note that PGD_U and IFGSM_U when "Attack without U" refers to PGD and IFGSM for computing the classical adversarial risk $R^{\texttt{ROB}}()$).

Datasets and algorithm description. We perform our experiment on six binary classification tasks from MNIST [LeCun et al., 1998] (1vs7, 4vs9, 5vs6) and Fashion MNIST [Xiao et al., 2017] (Coat vs Shirt, Sandal vs Ankle Boot, Top vs Pullover). We decompose the learning set into two disjoint subsets $\mathcal{S}'$ of around $7,000$ examples (to learn the prior and the voters) and $\mathcal{S}$ of exactly $5,000$ examples (to learn the posterior). We keep as test set $\mathcal{T}$ the original test set that contains around $2,000$ examples. Moreover, we need a perturbed test set, denoted by $\mathbf{T}$, to compute our averaged(-max) adversarial risks. Depending on the scenario, $\mathbf{T}$ is constructed from $\mathcal{T}$ by attacking the prior model $H_{\mathcal{P}}$ with PGD_U or IFGSM_U with $n{=}100$ (more details are given in Appendix). We run our Algorithm 1 for Equation (9) (Theorem 6), respectively Equation (12) (Theorem 7), and we compute our risk $R_{\mathbf{T}}(H_{\mathcal{Q}})$, respectively $A_{\mathbf{T}}(H_{\mathcal{Q}})$, the bound value and the usual adversarial risk associated to the model learned $R^{\texttt{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$. Note that, during the evaluation of the bounds, we have to compute our relaxed adversarial risks $R_{\mathbf{S}}(H_{\mathcal{Q}})$ and $A_{\mathbf{S}}(H_{\mathcal{Q}})$ on $\mathcal{S}$. For Step 1, the initial prior P_0 is fixed to the uniform distribution, the initial set of voters $\mathcal{H}_0$ is constructed with weights initialized with Xavier Initializer [Glorot and Bengio, 2010] and bias initialized at 0 (more details are given in Appendix). During Step 2, to optimize the bound, we fix the confidence parameter $\delta{=}0.05$, and we consider as the set of voters $\mathcal{H}$ two settings: $\mathcal{H}$ as it is output by Step 1, and the set $\mathcal{H}^{\text{SIGN}} = \{h'(\cdot) = \text{sign}(h(\cdot)) \mid h \in \mathcal{H}\}$ for which the theoretical results are still valid (we will see that in this latter situation we are able to better minimize the TV term of Theorem 7). For the two steps, we use Adam optimizer [Kingma and Ba, 2015] for $T{=}T'{=}20$ epochs with a learning rate at 10^{-2} and a batch size at 64.

Table 1: Test risks and bounds for **MNIST 1vs7** with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in *italic* the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (*i.e.*, we did *not* sample from the uniform distribution). Since Eq. (12) upper-bounds Eq. (11) thanks to the TV term, we compute the two bound values of Theorem 7.

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_Q)$		$R_{\mathbf{T}}(H_Q)$		Th. 6		$R^{\text{ROB}}_{\mathcal{T}}(H_Q)$		$A_{\mathbf{T}}(H_Q)$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.005*	*.005*	.005	.005	**.017**	.019	*.005*	*.005*	.005	.005	**.099**	0.100	**.099**	.100
—	PGD_U	***.245***	*.255*	**.263**	.276	.577	**.448**	*.315*	***.313***	**.325**	.326	**.801**	1.667	.684	**.515**
—	IFGSM_U	***.084***	*.086*	**.066**	.080	**.170**	.185	*.117*	***.113***	**.106**	.110	**.356**	1.431	.286	**.251**
UNIF	—	*.005*	*.005*	.005	.005	**.018**	.019	*.005*	*.005*	.005	.005	**.099**	0.100	**.099**	.100
UNIF	PGD_U	*.151*	***.146***	**.151**	.158	.355	**.292**	*.183*	***.178***	.190	**.189**	**.531**	1.620	.454	**.355**
UNIF	IFGSM_U	*.063*	***.061***	**.031**	.035	**.088**	.114	*.071*	***.070***	.056	**.054**	**.248**	1.405	.200	**.186**
PGD_U	—	***.006***	*.007*	**.006**	.007	**.023**	.024	***.006***	*.007*	**.006**	.007	**.102**	0.103	**.102**	.103
PGD_U	PGD_U	***.028***	*.030*	**.021**	.025	.065	**.064**	***.028***	*.029*	**.025**	.028	**.143**	1.389	.137	**.136**
PGD_U	IFGSM_U	***.021***	*.022*	**.013**	.016	**.043**	.045	*.022*	*.022*	**.018**	.019	**.125**	1.362	.121	**.119**
IFGSM_U	—	***.006***	*.007*	**.006**	.007	**.019**	.021	***.006***	*.007*	**.006**	.007	**.100**	0.102	**.100**	.102
IFGSM_U	PGD_U	***.040***	*.041*	**.033**	.035	**.086**	.094	*.040*	***.039***	.040	**.038**	**.184**	1.368	.166	**.163**
IFGSM_U	IFGSM_U	***.021***	*.022*	**.013**	.014	**.039**	.049	***.021***	*.022*	**.018**	.021	**.131**	1.329	**.122**	.123

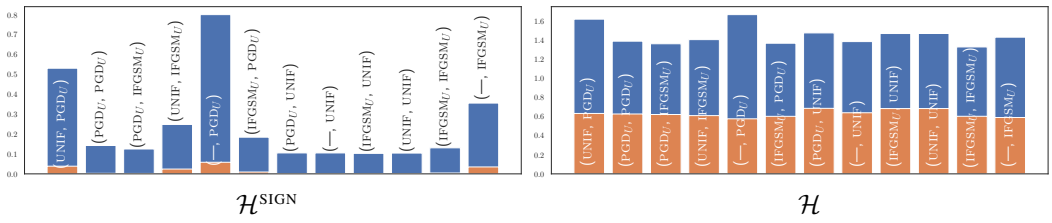

Figure 1: Visualization of the impact of the TV term in Equation (12). The left, respectively the right, bar plot show the bounds for the set of voters $\mathcal{H}^{\text{SIGN}}$, respectively $\mathcal{H}$. We plot the bounds for all the scenarios of Table 1 that use the TV distance, *i.e.*, all except the pairs $(\cdot, —)$. In orange we represent the value of the TV term while in blue we represent all the remaining terms of the bound.

Analysis of the results. For the sake of readability, we exhibit the detailed results for one task (MNIST:1vs7) and all the pairs (Defense,Attack) with ℓ_2-norm in Table 1, and we report in Figure 1 the influence of the TV term in the bound of Theorem 7 (Equation (12)). The detailed results on the other tasks are reported in Appendix; We provide in Figure 2 an overview of the results we obtained on all the tasks for the pairs (Defense,Attack) where "Defense=Attack" and with $\mathcal{H}^{\text{SIGN}}$.

First of all, from Table 1 the bounds of Theorem 6 are tighter than the ones of Theorem 7: this is an expected result since we showed that the averaged-max adversarial risk $A_{\mathbf{D}^n}(H_Q)$ is more pessimistic than its averaged counterpart $R_{\mathbf{D}}(H_Q)$. Note that the bound values of Equation (11) are tighter than the ones of Equation (12). This is expected since Equation (11) is a lower bound on Equation (12).

Second, the bounds with $\mathcal{H}^{\text{SIGN}}$ are all informative (lower than 1) and give insightful guarantees for our models. For Theorem 7 (Equation (12)) with $\mathcal{H}$, while the risks are comparable to the risks obtained with $\mathcal{H}^{\text{SIGN}}$, the bound values are greater than 1, meaning that we have no more guarantee on the model learned. As we can observe in Figure 1, this is due to the TV term involved in the bound. Considering $\mathcal{H}^{\text{SIGN}}$ when optimizing $A(\cdot)$ helps to control the TV term. Even if the bounds are non-vacuous for Theorem 6 with $\mathcal{H}$, the best models with the best guarantees are obtained with $\mathcal{H}^{\text{SIGN}}$. This is confirmed by the columns $R^{\text{ROB}}_{\mathcal{T}}(H_Q)$ that are always worse than $R_{\mathbf{T}}(H_Q)$ and mostly worse than $A_{\mathbf{T}}(H_Q)$ with $\mathcal{H}^{\text{SIGN}}$. The performance obtained with $\mathcal{H}^{\text{SIGN}}$ can be explained by the fact that the sign "saturates" the output of the voters which makes the majority vote more robust to noises. Thus, we focus the rest of the analysis on results obtained with $\mathcal{H}^{\text{SIGN}}$.

Third, we observe that the naive defense UNIF is able to improve the risks $R_{\mathbf{T}}(H_Q)$ and $A_{\mathbf{T}}(H_Q)$, but the improvement with the defenses based on PGD_U and IFGSM_U is much more significant specifically against a PGD_U attack (up to 13 times better). We observe the same phenomenon for both bounds

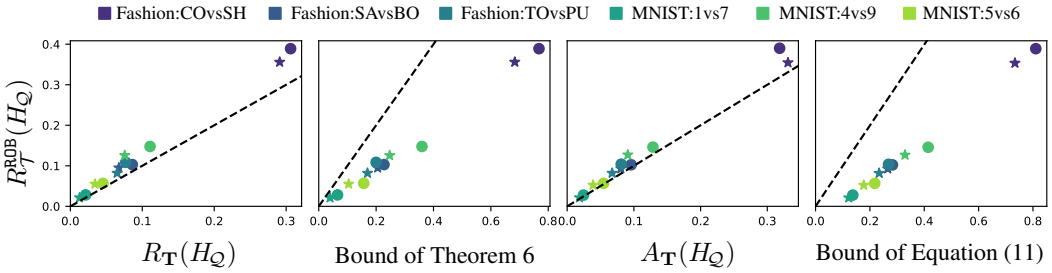

Figure 2: Visualization of the risk and bound values when "Defense=Attack" when the set of voters is $\mathcal{H}^{\text{SIGN}}$. Results obtained with the PGD_{U}, respectively IFGSM_{U}, defense are represented by a star ★, respectively a circle ● *(reminder:* $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$ *is computed with a* PGD*, respectively* IFGSM*, attack).* The dashed line corresponds to bisecting line $y{=}x$. For $R_{\mathbf{T}}(H_{\mathcal{Q}})$ and $A_{\mathbf{T}}(H_{\mathcal{Q}})$, the closer the datasets are to the bisecting line, the more accurate our relaxed risk is compared to the classical adversarial risk $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$. For the bounds, the closer the datasets are to the bisecting line, the tighter the bound.

(Theorems 6 and 7). This is an interesting fact because this behavior confirms that we are able to learn models that are robust against the attacks tested with theoretical guarantees.

Lastly, from Figure 2 and Table 1, it is important to notice that the gap between the classical risk and our relaxed risks is small, meaning that our relaxation are not too optimistic. Despite the pessimism of the classical risk $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$, it remains consistent with our bounds, *i.e.*, it is lower than the bounds. In other words, in addition to giving upper bounds for our risks $R_{\mathbf{T}}(H_{\mathcal{Q}})$ and $A_{\mathbf{T}}(H_{\mathcal{Q}})$, our bounds give non-vacuous guarantees on the classical risks $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$.

5 Conclusion

To the best of our knowledge, our work is the first one that studies from a general standpoint adversarial robustness through the lens of the PAC-Bayesian framework. We have started by formalizing a new adversarial robustness setting (for binary classification) specialized for models that can be expressed as a weighted majority vote; we referred to this setting as Adversarially Robust PAC-Bayes. This formulation allowed us to derive PAC-Bayesian generalization bounds on the adversarial risk of general majority votes. We illustrated the usefulness of this setting on the training of (differentiable) decision trees. Our contribution is mainly theoretical and it does not appear to directly lead to potentially negative social impact.

This work gives rise to many interesting questions and lines of future research. Some perspectives will focus on extending our results to other classification settings such as multiclass or multilabel. Another line of research could focus on taking advantage of other tools of the PAC-Bayesian literature. Among them, we can make use of other bounds on the risk of the majority vote that take into consideration the diversity between the individual voters; For example, the C-bound [Lacasse et al., 2006], or more recently the tandem loss [Masegosa et al., 2020]. Another very recent PAC-Bayesian bound for majority votes that needs investigation in the case of adversarial robustness is the one proposed by Zantedeschi et al. [2021] that has the advantage to be directly optimizable with the 0-1 loss. Last but not least, in real-life applications, one often wants to combine different input sources (from different sensors, cameras, etc). Being able to combine these sources in an effective way is then a key issue. We believe that our new adversarial robustness setting can offer theoretical guarantees and well-founded algorithms when the model we learn is expressed as a majority vote, whether for ensemble methods with weak voters [*e.g.* Roy et al., 2011, Lorenzen et al., 2019], or for fusion of classifiers [*e.g.* Morvant et al., 2014], or for multimodal/multiview learning [*e.g.* Sun et al., 2017, Goyal et al., 2019].

Acknowledgments and Disclosure of Funding

This work was partially funded supported by the French Project APRIORI ANR-18-CE23-0015. G. Vidot is supported by the ANRT with the convention "CIFRE" N°2019/0507. We also thank all anonymous reviewers for their constructive comments, and the time they took to review our work.

References

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *ICML*, 2019.

Gintare Karolina Dziugaite and Daniel Roy. Data-dependent PAC-Bayes priors via differential privacy. In *NeurIPS*, 2018.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In *AISTATS*, 2021.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable Adversarial Training via Spectral Normalization. In *ICLR*, 2019.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation. In *IEEE Symposium on Security and Privacy*, 2018.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. *JMLR*, 2015.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015.

Anil Goyal, Emilie Morvant, Pascal Germain, and Massih-Reza Amini. Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters. *Neurocomputing*, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019.

Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety Verification of Deep Neural Networks. In *CAV*, 2017.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *CAV*, 2017.

Justin Khim and Po-Ling Loh. Adversarial Risk Bounds for Binary Classification via Function Transformation. *CoRR*, 2018.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulò. Deep Neural Decision Forests. In *IJCAI*, 2016.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. In *ICLR*, 2017.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. In *NIPS*, 2006.

Yann LeCun, Corinna Cortes, and Christopher Burges. THE MNIST DATASET of handwritten digits, 1998. URL `http://yann.lecun.com/exdb/mnist/`.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 2013.

Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018.

Andrés Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second Order PAC-Bayesian Bounds for the Weighted Majority Vote. In *NeurIPS*, 2020.

David McAllester. Some PAC-Bayesian Theorems. In *COLT*, 1998.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC Classes are Adversarially Robustly Learnable, but Only Improperly. In *COLT*, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing Adversarially Robust Learning to Non-Robust PAC Learning. In *NeurIPS*, 2020.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority Vote of Diverse Classifiers for Late Fusion. In *S+SSPR*, 2014.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, 2019.

Yuki Ohnishi and Jean Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation. In *AISTATS*, 2021.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. In *IEEE EuroS&P*, 2016.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *JMLR*, 2012.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. *JMLR*, 2010.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds. In *NeurIPS*, 2018.

Kui Ren, Tianhang Zheng, and Xue Liu. Adversarial Attacks and Defenses in Deep Learning. *Engineering*, 2020.

Jean-Francis Roy, François Laviolette, and Mario Marchand. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. In *ICML*, 2011.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In *NeurIPS*, 2019.

Edward Scheinerman and Daniel Ullman. *Fractional Graph Theory: A Rational Approach to the Theory of Graphs*. Courier Corporation, 2011.

Matthias Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *JMLR*, 2002.

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. In *COLT*, 1997.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. Boosting Robustness Certification of Neural Networks. In *ICLR*, 2019.

Shiliang Sun, John Shawe-Taylor, and Liang Mao. PAC-Bayes analysis of multi-view learning. *Inf. Fusion*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks. In *NeurIPS*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, 2017.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher Complexity for Adversarially Robust Generalization. In *ICML*, 2019.

Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient Defenses Against Adversarial Attacks. In *ACM Workshop on Artificial Intelligence and Security, AISec@CCS*, 2017.

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. In *NeurIPS*, 2021.

A PAC-Bayes Analysis of Adversarial Robustness
Supplementary Material

The supplementary material is structured as follows. The sections from A to E are devoted to our proofs. We give details on Algorithm 1 and on the computation of the bounds in Section G. We discuss, in Section F, the validity of the bound when we select a prior with $\mathcal{S}$ and have a distribution on perturbations depending on this selected prior. We introduce, in Section H, the voters that we use in our majority vote. Finally, we present additional experiments in Section I.

A Proof of Proposition 3

Proposition 3. *For any distribution* $\mathbf{D}$ *on* $(X\times Y)\times B$*, for any distribution* $\mathcal{Q}$ *on* $\mathcal{H}$*, for any* $(n, n') \in \mathbb{N}^2$*, with* $n \geq n' \geq 1$*, we have*

$$R_{\mathbf{D}}(H_{\mathcal{Q}}) \;\leq\; A_{\mathbf{D}^{n'}}(H_{\mathcal{Q}}) \;\leq\; A_{\mathbf{D}^{n}}(H_{\mathcal{Q}}) \;\leq\; R_D^{\texttt{ROB}}(H_{\mathcal{Q}}). \tag{13}$$

For any distribution $\mathbf{D}$ *on* $(X\times Y)\times B$*, for any distribution* $\mathcal{Q}$ *on* $\mathcal{H}$*, for any* $(n, n') \in \mathbb{N}^2$*, with* $n \geq n' \geq 1$*, we have*

$$R_{\mathbf{D}}(H_{\mathcal{Q}}) \leq A_{\mathbf{D}^{n'}}(H_{\mathcal{Q}}) \leq A_{\mathbf{D}^{n}}(H_{\mathcal{Q}}) \leq R_D^{\texttt{ROB}}(H_{\mathcal{Q}}).$$

Proof. First, we prove $A_{\mathbf{D}^1}(H_{\mathcal{Q}})=R_{\mathbf{D}}(H_{\mathcal{Q}})$. We have

$$\begin{aligned}
A_{\mathbf{D}^1}(H_{\mathcal{Q}}) &= 1 - \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^1} \left(\forall \epsilon \in \boldsymbol{\mathcal{E}}, H_{\mathcal{Q}}(x+\epsilon) = y\right) \\
&= 1 - \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^1} \left(\forall \epsilon \in \{\epsilon_1\}, H_{\mathcal{Q}}(x+\epsilon) = y\right) \\
&= 1 - \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^1} \left(H_{\mathcal{Q}}(x+\epsilon_1) = y\right) = R_{\mathbf{D}}(H_{\mathcal{Q}}).
\end{aligned}$$

Then, we prove the inequality $A_{\mathbf{D}^{n'}}(H_{\mathcal{Q}}) \leq A_{\mathbf{D}^{n}}(H_{\mathcal{Q}})$ from the fact that the indicator function $\mathbf{I}(\cdot)$ is upper-bounded by 1. Indeed, from Definition 2 we have

$$\begin{aligned}
1 - A_{\mathbf{D}^{n}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x,y)\sim D} \; \mathop{\mathbb{E}}_{\boldsymbol{\mathcal{E}}\sim\omega_{(x,y)}^{n}} \mathbf{I}\left(\forall \epsilon \in \boldsymbol{\mathcal{E}}, H_{\mathcal{Q}}(x+\epsilon) = y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim D} \left[\prod_{i=1}^{n} \mathop{\mathbb{E}}_{\epsilon_i\sim\omega_{(x,y)}} \mathbf{I}\left(H_{\mathcal{Q}}(x+\epsilon_i) = y\right) \right] \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim D} \left[\prod_{i=1}^{n'} \mathop{\mathbb{E}}_{\epsilon_i\sim\omega_{(x,y)}} \mathbf{I}\left(H_{\mathcal{Q}}(x+\epsilon_i) = y\right) \right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim D} \; \mathop{\mathbb{E}}_{\boldsymbol{\mathcal{E}}'\sim\omega_{(x,y)}^{n'}} \mathbf{I}\left(\forall \epsilon \in \boldsymbol{\mathcal{E}}', H_{\mathcal{Q}}(x+\epsilon) = y\right) \\
&= 1 - A_{\mathbf{D}^{n'}}(H_{\mathcal{Q}}).
\end{aligned}$$

Lastly, to prove the rightmost inequality, we have to use the fact that the expectation over the set B is bounded by the maximum over the set B. We have

$$\begin{aligned}
A_{\mathbf{D}^{n}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x,y)\sim D} \; \mathop{\mathbb{E}}_{\epsilon_1\sim\omega_{(x,y)}} \cdots \mathop{\mathbb{E}}_{\epsilon_n\sim\omega_{(x,y)}} \mathbf{I}\left(\exists \epsilon \in \{\epsilon_1, \ldots, \epsilon_n\}, H_{\mathcal{Q}}(x+\epsilon) \neq y\right) \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim D} \max_{\epsilon_1\in B} \ldots \max_{\epsilon_n\in B} \mathbf{I}\left(\exists \epsilon \in \{\epsilon_1, \ldots, \epsilon_n\}, H_{\mathcal{Q}}(x+\epsilon) \neq y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim D} \max_{\epsilon_1\in B} \ldots \max_{\epsilon_{n-1}\in B} \mathbf{I}\left(\exists \epsilon \in \{\epsilon_1, \ldots, \epsilon^*\}, H_{\mathcal{Q}}(x+\epsilon) \neq y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim D} \mathbf{I}\left(H_{\mathcal{Q}}(x+\epsilon^*) \neq y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim D} \max_{\epsilon\in B} \mathbf{I}\left(H_{\mathcal{Q}}(x+\epsilon) \neq y\right) \;=\; R_D^{\texttt{ROB}}(H_{\mathcal{Q}}).
\end{aligned}$$

Merging the three equations proves the claim. □

B Proof of Proposition 4

In this section, we provide the proof of Proposition 4 that relies on Lemmas 8 and 9 which are also described and proved. Lemma 8 shows that $R_{\mathbf{D}}(H_{\mathcal{Q}})$ is equivalent to $R_{\Delta}(H_{\mathcal{Q}})$.

Lemma 8. *For any distribution* $\mathbf{D}$ *on* $(X\times Y)\times B$ *and its associated distribution* Δ*, for any posterior* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$R_{\mathbf{D}}(H_{\mathcal{Q}}) = \Pr_{(x+\epsilon,y)\sim\Delta}[H_{\mathcal{Q}}(x+\epsilon)\neq y] = R_{\Delta}(H_{\mathcal{Q}}).$$

Proof. Starting from the averaged adversarial risk $R_{\mathbf{D}}(H_{\mathcal{Q}}) = \mathbb{E}_{((x,y),\epsilon)\sim\mathbf{D}}\,\mathbf{I}\left[H_{\mathcal{Q}}(x+\epsilon)\neq y\right]$, we have

$$\begin{aligned} R_{\mathbf{D}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta} \tfrac{1}{\Delta(x'+\epsilon',y')}\left[\Pr_{((x,y),\epsilon)\sim\mathbf{D}}[H_{\mathcal{Q}}(x+\epsilon)\neq y, x'+\epsilon'=x+\epsilon, y'=y]\right] \\ &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta} \tfrac{1}{\Delta(x'+\epsilon',y')}\left[\mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}}\mathbf{I}[H_{\mathcal{Q}}(x+\epsilon)\neq y]\,\mathbf{I}[x'+\epsilon'=x+\epsilon, y'=y]\right]. \end{aligned}$$

In other words, the double expectation only rearranges the terms of the original expectation: given an example $(x'+\epsilon',y')$, we gather probabilities such that $H_{\mathcal{Q}}(x+\epsilon)\neq y$ with $(x+\epsilon,y)=(x'+\epsilon',y')$ in the inner expectation, while integrating over all couple $(x'+\epsilon', y') \in X\times Y$ in the outer expectation. Then, from the fact that when $x'+\epsilon'=x+\epsilon$ and $y'=y$, $\mathbf{I}[H_{\mathcal{Q}}(x+\epsilon)\neq y] = \mathbf{I}[H_{\mathcal{Q}}(x'+\epsilon')\neq y']$, we have

$$\begin{aligned} R_{\mathbf{D}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta} \tfrac{1}{\Delta(x'+\epsilon',y')}\left[\mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}}\mathbf{I}[H_{\mathcal{Q}}(x'+\epsilon')\neq y']\mathbf{I}[x'+\epsilon'=x+\epsilon, y'=y]\right] \\ &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta} \tfrac{1}{\Delta(x'+\epsilon',y')}\left[\mathbf{I}[H_{\mathcal{Q}}(x'+\epsilon')\neq y'] \mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}}\mathbf{I}[x'+\epsilon'=x+\epsilon, y'=y]\right]. \end{aligned}$$

Finally, by definition of $\Delta(x'+\epsilon',y')$, we can deduce that

$$\begin{aligned} R_{\mathbf{D}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta} \tfrac{1}{\Delta(x'+\epsilon',y')}\left[\mathbf{I}[H_{\mathcal{Q}}(x'+\epsilon')\neq y']\,\Delta(x'+\epsilon',y')\right] \\ &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Delta}\mathbf{I}[H_{\mathcal{Q}}(x'+\epsilon')\neq y'] = R_{\Delta}(H_{\mathcal{Q}}). \end{aligned}$$

□

Similarly, Lemma 9 shows that $R_D^{\mathtt{ROB}}(H_{\mathcal{Q}})$ is equivalent to $R_{\Pi}(H_{\mathcal{Q}})$.

Lemma 9. *For any distribution* D *on* $X\times Y$ *and its associated distribution* Π*, for any posterior* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$R_D^{\mathtt{ROB}}(H_{\mathcal{Q}}) = \Pr_{(x+\epsilon,y)\sim\Pi}[H_{\mathcal{Q}}(x+\epsilon)\neq y] = R_{\Pi}(H_{\mathcal{Q}}).$$

Proof. The proof is similar to the one of Lemma 8. Indeed, starting from the definition of $R_D^{\mathtt{ROB}}(H_{\mathcal{Q}}) = \mathbb{E}_{(x,y)\sim D}\,\mathbf{I}[H_{\mathcal{Q}}(x+\epsilon^*(x,y))\neq y]$, we have

$$\begin{aligned} R_D^{\mathtt{ROB}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Pi} \tfrac{1}{\Pi(x'+\epsilon',y')}\left[\mathop{\mathbb{E}}_{(x,y)\sim D}\mathbf{I}\left[H_{\mathcal{Q}}(x+\epsilon^*(x,y))\neq y\right]\mathbf{I}[x'+\epsilon'=x+\epsilon^*(x,y), y'=y]\right] \\ &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Pi} \tfrac{1}{\Pi(x'+\epsilon',y')}\left[\mathop{\mathbb{E}}_{(x,y)\sim D}\mathbf{I}\left[H_{\mathcal{Q}}(x'+\epsilon')\neq y'\right]\mathbf{I}[x'+\epsilon'=x+\epsilon^*(x,y), y'=y]\right]. \end{aligned}$$

Finally, by definition of $\Pi(x'+\epsilon', y')$, we can deduce that

$$\begin{aligned} R_D^{\mathtt{ROB}}(H_{\mathcal{Q}}) &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Pi} \tfrac{1}{\Pi(x'+\epsilon',y')}\left[\mathbf{I}\left[H_{\mathcal{Q}}(x'+\epsilon')\neq y'\right]\Pi(x'+\epsilon',y')\right] \\ &= \mathop{\mathbb{E}}_{(x'+\epsilon',y')\sim\Pi}\mathbf{I}\left[H_{\mathcal{Q}}(x'+\epsilon')\neq y'\right] = R_{\Pi}(H_{\mathcal{Q}}). \end{aligned}$$

□

We can now prove Proposition 4.

Proposition 4. *For any distribution* $\mathbf{D}$ *on* $(X \times Y) \times B$*, for any distribution* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$R_D^{\texttt{ROB}}(H_{\mathcal{Q}}) - \mathrm{TV}(\Pi \| \Delta) \;\leq\; R_{\mathbf{D}}(H_{\mathcal{Q}}).$$

Proof. From Lemmas 8 and 9, we have

$$R_{\mathbf{D}}(H_{\mathcal{Q}}) = R_{\Delta}(H_{\mathcal{Q}}), \quad \text{and} \quad R_D^{\texttt{ROB}}(H_{\mathcal{Q}}) = R_{\Pi}(H_{\mathcal{Q}}).$$

Then, we apply Lemma 4 of Ohnishi and Honorio [2021], we have

$$R_{\Pi}(H_{\mathcal{Q}}) \leq \mathrm{TV}(\Pi \| \Delta) + R_{\Delta}(H_{\mathcal{Q}}) \quad \iff \quad R_D^{\texttt{ROB}}(H_{\mathcal{Q}}) \leq \mathrm{TV}(\Pi \| \Delta) + R_{\mathbf{D}}(H_{\mathcal{Q}}).$$

□

C Proof of Theorem 5

Theorem 5. *For any distributions* $\mathbf{D}$ *on* $(X \times Y) \times B$ *and* $\mathcal{Q}$ *on* $\mathcal{H}$*, for any* $n > 1$*, we have*

$$R_{\mathbf{D}}(H_{\mathcal{Q}}) \leq 2\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}}), \qquad \text{and} \qquad A_{\mathbf{D}^n}(H_{\mathcal{Q}}) \leq 2\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}}).$$

Proof. By the definition of the majority vote, we have

$$\begin{aligned}
\frac{1}{2} R_{\mathbf{D}}(H_{\mathcal{Q}}) &= \frac{1}{2} \Pr_{((x,y),\epsilon)\sim\mathbf{D}} \left(y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \leq 0 \right) \\
&= \frac{1}{2} \Pr_{((x,y),\epsilon)\sim\mathbf{D}} \left(1 - y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \geq 1 \right) \\
&\leq \mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}} \frac{1}{2} \Big[1 - y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \Big] \qquad \text{(Markov's ineq. on } y \,\mathbb{E}\, h(x+\epsilon)\text{)}.
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\frac{1}{2} A_{\mathbf{D}^n}(H_{\mathcal{Q}}) &= \frac{1}{2} \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^n} \left(\exists \epsilon \in \boldsymbol{\mathcal{E}}, y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \leq 0 \right) \\
&= \frac{1}{2} \Pr_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^n} \left(\min_{\epsilon\in\boldsymbol{\mathcal{E}}} \left(y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \right) \leq 0 \right) \\
&= \frac{1}{2} \Pr_{((x,y),\epsilon)\sim\mathbf{D}} \left(1 - \min_{\epsilon\in\boldsymbol{\mathcal{E}}} \left(y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \right) \geq 1 \right) \\
&\leq \mathop{\mathbb{E}}_{((x,y),\epsilon)\sim\mathbf{D}} \frac{1}{2} \Big[1 - \min_{\epsilon\in\boldsymbol{\mathcal{E}}} \left(y \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x+\epsilon) \right) \Big] \quad \text{(Markov's ineq. on } \min y \,\mathbb{E}\, h(x+\epsilon)\text{)}.
\end{aligned}$$

□

D Proof of Theorem 6

Theorem 6. *For any distribution* $\mathbf{D}$ *on* $(X \times Y) \times B$*, for any set of voters* $\mathcal{H}$*, for any prior* $\mathcal{P}$ *on* $\mathcal{H}$*, for any* n*, with probability at least* $1-\delta$ *over* $\mathbf{S}$*, for all posteriors* $\mathcal{Q}$ *on* $\mathcal{H}$*, we have*

$$\mathrm{kl}(\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) \| \overline{R_{\mathbf{D}}}(H_{\mathcal{Q}})) \leq \frac{1}{m} \left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln \frac{m+1}{\delta} \right], \tag{14}$$

$$\text{and} \quad \overline{R_{\mathbf{D}}}(H_{\mathcal{Q}}) \leq \overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) + \sqrt{\frac{1}{2m} \left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln \frac{m+1}{\delta} \right]}, \tag{15}$$

$$\text{where} \quad \overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{2} \left[1 - y_i \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x_i + \epsilon_j^i) \right],$$

$\mathrm{kl}(a\|b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$*, and* $\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) = \mathop{\mathbb{E}}_{h\sim\mathcal{P}} \ln \frac{\mathcal{P}(h)}{\mathcal{Q}(h)}$ *the KL-divergence between* $\mathcal{P}$ *and* $\mathcal{Q}$*.*

Proof. Let $\Gamma{=}(V,E)$ be the graph representing the dependencies between the random variables where *(i)* the set of vertices is $V{=}\mathbf{S}$, *(ii)* the set of edges E is defined such that $(((x,y),\epsilon),((x',y'),\epsilon')) \notin E \Leftrightarrow x \neq x'$. Then, applying Th. 8 of Ralaivola et al. [2010] with our notations gives

$$\mathrm{kl}(\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}})) \leq \frac{\chi(\Gamma)}{mn}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{mn+\chi(\Gamma)}{\delta\chi(\Gamma)}\right],$$

where $\chi(\Gamma)$ is the fractional chromatic number of Γ. From a property of Scheinerman and Ullman [2011], we have

$$c(\Gamma) \leq \chi(\Gamma) \leq \Delta(\Gamma)+1,$$

where $c(\Gamma)$ is the order of the largest clique in Γ and $\Delta(\Gamma)$ is the maximum degree of a vertex in Γ. By construction of Γ, $c(\Gamma){=}n$ and $\Delta(\Gamma){=}n{-}1$. Thus, $\chi(\Gamma){=}n$ and rearranging the terms proves Equation (9). Finally, by applying Pinsker's inequality (*i.e.*, $|a{-}b|{\leq}\sqrt{\frac{1}{2}\mathrm{kl}(a\|b)}$), we obtain Equation (10). □

E Proof of Theorem 7

Theorem 7. *For any distribution* $\mathbf{D}$ *on* $(X{\times}Y){\times}B$*, for any set of voters* $\mathcal{H}$*, for any prior* $\mathcal{P}$ *on* $\mathcal{H}$*, for any* n*, with probability at least* $1{-}\delta$ *over* $\mathbf{S}$*, for all posteriors* $\mathcal{Q}$ *on* $\mathcal{H}$*, for all* $i \in \{1,\ldots,m\}$*, for all distributions* π_i *on* $\boldsymbol{\mathcal{E}}_i$ *independent from a voter* $h \in \mathcal{H}$*, we have*

$$\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}}) \leq \frac{1}{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\frac{1}{2}\left(1-y_i h(x_i{+}\epsilon)\right) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\tfrac{2\sqrt{m}}{\delta}\right]} \tag{16}$$

$$\leq \overline{A_{\mathbf{S}}}(H_{\mathcal{Q}}) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\mathrm{TV}(\rho_i^h\|\pi_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\tfrac{2\sqrt{m}}{\delta}\right]}, \tag{17}$$

where $\overline{A_{\mathbf{S}}}(H_{\mathcal{Q}}) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\frac{1}{2}\Big[1{-}\min_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\Big(y_i\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}h(x_i{+}\epsilon)\Big)\Big]$*, and* $\mathrm{TV}(\rho\|\pi) = \displaystyle\mathop{\mathbb{E}}_{\epsilon\sim\pi}\frac{1}{2}\left|\left[\frac{\rho(\epsilon)}{\pi(\epsilon)}\right]-1\right|$.

Proof. Let $L_{h,(x,y),\epsilon}{=}\frac{1}{2}\left[1{-}yh(x{+}\epsilon)\right]$ for the sake of readability. The losses $\max_{\epsilon\in\boldsymbol{\mathcal{E}}_1}L_{h,(x_1,y_1),\epsilon}$, $\ldots$ $\max_{\epsilon\in\boldsymbol{\mathcal{E}}_1}L_{h,(x_m,y_m),\epsilon}$ are *i.i.d.* for any $h{\in}\mathcal{H}$. Hence, we can apply Theorem 20 of Germain et al. [2015] and Pinsker's inequality (*i.e.*, $|q{-}p|\leq\sqrt{\frac{1}{2}\mathrm{kl}(q\|p)}$) to obtain

$$\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\mathop{\mathbb{E}}_{(x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^n}\max_{\epsilon\in\boldsymbol{\mathcal{E}}}L_{h,(x,y),\epsilon} \leq \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}L_{h,(x_i,y_i),\epsilon} + \sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}.$$

Then, we lower-bound the left-hand side of the inequality with $\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}})$, we have

$$\overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}}) \leq \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\mathop{\mathbb{E}}_{((x,y),\boldsymbol{\mathcal{E}})\sim\mathbf{D}^n}\max_{\epsilon\in\boldsymbol{\mathcal{E}}}L_{h,(x,y),\epsilon}.$$

Finally, from the definition of ρ_i^h, and from Lemma 4 of Ohnishi and Honorio [2021], we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}L_{h,(x_i,y_i),\epsilon} &= \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\rho_i^h}L_{h,(x_i,y_i),\epsilon}\\
&\leq \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\rho_i^h\|\pi_i) + \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\pi_i}L_{h,(x_i,y_i),\epsilon}\\
&= \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\rho_i^h\|\pi_i) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\pi_i}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}L_{h,(x_i,y_i),\epsilon}\\
&\leq \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\rho_i^h\|\pi_i) + \overline{A_{\mathbf{S}}}(H_{\mathcal{Q}}).
\end{aligned}$$

□

F Details on the Validity of the Bounds

In this section, we discuss about the validity of the bound when *(i)* generating perturbed sets such as $\mathbf{S}$ from a distribution $\mathbf{D}$ dependent on the prior $\mathcal{P}$ *(ii)* selecting the prior $\mathcal{P}$ with $\mathcal{S}_t$.

Actually, computing the bounds implies perturbing examples, *i.e.*, generating examples from $\mathbf{D}$ that is defined as $\mathbf{D}((x,y),\epsilon) = D(x,y)\cdot\omega_{(x,y)}(\epsilon)$. However, in order to obtain valid bounds, $\omega_{(x,y)}$ must be defined *a priori*. Since the prior $\mathcal{P}$ is defined *a priori* as well, $\omega_{(x,y)}$ can be dependent on $\mathcal{P}$. Hence, $\omega_{(x,y)}$ boils down to generating perturbed example $(x+\epsilon, y)$ by attacking the prior majority vote $H_{\mathcal{P}}$ with PGD_{U} or IFGSM_{U}. Nevertheless, our selection of the prior $\mathcal{P}$ with $\mathcal{S}$ may seem like "cheating", but this remains a valid strategy when we perform a union bound.

We explain the union bound for Theorem 6, and the same technique can be applied for Theorem 7. Let $\mathbf{D}_1,\ldots,\mathbf{D}_T$ be T distributions defined as $\mathbf{D}_1 = D(x,y)\cdot\omega^1_{(x,y)}(\epsilon),\ldots,\mathbf{D}_T = D(x,y)\cdot\omega^T_{(x,y)}(\epsilon)$ on $(X\times Y)\times B$ where each distribution $\omega^t_{(x,y)}$ depends on the example (x,y) and possibly on the fixed prior $\mathcal{P}_t$. Furthermore, we denote as $(\mathbf{D}_t^n)^m$ the distribution on the perturbed learning sample consisted of m examples and n perturbations for each example. Then, for all distributions $\mathbf{D}_t$, we can derive a bound on the risk $\overline{R_{\mathbf{D}_t}}(H_{\mathcal{Q}})$ which holds with probability at least $1-\frac{\delta}{T}$, we have

$$\Pr_{\mathbf{S}_t\sim(\mathbf{D}_t^n)^m}\left[\forall\mathcal{Q},\ \mathrm{kl}(\overline{R_{\mathbf{S}_t}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}_t}}(H_{\mathcal{Q}}))\leq\frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_t)+\ln\frac{T(m+1)}{\delta}\right]\right]$$

$$=\Pr_{\mathbf{S}_1\sim(\mathbf{D}_1^n)^m,\ldots,\mathbf{S}_T\sim(\mathbf{D}_T^n)^m}\left[\forall\mathcal{Q},\ \mathrm{kl}(\overline{R_{\mathbf{S}_t}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}_t}}(H_{\mathcal{Q}}))\leq\frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_t)+\ln\frac{T(m+1)}{\delta}\right]\right]\geq 1-\tfrac{\delta}{T}.$$

Then, from a union bound argument, we have

$$\Pr_{\mathbf{S}_1\sim(\mathbf{D}_1^n)^m,\ldots,\mathbf{S}_T\sim(\mathbf{D}_T^n)^m}\Bigg[\forall\mathcal{Q},\ \mathrm{kl}(\overline{R_{\mathbf{S}_1}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}_1}}(H_{\mathcal{Q}}))\leq\frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_t)+\ln\frac{T(m+1)}{\delta}\right],$$

$$\text{and }\ldots,$$

$$\text{and }\mathrm{kl}(\overline{R_{\mathbf{S}_T}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}_T}}(H_{\mathcal{Q}}))\leq\frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_T)+\ln\frac{T(m+1)}{\delta}\right]\Bigg]\geq 1-\delta.$$

Hence, we can select $\mathcal{P}\in\{\mathcal{P}_1,\ldots,\mathcal{P}_T\}$ with $\mathcal{S}$, and let $\mathbf{D}((x,y),\epsilon) = D(x,y)\cdot\omega_{(x,y)}(\epsilon)$ be the distributions on $(X\times Y)\times B$ where $\omega_{(x,y)}(\epsilon)$ is dependent on $\mathcal{P}$ and on the example (x,y), we can say that

$$\Pr_{\mathbf{S}\sim(\mathbf{D}^n)^m}\left[\forall\mathcal{Q},\ \mathrm{kl}(\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}})\|\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}}))\leq\frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P})+\ln\frac{T(m+1)}{\delta}\right]\right]\geq 1-\delta. \tag{18}$$

Additionally, when applying the same process for Equations (11) and (12) in Theorem 7, we have

$$\Pr_{\mathbf{S}\sim(\mathbf{D}^n)^m}\Bigg[\forall\mathcal{Q},\ \overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}})\leq\frac{1}{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\sum_{i=1}^{m}\max_{\epsilon\in\boldsymbol{\mathcal{E}}_i}\frac{1}{2}\left(1-y_i h(x_i+\epsilon)\right)$$

$$+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P})+\ln\tfrac{2T\sqrt{m}}{\delta}\right]}\Bigg]\geq 1-\delta, \tag{19}$$

and

$$\Pr_{\mathbf{S}\sim(\mathbf{D}^n)^m}\Bigg[\forall\mathcal{Q},\ \overline{A_{\mathbf{D}^n}}(H_{\mathcal{Q}})\leq\overline{A_{\mathbf{S}}}(H_{\mathcal{Q}})+\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\mathrm{TV}(\rho_i^h\|\pi_i)$$

$$+\sqrt{\frac{1}{2m}\Big[\mathrm{KL}(\mathcal{Q}\|\mathcal{P})+\ln\tfrac{2T\sqrt{m}}{\delta}\Big]}\Bigg]\geq 1-\delta. \tag{20}$$

G Details on Algorithm 1 and on the computation of the bounds

In this section, we explain how we attack the examples and optimize the bounds in Algorithm 1. Moreover, we present the computation of the bounds after the optimization. Furthermore, remark that the bounds involve the number of epochs T (see Section F for more details).

Computing the bounds. Unlike Equation (19) or Equation (20), Equation (18) is not directly optimizable since we upper-bound a deviation (the kl) between the empirical and true risk. Hopefully, we can compute the bound when it is expressed with the inverse binary kl divergence kl^{-1} defined as $\mathrm{kl}^{-1}(a|\epsilon) = \max_{b\in[0,1]}\{\mathrm{kl}(a\|b) \leq \epsilon\}$. Equation (18) can be rewritten as

$$\overline{R_{\mathbf{D}}}(H_{\mathcal{Q}}) \leq \mathrm{kl}^{-1}\left(\overline{R_{\mathbf{S}}}(H_{\mathcal{Q}}) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{T(m+1)}{\delta}\right]\right).$$

Optimizing the bounds. During each epoch t of Step 2 in Algorithm 1, the posterior distribution $\mathcal{Q}_t$ is updated with $\nabla_{\mathcal{Q}_t}\mathtt{B}_{\mathbb{S}}(H_{\mathcal{Q}_t})$. The objective function associated to Equation (18) of Theorem 6 is

$$\mathtt{B}_{\mathbb{S}}(H_{\mathcal{Q}_t}) = \mathrm{kl}^{-1}\left(\overline{R_{\mathbb{S}}}(H_{\mathcal{Q}_t}) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}_t\|\mathcal{P}) + \ln\frac{T(m+1)}{\delta}\right]\right).$$

Note that the derivative of kl^{-1} and its computation can be found in [Reeb et al., 2018, Appendix A]. On the other hand, the objective function to optimize Equation (20) of Theorem 7 is defined as

$$\mathtt{B}_{\mathbb{S}}(H_{\mathcal{Q}_t}) = \overline{A_{\mathbb{S}}}(H_{\mathcal{Q}_t}) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathcal{Q}_t\|\mathcal{P}) + \ln\tfrac{2T\sqrt{m}}{\delta}\right]}.$$

Note that the TV distance is 0 when we sample one noise for each example, *i.e.*, when $n{=}1$. In consequence, the distance is not optimized in Algorithm 1. However, if we had $n{>}1$, we would have to minimize it.

Attacking the examples. The `attack` function used in Algorithm 1 differs from the attack that generates the perturbed set $\mathbf{S}$ (for the bound). Indeed, at each iteration (in both steps), the function attacks an example with the current model while $\mathbf{S}$ is generated with the prior majority vote $H_{\mathcal{P}}$ (the output of Step 1). Note that for all attacks, in order to be differentiable with respect to the input, we remove the sign function on the voters' outputs.

H About the (differentiable) decision trees

In this section, we introduce the differentiable decision trees, *i.e.*, the voters of our majority vote. Note that we adapt the model of Kontschieder et al. [2016] in order to fit with our framework: a voter must output a real between -1 and $+1$. An example of such a tree is represented in Figure 3.

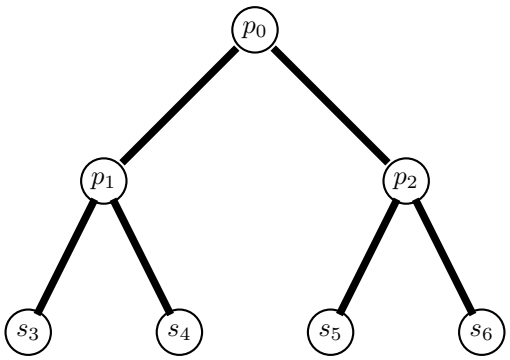

Figure 3: Representation of a (differentiable) decision tree of depth $l = 2$; The root is the node 0 and the leafs are 4; 5; 6 and 7. The probability $p_i(x)$ (respectively $1{-}p_i(x)$) to go left (respectively right) at the node i is represented by p_i (we omitted the dependence on x for simplicity). Similarly, the predicted label (a "score" between -1 and $+1$) at the leaf i is represented by s_i.

This differentiable decision tree is stochastic by nature: at each node i of the tree, we continue recursively to the left sub-tree with a probability of $p_i(x)$ and to the right sub-tree with a probability of $1{-}p_i(x)$; When we attain a leaf j, the tree predicts the label s_j. Precisely, the probability $p_i(x)$ is constructed by *(i)* selecting randomly 50% of the input features x and applying a random mask $M_i \in \mathbb{R}^d$ on x (where the k-th entry of the mask is 1 if the k-th feature is selected and 0 otherwise),

by *(ii)* multiplying this quantity by a learned weight vector $v_i \in \mathbb{R}^d$, and by *(iii)* applying a sigmoid function to output a probability. Indeed, we have

$$p_i(x) = \sigma\Big(\langle v_i, M_i \odot x \rangle\Big),$$

where $\sigma(a) = \left[1 + e^{-a}\right]^{-1}$ is the sigmoid function; $\langle a, b \rangle$ is the dot product between the vector a and b and $a \odot b$ is the elementwise product between the vector a and b. Moreover, s_i is obtained by learning a parameter $u_i \in \mathbb{R}$ and applying a $\tanh$ function, *i.e.*, we have

$$s_i = \tanh\Big(u_i\Big).$$

Finally, instead of having a stochastic voter, h will output the expected label predicted by the tree (see Kontschieder et al. [2016] for more details). It can be computed by $h(x) = f(x, 0, 0)$ with

$$f(x, i, l') = \begin{cases} s_i & \text{if } l' = l \\ p_i(x) f(x, 2i{+}1, l'{+}1) + (1 - p_i(x)) f(x, 2i{+}2, l'{+}1) & \text{otherwise} \end{cases}.$$

I Additional experimental results

In this section, we present the detailed results for the 6 tasks (3 on MNIST and 3 on Fashion MNIST) on which we perform experiments that show the test risks and the bounds for the different scenarios of (Defense, Attack). We train all the models using the same parameters as described in Section 4.2. Table 2 and Table 3 complement Table 1 to present the results for all the tasks when using the ℓ_2-norm with $b = 1$ (the maximum noise allowed by the norm). Then, we run again the same experiment but we use the ℓ_∞-norm with $b = 0.1$ and exhibit the results in Table 4 and Table 5. For the experiments on the 5 other tasks using the ℓ_2-norm, we have a similar behavior than **MNIST 1vs7** (presented in the paper). Indeed, using the attacks PGD_U and IFGSM_U as defense mechanism allows to obtain better risks and also tighter bounds compared to the bounds obtained with a defense based on UNIF (which is a naive defense). For the experiments on the 6 tasks using the ℓ_∞-norm, the trend is the same as with the ℓ_2-norm, *i.e.*, the appropriate defense leads to better risks and bounds.

We also run experiments that do not rely on the PAC-Bayesian framework. In other words, we train the models following only Step 1 of our adversarial training procedure (*i.e.*, Algorithm 1) using classical attacks (PGD or IFGSM): we refer to this experiment as a baseline. In our cases, it means learning a majority vote $H_{\mathcal{P}'}$ that follows a distribution $\mathcal{P}'$. As a reminder, the studied scenarios for the baseline are all the pairs $(\text{Defense}, \text{Attack})$ belonging to the set $\{—, \text{UNIF}, \text{PGD}, \text{IFGSM}\} \times \{—, \text{PGD}, \text{IFGSM}\}$. We report the results in Table 6 and Table 7. With this experiment, we are now able to compare our defense based on PGD_U or IFGSM_U and a classical defense based on PGD and IFGSM. Hence, considering the test risks $R^{\texttt{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$ (columns "Attack without U" of Tables 1 to 5) and $R^{\texttt{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$ (in Tables 6 and 7) , we observe similar results between the baseline and our framework.

Table 2: Test risks and bounds for 2 tasks of **MNIST** with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in *italic* the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (*i.e.*, we did *not* sample from the uniform distribution). Since Eq. (12) upperbounds Eq. (11) thanks to the TV term, we compute the two bound values of Theorem 7.

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.015*	*.015*	.015	.015	**0.060**	.067	*.015*	*.015*	.015	.015	**0.129**	0.135	**0.129**	.135
—	PGD_U	*.632*	***.628***	**.520**	.526	1.059	**.847**	*.672*	***.641***	**.683**	.684	**1.718**	2.405	1.392	**.962**
—	IFGSM_U	*.447*	***.443***	**.157**	.166	**0.387**	.572	*.461*	***.451***	**.337**	.345	**1.137**	2.090	0.776	**.669**
UNIF	—	*.024*	*.024*	.024	.024	**0.073**	.083	*.024*	*.024*	.024	.024	**0.140**	0.148	**0.140**	.148
UNIF	PGD_U	*.646*	***.619***	**.486**	.500	1.016	**.809**	*.649*	***.626***	**.648**	.650	**1.646**	2.417	1.338	**.915**
UNIF	IFGSM_U	*.442*	*.442*	**.128**	.139	**0.316**	.528	*.442*	*.442*	**.281**	.293	**0.907**	2.118	0.633	**.617**
PGD_U	—	***.024***	*.025*	**.024**	.025	**0.094**	.101	***.024***	*.025*	**.024**	.025	**0.158**	0.163	**0.158**	.163
PGD_U	PGD_U	*.148*	***.135***	.111	**.103**	0.360	**.355**	*.146*	***.136***	.129	**.120**	**0.442**	2.062	0.414	**.403**
PGD_U	IFGSM_U	*.104*	***.103***	.072	.072	0.277	.277	*.102*	*.102*	.090	**.084**	**0.358**	1.954	0.335	**.328**
IFGSM_U	—	*.027*	***.025***	.027	**.025**	**0.080**	.091	*.027*	***.025***	.027	**.025**	**0.146**	0.154	**0.146**	.154
IFGSM_U	PGD_U	*.188*	***.178***	**.111**	.119	**0.383**	.405	*.190*	***.178***	**.126**	.134	**0.501**	2.063	0.454	.454
IFGSM_U	IFGSM_U	*.126*	***.115***	.076	**.070**	**0.248**	.290	*.127*	***.115***	.091	**.085**	**0.371**	1.918	**0.329**	.342

(a) MNIST 4vs9

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.015*	*.015*	.015	.015	**.043**	.045	*.015*	*.015*	.015	.015	**.117**	0.118	**.117**	.118
—	PGD_U	*.279*	***.271***	**.232**	.234	.600	**.453**	*.284*	***.274***	.284	.284	**.829**	1.929	.724	**.524**
—	IFGSM_U	*.143*	***.137***	**.089**	.090	**.204**	.227	*.144*	***.139***	**.125**	.127	**.422**	1.662	.337	**.293**
UNIF	—	*.017*	*.017*	.017	.017	**.054**	.055	*.017*	*.017*	.017	.017	**.124**	0.125	**.124**	.125
UNIF	PGD_U	*.219*	***.201***	**.172**	.177	.433	**.350**	*.219*	***.209***	**.217**	.218	**.671**	1.810	.565	**.419**
UNIF	IFGSM_U	*.122*	*.122*	**.052**	.055	**.119**	.181	***.122***	*.123*	**.077**	.082	**.307**	1.554	**.242**	.248
PGD_U	—	***.013***	*.015*	**.013**	.015	.061	.061	***.013***	*.015*	**.013**	.015	**.131**	0.130	.131	**.130**
PGD_U	PGD_U	*.057*	*.057*	.045	**.041**	**.157**	.160	*.057*	*.057*	.055	**.045**	**.227**	1.536	.218	.218
PGD_U	IFGSM_U	*.043*	*.043*	**.027**	.031	**.114**	.119	***.042***	*.043*	.037	**.035**	**.187**	1.433	**.179**	.181
IFGSM_U	—	*.014*	***.012***	.014	**.012**	.057	.057	*.014*	***.013***	.014	**.013**	.128	**0.127**	.128	**.127**
IFGSM_U	PGD_U	*.077*	***.072***	.054	**.043**	**.170**	.174	*.076*	***.075***	.055	**.052**	**.252**	1.510	**.233**	.236
IFGSM_U	IFGSM_U	*.055*	***.048***	.034	**.030**	**.105**	.121	*.052*	***.051***	.039	**.032**	**.191**	1.379	**.177**	.185

(b) MNIST 5vs6

Table 3: Test risks and bounds for 3 tasks **Fashion MNIST** with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in *italic* the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (*i.e.*, we did *not* sample from the uniform distribution). Since Eq. (12) upperbounds Eq. (11) thanks to the TV term, we compute the two bound values of Theorem 7.

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.021*	***.020***	.021	**.020**	**0.060**	0.070	*.019*	*.019*	**.019**	**.019**	**0.130**	0.139	**0.130**	0.139
—	PGD_U	*.695*	***.650***	**.494**	.568	**1.042**	1.090	***.677***	*.686*	**.588**	.674	**1.326**	2.307	1.152	**1.082**
—	IFGSM_U	*.451*	*.451*	**.269**	.328	**0.585**	0.731	***.405***	*.438*	**.295**	.381	**0.878**	1.971	**0.730**	0.746
UNIF	—	*.071*	*.071*	.071	.071	**0.185**	0.191	*.071*	*.071*	.071	.071	**0.236**	0.241	**0.236**	0.241
UNIF	PGD_U	***.423***	*.477*	**.418**	.425	0.957	**0.755**	*.486*	*.486*	.513	.513	**1.372**	2.173	1.151	**0.869**
UNIF	IFGSM_U	***.326***	*.331*	.105	.105	**0.273**	0.422	*.333*	***.331***	.144	**.142**	**0.496**	1.642	**0.397**	0.504
PGD_U	—	*.034*	***.032***	.034	**.032**	**0.094**	0.114	*.034*	***.032***	.034	**.032**	**0.158**	0.174	**0.158**	0.174
PGD_U	PGD_U	***.103***	*.115*	**.086**	.091	**0.227**	0.289	***.102***	*.115*	**.096**	.101	**0.299**	1.985	**0.283**	0.338
PGD_U	IFGSM_U	***.092***	*.099*	**.073**	.076	**0.195**	0.248	***.092***	*.099*	.082	.082	**0.266**	1.914	**0.253**	0.299
IFGSM_U	—	***.028***	*.030*	**.028**	.030	**0.091**	0.105	***.027***	*.030*	**.027**	.030	**0.155**	0.166	**0.155**	0.166
IFGSM_U	PGD_U	*.115*	***.114***	.085	.085	**0.254**	0.287	***.112***	*.114*	**.096**	.101	**0.331**	2.026	**0.313**	0.337
IFGSM_U	IFGSM_U	***.095***	*.097*	**.067**	.068	**0.206**	0.232	***.093***	*.097*	**.080**	.081	**0.282**	1.927	**0.266**	0.285

(a) Fashion MNIST Sandall vs Ankle Boot

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.038*	***.037***	.038	**.037**	**.088**	.091	*.038*	***.037***	.038	**.037**	**.153**	0.155	**.153**	.155
—	PGD_U	*.292*	***.248***	.233	**.112**	.452	**.363**	*.289*	***.272***	.287	**.246**	**.578**	1.314	.525	**.479**
—	IFGSM_U	*.194*	***.154***	.132	**.075**	.300	**.262**	*.193*	***.181***	.176	**.148**	**.423**	1.103	.376	**.359**
UNIF	—	*.039*	*.039*	**.039**	**.039**	**.091**	.093	*.041*	***.039***	.041	**.039**	**.155**	0.157	**.155**	.157
UNIF	PGD_U	*.240*	***.220***	**.099**	.117	.346	**.332**	*.250*	***.231***	.250	**.245**	**.553**	1.228	.490	**.443**
UNIF	IFGSM_U	*.177*	***.171***	**.070**	.078	**.228**	.247	*.197*	***.185***	.186	**.164**	**.445**	1.046	.371	**.346**
PGD_U	—	*.045*	***.044***	.045	**.044**	.108	**.105**	*.046*	***.045***	.046	**.045**	.172	**0.167**	.172	**.167**
PGD_U	PGD_U	*.108*	***.100***	**.077**	.082	**.203**	.211	*.104*	***.100***	**.081**	.087	**.279**	1.118	.269	**.264**
PGD_U	IFGSM_U	*.094*	***.086***	.071	**.069**	**.184**	.186	*.090*	***.086***	.076	**.073**	**.257**	1.015	.248	**.241**
IFGSM_U	—	***.041***	*.043*	**.041**	.043	**.094**	.101	***.039***	*.042*	**.039**	.042	**.158**	0.163	**.158**	.163
IFGSM_U	PGD_U	***.106***	*.114*	**.078**	.092	**.220**	.226	***.109***	*.113*	**.084**	.095	**.293**	1.052	.279	**.275**
IFGSM_U	IFGSM_U	***.082***	*.087*	**.065**	.072	**.171**	.176	***.082***	*.089*	**.068**	.078	**.247**	0.927	.234	**.232**

(b) Fashion MNIST Top vs Pullover

ℓ_2-norm $b=1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.122*	*.122*	.122	.122	**0.276**	0.286	*.122*	*.122*	.122	.122	**0.318**	0.328	**0.318**	0.328
—	PGD_U	*.744*	***.738***	**.674**	.689	1.386	**1.066**	*.745*	***.740***	**.767**	.768	**1.773**	2.386	1.576	**1.180**
—	IFGSM_U	*.652*	***.646***	**.454**	.474	0.947	**0.887**	*.659*	***.648***	**.618**	.632	**1.597**	2.214	1.276	**0.992**
UNIF	—	*.204*	*.204*	.204	.204	0.444	0.444	*.204*	*.204*	.204	.204	**0.475**	0.476	**0.475**	0.476
UNIF	PGD_U	*.750*	***.714***	.682	**.671**	1.350	**1.069**	*.750*	***.719***	.752	**.749**	**1.732**	2.063	1.524	**1.189**
UNIF	IFGSM_U	*.605*	***.575***	**.423**	.431	0.871	**0.866**	*.605*	***.578***	.530	**.526**	**1.304**	1.860	1.091	**0.956**
PGD_U	—	*.168*	***.165***	.168	**.165**	**0.423**	0.428	*.167*	***.165***	.167	**.165**	0.463	**0.461**	0.463	**0.460**
PGD_U	PGD_U	***.389***	*.402*	**.306**	.369	0.768	**0.719**	***.390***	*.402*	**.319**	.403	**0.847**	2.354	0.810	**0.755**
PGD_U	IFGSM_U	***.361***	*.368*	**.298**	.324	0.693	**0.672**	***.362***	*.368*	**.320**	.361	**0.799**	2.258	0.754	**0.707**
IFGSM_U	—	***.150***	*.163*	**.150**	.163	**0.424**	0.428	***.149***	*.163*	**.149**	.163	**0.458**	0.461	**0.458**	0.461
IFGSM_U	PGD_U	***.391***	*.428*	.347	**.292**	0.778	**0.757**	***.390***	*.426*	.371	**.298**	**0.856**	2.327	0.820	**0.791**
IFGSM_U	IFGSM_U	***.356***	*.382*	.291	**.273**	**0.685**	0.689	***.354***	*.382*	.331	**.278**	**0.772**	2.218	0.734	**0.723**

(c) Fashion MNIST Coat vs Shirt

Table 4: Test risks and bounds for 3 tasks of **MNIST** with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in *italic* the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (*i.e.*, we did *not* sample from the uniform distribution). Since Eq. (12) upperbounds Eq. (11) thanks to the TV term, we compute the two bound values of Theorem 7.

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.005*	*.005*	.005	.005	**.017**	.019	*.005*	*.005*	.005	.005	**0.099**	0.100	**.099**	.100
—	PGD_U	*.454*	*.454*	**.375**	.384	.770	**.638**	*.492*	***.484***	.480	**.476**	**1.127**	2.031	.946	**.716**
—	IFGSM_U	*.428*	***.423***	**.350**	.361	.727	**.610**	*.474*	***.465***	.448	**.443**	**1.061**	2.008	.886	**.686**
UNIF	—	*.004*	*.004*	.004	.004	**.018**	.019	*.004*	*.004*	.004	.004	**0.099**	0.100	**.099**	.100
UNIF	PGD_U	***.487***	*.491*	**.369**	.392	.779	**.667**	*.512*	***.507***	**.484**	.487	**1.179**	2.083	.972	**.739**
UNIF	IFGSM_U	***.436***	*.442*	**.325**	.337	.664	**.598**	*.466*	***.459***	.417	.417	**1.023**	1.959	.841	**.671**
PGD_U	—	*.006*	*.006*	.006	.006	.024	.024	***.005***	*.006*	**.005**	.006	0.103	0.103	.103	.103
PGD_U	PGD_U	***.018***	*.020*	**.013**	.016	**.046**	.050	***.018***	*.020*	**.015**	.020	**0.127**	1.461	**.122**	.123
PGD_U	IFGSM_U	***.020***	*.021*	**.012**	.016	**.048**	.054	***.019***	*.021*	**.015**	.020	**0.130**	1.455	**.125**	.127
IFGSM_U	—	***.006***	*.007*	**.006**	.007	**.023**	.024	***.006***	*.007*	**.006**	.007	**0.102**	0.103	**.102**	.103
IFGSM_U	PGD_U	***.018***	*.019*	.016	.016	**.046**	.051	***.018***	*.019*	**.018**	.019	**0.126**	1.489	**.122**	.124
IFGSM_U	IFGSM_U	*.020*	*.020*	**.015**	.016	**.050**	.055	*.020*	*.020*	.020	**.019**	**0.131**	1.481	**.126**	.127

(a) MNIST 1 vs 7

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.015*	*.015*	.015	.015	**0.060**	0.067	*.015*	*.015*	.015	.015	**0.129**	0.135	**0.129**	0.135
—	PGD_U	***.929***	*.930*	**.651**	.662	1.367	**1.125**	***.920***	*.925*	**.874**	.880	**2.213**	2.661	1.792	**1.266**
—	IFGSM_U	*.935*	*.935*	**.601**	.609	1.243	**1.088**	***.926***	*.928*	**.800**	.806	**2.047**	2.615	1.649	**1.224**
UNIF	—	*.017*	*.017*	.017	.017	**0.062**	0.072	*.017*	*.017*	.017	.017	**0.131**	0.139	**0.131**	0.139
UNIF	PGD_U	*.895*	*.895*	**.615**	.623	1.302	**1.078**	***.884***	*.888*	**.815**	.818	**2.035**	2.722	1.670	**1.208**
UNIF	IFGSM_U	*.898*	*.898*	**.516**	.528	1.112	**1.027**	***.884***	*.890*	**.697**	.706	**1.875**	2.658	1.497	**1.153**
PGD_U	—	*.039*	***.037***	.039	**.037**	**0.093**	0.094	*.039*	***.037***	.039	**.037**	**0.156**	0.157	**0.156**	0.157
PGD_U	PGD_U	***.108***	*.109*	.090	.090	**0.200**	0.209	***.108***	*.109*	**.110**	.112	**0.337**	1.874	0.290	**0.271**
PGD_U	IFGSM_U	***.121***	*.124*	**.101**	.103	**0.229**	0.235	***.121***	*.124*	.126	**.125**	**0.378**	1.890	0.326	**0.297**
IFGSM_U	—	*.046*	***.044***	.046	**.044**	**0.102**	0.119	*.046*	***.044***	.046	**.044**	**0.164**	0.178	**0.164**	0.178
IFGSM_U	PGD_U	*.105*	***.093***	.091	**.078**	**0.203**	0.214	*.105*	***.093***	.108	**.089**	**0.321**	1.810	0.286	**0.269**
IFGSM_U	IFGSM_U	*.119*	***.095***	.102	**.080**	**0.220**	0.229	*.119*	***.095***	.122	**.090**	**0.357**	1.821	0.309	**0.283**

(b) MNIST 4 vs 9

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.015*	*.015*	.015	.015	**.043**	.045	*.015*	*.015*	.015	.015	**0.117**	0.118	**0.117**	.118
—	PGD_U	*.500*	***.499***	**.387**	.390	.923	**.744**	*.502*	***.500***	**.474**	.475	**1.361**	2.275	1.146	**.830**
—	IFGSM_U	*.519*	***.505***	**.395**	.398	.915	**.762**	***.514***	*.516*	.481	.481	**1.335**	2.283	1.129	**.847**
UNIF	—	*.015*	*.015*	.015	.015	**.052**	.053	*.015*	*.015*	.015	.015	**0.123**	0.124	**0.123**	.124
UNIF	PGD_U	***.529***	*.544*	**.388**	.393	.925	**.761**	***.517***	*.532*	**.481**	.482	**1.342**	2.349	1.137	**.848**
UNIF	IFGSM_U	***.536***	*.544*	**.372**	.379	.881	**.774**	***.523***	*.544*	**.451**	.456	**1.268**	2.348	1.077	**.857**
PGD_U	—	*.015*	***.014***	.015	**.014**	**.060**	.064	*.015*	***.014***	.015	**.014**	**0.130**	0.133	**0.130**	.133
PGD_U	PGD_U	***.055***	*.058*	**.037**	.039	**.131**	.143	***.056***	*.057*	.046	.046	**0.219**	1.619	**0.202**	.204
PGD_U	IFGSM_U	***.061***	*.065*	**.040**	.043	**.146**	.154	***.059***	*.062*	.050	**.046**	**0.232**	1.626	0.216	**.214**
IFGSM_U	—	*.019*	***.014***	.019	**.014**	.069	**.064**	*.018*	***.014***	.018	**.014**	0.136	**0.132**	0.136	**.132**
IFGSM_U	PGD_U	*.061*	*.061*	**.040**	.050	.143	**.142**	*.061*	*.061*	**.045**	.061	**0.218**	1.694	0.208	**.205**
IFGSM_U	IFGSM_U	***.066***	*.069*	**.044**	.054	.154	**.152**	***.065***	*.069*	**.048**	.068	**0.228**	1.708	0.216	**.214**

(c) MNIST 5 vs 6

Table 5: Test risks and bounds for 3 tasks of **Fashion MNIST** with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in *italic* the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (*i.e.*, we did *not* sample from the uniform distribution). Since Eq. (12) upperbounds Eq. (11) thanks to the TV term, we compute the two bound values of Theorem 7.

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.021*	***.020***	.021	**.020**	**0.060**	0.070	*.019*	*.019*	.019	.019	**0.130**	0.139	**0.130**	0.139
—	PGD_U	*.951*	***.944***	**.606**	.719	**1.275**	1.333	*.935*	***.920***	**.762**	.864	**1.617**	2.503	1.421	**1.317**
—	IFGSM_U	*.957*	***.947***	**.588**	.718	**1.231**	1.336	*.950*	*.950*	**.734**	.851	**1.587**	2.495	1.395	**1.316**
UNIF	—	***.076***	*.077*	**.076**	.077	**0.178**	0.184	***.076***	*.077*	**.076**	.077	**0.230**	0.235	**0.230**	0.235
UNIF	PGD_U	*.964*	***.961***	**.714**	.719	1.496	**1.265**	*.966*	***.963***	**.853**	.859	**2.098**	2.417	1.785	**1.416**
UNIF	IFGSM_U	*.978*	***.976***	**.627**	.632	1.306	**1.259**	*.979*	*.979*	**.758**	.762	**1.914**	2.422	1.597	**1.396**
PGD_U	—	*.041*	***.040***	.041	**.040**	0.114	**0.111**	*.041*	***.040***	.041	**.040**	0.173	**0.171**	0.173	**0.171**
PGD_U	PGD_U	*.098*	***.097***	.089	**.086**	**0.207**	0.210	*.099*	***.097***	.101	**.100**	**0.306**	1.826	0.281	**0.267**
PGD_U	IFGSM_U	*.113*	***.112***	.105	**.101**	**0.244**	0.246	*.115*	***.112***	.120	**.113**	**0.353**	1.853	0.321	**0.302**
IFGSM_U	—	***.045***	*.047*	**.045**	.047	**0.131**	0.137	***.045***	*.047*	**.045**	.047	**0.188**	0.194	**0.188**	0.194
IFGSM_U	PGD_U	***.100***	*.102*	.089	**.085**	**0.203**	0.232	***.100***	*.102*	**.102**	**.102**	**0.298**	1.645	**0.274**	0.287
IFGSM_U	IFGSM_U	***.112***	*.116*	.099	**.096**	**0.232**	0.260	***.112***	*.116*	.114	**.112**	**0.328**	1.687	**0.301**	0.313

(a) Fashion MNIST Sandall vs Ankle Boot

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.038*	***.037***	.038	**.037**	**.088**	.091	*.038*	***.037***	.038	**.037**	**0.153**	0.155	**0.153**	.155
—	PGD_U	*.596*	***.515***	.477	**.218**	.844	**.662**	*.590*	***.576***	.570	**.502**	**1.049**	1.924	0.948	**.857**
—	IFGSM_U	*.723*	***.623***	.573	**.257**	.971	**.751**	*.716*	***.695***	.678	**.598**	**1.189**	2.031	1.080	**.980**
UNIF	—	*.032*	*.032*	.032	.032	**.083**	.085	***.032***	*.033*	**.032**	.033	**0.149**	0.151	**0.149**	.151
UNIF	PGD_U	***.438***	*.439*	.356	**.245**	.813	**.563**	*.435*	*.435*	.423	**.312**	**1.082**	1.867	0.959	**.688**
UNIF	IFGSM_U	***.546***	*.547*	.453	**.325**	.974	**.690**	***.544***	*.547*	.530	**.409**	**1.266**	2.009	1.128	**.823**
PGD_U	—	***.048***	*.053*	**.048**	.053	**.115**	.130	***.048***	*.053*	**.048**	.053	**0.177**	0.188	**0.177**	.188
PGD_U	PGD_U	***.102***	*.116*	**.089**	.099	**.205**	.223	***.102***	*.116*	**.096**	.115	**0.282**	1.323	**0.266**	.278
PGD_U	IFGSM_U	***.120***	*.135*	**.102**	.115	**.237**	.255	***.120***	*.135*	**.109**	.133	**0.318**	1.380	**0.299**	.309
IFGSM_U	—	*.051*	***.045***	.051	**.045**	.120	**.115**	*.051*	***.045***	.051	**.045**	0.179	**0.175**	0.179	**.175**
IFGSM_U	PGD_U	*.106*	***.094***	.091	**.085**	.211	**.193**	*.106*	***.094***	.102	**.097**	**0.292**	1.488	0.273	**.252**
IFGSM_U	IFGSM_U	*.120*	***.111***	**.101**	.102	.239	**.218**	*.119*	***.111***	.113	.113	**0.322**	1.546	0.299	**.277**

(b) Fashion MNIST Top vs Pullover

ℓ_∞-norm $b=0.1$		Algo.1 with Eq. (9)						Algo.1 with Eq. (12)							
		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$R_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 6		Attack without U $R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{Q}})$		$A_{\mathbf{T}}(H_{\mathcal{Q}})$		Th. 7 - Eq. (12)		Th. 7 - Eq. (11)	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	*.122*	*.122*	.122	.122	**0.276**	0.286	*.122*	*.122*	.122	.122	**0.318**	0.328	**0.318**	0.328
—	PGD_U	***.884***	*.887*	**.781**	.795	1.579	**1.268**	***.882***	*.886*	**.864**	.872	**2.020**	2.640	1.803	**1.390**
—	IFGSM_U	***.901***	*.902*	**.756**	.774	1.558	**1.272**	***.901***	*.902*	**.865**	.876	**2.032**	2.651	1.795	**1.393**
UNIF	—	*.166*	*.166*	.166	.166	**0.352**	0.357	.166	*.166*	*.166*	.166	**0.389**	0.394	**0.389**	0.394
UNIF	PGD_U	***.911***	*.914*	**.796**	.798	1.402	**1.326**	***.913***	*.914*	.896	**.888**	**1.934**	2.325	1.713	**1.447**
UNIF	IFGSM_U	***.935***	*.937*	**.787**	.798	1.392	**1.350**	***.934***	*.936*	.887	**.882**	**1.905**	2.378	1.693	**1.469**
PGD_U	—	*.163*	***.162***	.163	**.162**	**0.386**	0.395	*.163*	***.162***	.163	**.162**	**0.419**	0.430	**0.419**	0.430
PGD_U	PGD_U	***.394***	*.396*	.359	**.329**	0.764	**0.673**	***.394***	*.396*	.403	**.394**	**0.954**	2.321	0.865	**0.726**
PGD_U	IFGSM_U	***.475***	*.480*	.442	**.410**	0.910	**0.769**	***.477***	*.480*	.487	**.472**	**1.121**	2.411	1.020	**0.826**
IFGSM_U	—	***.167***	*.168*	**.167**	.168	0.411	**0.395**	***.167***	*.168*	**.167**	.168	0.445	**0.429**	0.445	**0.429**
IFGSM_U	PGD_U	*.396*	***.373***	.359	**.293**	0.772	**0.641**	*.396*	***.373***	.405	**.328**	**0.970**	2.368	0.877	**0.692**
IFGSM_U	IFGSM_U	*.465*	***.428***	.424	**.334**	0.891	**0.705**	*.465*	***.429***	.470	**.372**	**1.090**	2.425	0.995	**0.758**

(c) Fashion MNIST Coat vs Shirt

Table 6: Test risks for 6 tasks of **MNIST** and **Fashion MNIST** datasets for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$ using ℓ_2-norm. The results of these tables are computed considering defenses of the literature, *i.e.*, adversarial training using PGD or IFGSM. We also add an adversarial training using UNIF for the completeness of comparison between this baseline defense and our algorithm. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$.

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.005	.005
—	PGD	**.326**	.327
—	IFGSM	.122	**.121**
UNIF	—	.005	.005
UNIF	PGD	.191	**.190**
UNIF	IFGSM	**.071**	.072
PGD	—	.007	.007
PGD	PGD	.027	**.026**
PGD	IFGSM	.022	**.021**
IFGSM	—	**.005**	.006
IFGSM	PGD	.041	**.035**
IFGSM	IFGSM	.021	.021

(a) MNIST 1 vs 7

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.015	.015
—	PGD	.692	.692
—	IFGSM	.464	**.462**
UNIF	—	.024	.024
UNIF	PGD	.653	.653
UNIF	IFGSM	.441	**.438**
PGD	—	**.024**	.027
PGD	PGD	**.136**	.138
PGD	IFGSM	**.097**	.102
IFGSM	—	**.022**	.027
IFGSM	PGD	**.166**	.186
IFGSM	IFGSM	**.113**	.124

(b) MNIST 4 vs 9

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.015	.015
—	PGD	.283	.283
—	IFGSM	.144	.144
UNIF	—	.017	.017
UNIF	PGD	.220	**.219**
UNIF	IFGSM	.122	.122
PGD	—	.014	**.013**
PGD	PGD	.056	**.055**
PGD	IFGSM	.045	**.041**
IFGSM	—	**.013**	.014
IFGSM	PGD	.077	**.070**
IFGSM	IFGSM	.053	**.047**

(c) MNIST 5 vs 6

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.019	.019
—	PGD	.709	**.708**
—	IFGSM	.426	**.414**
UNIF	—	**.071**	.072
UNIF	PGD	.531	.531
UNIF	IFGSM	.331	**.329**
PGD	—	**.034**	.036
PGD	PGD	.107	**.103**
PGD	IFGSM	.091	**.087**
IFGSM	—	.031	**.029**
IFGSM	PGD	.125	**.108**
IFGSM	IFGSM	.104	**.090**

(d) Fashion MNIST
Sandall vs Ankle Boot

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.038	.038
—	PGD	.286	**.285**
—	IFGSM	.188	**.186**
UNIF	—	.041	**.039**
UNIF	PGD	.249	**.248**
UNIF	IFGSM	.197	**.192**
PGD	—	**.043**	.045
PGD	PGD	**.102**	.117
PGD	IFGSM	**.090**	.094
IFGSM	—	**.038**	.040
IFGSM	PGD	.120	**.106**
IFGSM	IFGSM	.092	**.080**

(e) Fashion MNIST
Top vs Pullover

ℓ_2-norm, $b=1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.122	.122
—	PGD	.768	**.767**
—	IFGSM	.683	**.680**
UNIF	—	.204	.204
UNIF	PGD	**.753**	.754
UNIF	IFGSM	.607	**.606**
PGD	—	.182	**.178**
PGD	PGD	.453	**.412**
PGD	IFGSM	.408	**.379**
IFGSM	—	.148	**.146**
IFGSM	PGD	**.405**	.411
IFGSM	IFGSM	.369	**.364**

(f) Fashion MNIST
Coat vs Shirt

Table 7: Test risks for 6 tasks of **MNIST** and **Fashion MNIST** datasets for all pairs (Defense,Attack) with the two voters' set $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$ using ℓ_∞-norm. The results of these tables are computed considering defenses of the literature, *i.e.*, adversarial training using PGD or IFGSM. We also add an adversarial training using UNIF for the completeness of comparison between this baseline defense and our algorithm. The results in **bold** correspond to the best values between results for $\mathcal{H}$ and $\mathcal{H}^{\text{SIGN}}$.

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.005	.005
—	PGD	.499	**.498**
—	IFGSM	**.479**	.480
UNIF	—	.004	.004
UNIF	PGD	.516	**.515**
UNIF	IFGSM	.467	.467
PGD	—	**.006**	.007
PGD	PGD	.019	.019
PGD	IFGSM	.021	.021
IFGSM	—	.007	.007
IFGSM	PGD	**.017**	.018
IFGSM	IFGSM	**.019**	.020

(a) MNIST 1 vs 7

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.015	.015
—	PGD	.921	.921
—	IFGSM	.923	.923
UNIF	—	.017	.017
UNIF	PGD	.877	**.876**
UNIF	IFGSM	.877	.877
PGD	—	.041	**.040**
PGD	PGD	**.108**	.109
PGD	IFGSM	**.122**	.123
IFGSM	—	.057	**.044**
IFGSM	PGD	.109	**.101**
IFGSM	IFGSM	.119	**.108**

(b) MNIST 4 vs 9

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.015	.015
—	PGD	.498	.498
—	IFGSM	.511	**.510**
UNIF	—	.015	.015
UNIF	PGD	.512	**.511**
UNIF	IFGSM	.511	.511
PGD	—	.014	.014
PGD	PGD	.065	**.058**
PGD	IFGSM	.068	**.065**
IFGSM	—	.018	**.017**
IFGSM	PGD	**.061**	.063
IFGSM	IFGSM	**.069**	.071

(c) MNIST 5 vs 6

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.019	.019
—	PGD	.938	.938
—	IFGSM	**.948**	.949
UNIF	—	**.076**	.077
UNIF	PGD	.970	**.969**
UNIF	IFGSM	.981	.981
PGD	—	.041	**.040**
PGD	PGD	.098	**.097**
PGD	IFGSM	.115	**.111**
IFGSM	—	.112	**.047**
IFGSM	PGD	**.045**	.100
IFGSM	IFGSM	**.101**	.114

(d) Fashion MNIST
Sandall vs Ankell Boot

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.038	.038
—	PGD	**.574**	.577
—	IFGSM	.700	**.696**
UNIF	—	**.032**	.033
UNIF	PGD	**.428**	.435
UNIF	IFGSM	**.540**	.550
PGD	—	**.047**	.049
PGD	PGD	.101	**.097**
PGD	IFGSM	.118	**.112**
IFGSM	—	.049	**.048**
IFGSM	PGD	.100	**.090**
IFGSM	IFGSM	.112	**.108**

(e) Fashion MNIST
Top vs Pullover

ℓ_∞-norm, $b = 0.1$		$R^{\text{ROB}}_{\mathcal{T}}(H_{\mathcal{P}'})$	
Defense	Attack	$\mathcal{H}^{\text{SIGN}}$	$\mathcal{H}$
—	—	.122	.122
—	PGD	.879	.879
—	IFGSM	.898	.898
UNIF	—	.166	.166
UNIF	PGD	.913	**.911**
UNIF	IFGSM	.934	**.933**
PGD	—	**.164**	.167
PGD	PGD	.398	**.395**
PGD	IFGSM	**.479**	.481
IFGSM	—	**.163**	.169
IFGSM	PGD	**.356**	.391
IFGSM	IFGSM	**.422**	.461

(f) Fashion MNIST
Coat vs Shirt